DYNAMIC ANALYSIS OF FALSE INFORMATION SPREAD OVER SOCIAL MEDIA: 5G-COVID 19 CONSPIRACY THEORY

by

Orkun İrsoy

B.S., Industrial Engineering, Boğaziçi University, 2019

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Industrial Engineering

Boğaziçi University

2022

DYNAMIC ANALYSIS OF FALSE INFORMATION SPREAD OVER SOCIAL MEDIA: 5G-COVID 19 CONSPIRACY THEORY

APPROVED BY:

Assoc. Prof. Gönenç Yücel .....
(Thesis Supervisor)

Prof. Yaman Barlas .....
(Thesis Co-supervisor)

Assist. Prof. Özge Karanfil .....

Prof. Ali Kerem Saysel .....

DATE OF APPROVAL: 03.10.2022

ACKNOWLEDGEMENTS

I am obliged to express my gratitude to my thesis supervisors and mentors Yaman Barlas and Gönenç Yücel. Apart from their guidance and motivational support, they have always inspired me to improve academically and personally with their fellowship and affinity. I should also thank Özge Karanfil and Ali Kerem Saysel for both their participation in my thesis committee as jury members and their mentorship during my graduate studies.

A unique part of this acknowledgment should be dedicated to Naz Beril Akan and Şanser Güz as we spent most of our time in the department together. I feel genuinely grateful to have them by my side in this long and bumpy journey as friends and co-authors.

I wish to thank my dearest Seray Işık, Merve Erdoğan, and İdil Cömert as they always provide the support I need in my life whether it is a pad on the back, a shoulder to cry on, or a slap on the face.

I wish to thank my friends in the SESDYN Lab Feyyaz Şentürk, Gizem Aktaş, Gizem Taş, Zeynep Hasgül, Elif Bal, Mehmet Can Tunca, Nefel Tellioğlu. Being a member of SESDYN always felt like being a part of a family, thus I am looking forward to meeting with them in manifold alumni meetings, gatherings, and conferences.

Lastly, I want to my colleagues Tuğçe Türkmendağ, Buğra Çınar, Tarkan Temizöz, Ekin Özgürbüz, Çiğdem Karademir, İlayda Çelenk, and Ezgi Topuz for their usual support in my burnout state and also for the joyful conversations during the overextended coffee breaks or Wednesday night gatherings.

I thank TUBİTAK for their financial support as part of their BİDEB 2210-A graduate scholarship program.

ABSTRACT

DYNAMIC ANALYSIS OF FALSE INFORMATION SPREAD OVER SOCIAL MEDIA: 5G-COVID 19 CONSPIRACY THEORY

The spread of false information via online social networks is a critical societal issue with various potential harms. Although there are huge efforts both in research and application to mitigate this problem, it persists with an increasing magnitude of results ranging from political manipulation to violent attacks. In our research, we built a causal simulation model to combine the existing accumulated knowledge in the literature and provide a formal model to evaluate the governing dynamics for the specific case of the viral spread of the 5G-COVID-19 conspiracy theory. The model makes use of both qualitative and quantitative data and successfully generates the observed dynamics for the 5g narrative. Results from the base run suggest that the dominance of believers in the active discussion on social media is overrepresented relative to the total population. Moreover, common mitigation strategies proposed in the literature such as limiting the interaction with believers of the misinformation often seem to produce worse outcomes for specific cases which indicates policy resistance. In addition, scenario analysis suggests that the involvement of neutral people in sharing misinformation or superspreader actors might be enough to induce the system to pass the tipping point and generate an infodemic. The current analysis presents several trade-offs while discussing the underlying reasons through posterior analysis. In further research, we plan to expand our analysis by the inclusion of other user profiles, experiment with other mitigation strategies, and discuss the potential similarities and differences of our case with other types of false information dynamics.

ÖZET

SOSYAL MEDYADA YANLIŞ BİLGİ YAYILIMININ DİNAMİK İNCELENMESİ: 5G-COVID 19 KOMPLO TEORİSİ

Sosyal medya aracılığıyla yanlış bilgi yayılımı çeşitli tehlikeler barındıran ciddi bir sosyal meseledir. Bu durumu hafifletmek için teoride ve pratikte büyük çaba harcanmasına rağmen, yanlış bilgi yayılımı endişe verici bir hızda artmakta ve politik manipülasyondan fiili saldırılara varan çeşitli sonuçlara sebep olmaktadır. Araştırmamızda, literatürdeki bilgileri birleştirmek ve spesifik olarak 5G-COVID 19 komplo teorisinin viral yayılım dinamiklerini incelemek için nedensel bir benzetim modeli oluşturduk. Yapılan model hem nitel hem de nicel verileri kullanarak 5G kurgusunun dinamiklerini başarıyla temsil etmektedir. İlk analizden elde edilen sonuçlar sosyal medyada aktif olarak süren tartışmada 5G kurgusuna inananların baskınlığının toplam nüfusa göre daha fazla temsil edildiğini göstermektedir. Ek olarak, literatürde önerilen yanlış bilgiye inanan kesimle etkileşimin sınırlandırılması gibi stratejilerin, belirli vakalarda çok daha kötü sonuçlar yarattığı görülmektedir. Yapılan senaryo analizi, tarafsız kişilerin yanlış bilgi paylaşımına dahiliyetinin veya süper yayıcı aktörlerin etkinliğinin tetikleyici bir unsur olarak yanlış bilgi salgını oluşturması için yeterli olabileceğine işaret etmektedir. Mevcut çalışma arka plan analizi yoluya yapısal nedenler arasındaki dengelere dair çıkarımlar sunmaktadır. İlerleyen çalışmalarda, başka kullanıcı profillerinin dahil edilmesi, başka stratejilerinin denenmesi ve bizim örneğimizle diğer türlerdeki yanlış bilgi dinamiklerinin olası benzerlik ve farklılıklarının tartışılmasıyla analizi genişletmeyi planlıyoruz.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Following the increased accessibility of communication technologies and development of novel media tools, communication methods have shifted significantly toward digital communication. The vast majority of people use social media, including people of all ages and socioeconomic backgrounds. As a result, an increasing number of people are using social media to gather and disseminate information on a variety of topics, including critical information. For instance, according to Reuters, nearly two-thirds of adult Americans use social media as a source of news [1].

This new mode of communication offers numerous benefits, including the promotion of engagement and the reduction of barriers among people all over the world by providing an alternative to face-to-face socialization. Information spreads faster and to a larger audience on these social networks. However, because the content is created by users without any review process, unlike traditional media, the content's validity cannot be verified. As a result, whether willingly or unwillingly, people may spread false information. One of the most recent example demonstrating the potential harms of this phenomenon is the "infodemic" during the COVID-19 crisis, with results such as various ineffective and possibly harmful remedies, to outright rejection of the existence of the virus [2].

A recent example of such viral false information spread is 5G technology being one of the causes of COVID-19 or increasing its spread. The debate over the topic quickly erupted in the United Kingdom, particularly on social media platforms. Although fact-checking organizations or experts falsified the concerns related to this link, corrections were insufficient to alleviate the concerns, resulting in 5G tower arsons in Birmingham and Merseyside, United Kingdom [3].

Given the seriousness of the repercussions of misinformation dissemination, a massive body of research is carried out to tackle various facets of the problem. Researchers from various fields attempt to comprehend the psychological and cognitive

drivers underlying the phenomenon, analyze the data at hand to deduce why and how false information spreads, develop novel methods to detect such information, and develop mitigation strategies to combat misinformation.

Unfortunately, due to the complexity of the problem, mitigation measures used today are far from creating a structural solution but instead serve as symptom relief. The use of warning labels, which is one of the dominant tactics used by social media platforms, produces an "Implied Truth Effect" on unlabeled information [2] or may increase online traffic for the labeled content [4]. Fact-checking services attempt to verify the accuracy of the contents, although the rate of information production has increased far faster than the capacity of confirmation services has expanded [5]. Extensive data science research on false news detecting methods lays the door for the development of smart bots [6].

Despite great efforts in both research and application, the failure to develop effective mitigation techniques indicates the requirement for a dynamic systems approach. As a result, a systems perspective of the situation that combines current literature findings might identify potential leverage points and policy resistances to obtaining structural solutions to the problem at hand. In this regard, we argue that constructing a causal simulation model will constitute a theoretical basis to discuss the structural properties and allow us to derive practical insights and experiment with different scenarios.

In the following sections, we provide a brief overview of both the general problem of misinformation spread and the 5G-COVID 19 misinformation, in particular. Then we scrutinize the System Dynamics model that is built for the 5G-COVID 19 misinformation. The credibility and the validity of the model and the resulting base behavior are discussed. Finally, by conducting numerous simulation experiments, we infer some characteristics of the problem, explore what-if scenarios, and discuss effectiveness of various policy interventions.

2. LITERATURE REVIEW AND RESEARCH OBJECTIVE

2.1. Background on Misinformation and 5G-COVID 19 Conspiracy

Although false information dissemination is not a contemporary phenomenon, as evidenced by the 'Great Moon Hoax' in 1835 [5], the availability of highly linked world-wide platforms in the present world, allows anybody to transmit information to millions of individuals in a matter of minutes [7] further increasing the reach and severity of the problem. False information causes issues ranging from political manipulation of large groups of people [8] and stifling rescue efforts during a crisis to even a terror strike [7]. One such instance is the Pizza Gate conspiracy, which resulted in a person firing a gun at a neighborhood shop in response to reports of child trafficking [9]. Another example is Facebook's claim of voter tampering in the 2016 Presidential Election [10,11]. Given the gravity of the effects, the World Economic Forum has identified false information spread on digital platforms as one of the main challenges to society [12].

Many definitions and classifications of false information exist in the literature from rumors to "Fake News". One prevalent classification dimension is the intention of the agent where "misinformation" refers to unintentionally spreading information whereas "disinformation" is intentional [7,9,13]. Another categorization is the knowledge-based differentiation i.e., whether the information is purely factual or opinion-based [7, 14]. Such classifications are not made solely for simplification purposes, rather they are necessities as each different category causes different problems, stems from various causes, and requires unique research perspective to tackle.

The problem of false information spread on social media has various drivers, both at the individual and aggregate levels. At the individual level, various psychological and cognitive factors are thought to be effective, and many researchers are trying to find answers to questions: Do political motives drive susceptibility to misinformation, does repeated exposure lead to higher susceptibility to false beliefs, which cognitive

processes are influential on vulnerability to misinformation, and how can we design better corrective messages [5, 13, 15, 16]. From a more holistic perspective, another line of research focuses on the properties of these social networks, such as whether preferential attachment in these networks forms echo chambers (repetitive exposure of specific information due to homogeneous social clusters), whether any structural reasons make specific networks more susceptible to misinformation spread, and whether any distinguishing characteristics differentiate the propagation dynamics of misinformation compared to true information [14, 17–19]. A huge effort is put into misinformation detection with machine learning using either content or context-based cues to design early interventions [9]. Finally, simulation studies act as a testing platform to test the effectiveness of various intervention strategies or develop novel hypotheses about the underlying mechanisms of the problem [6, 20, 21].

Perhaps the most recent and critical forms of misinformation are experienced during the COVID-19 outbreak. Because of the ambiguity surrounding the situation, misleading information swiftly disseminates across borders, including conspiracy theories, fictitious miracle cures, and material that trivializes the infection [22]. One such case that emerged in early January 2020 was the conspiracy theory suggesting a link between the installation of new 5G towers with the spread of the virus [3, 23]. Unfortunately, the spread of rumors did not solely become an instance of misinformation but the escalated panic yielded multiple attacks on 5G towers in the UK [24, 25].

Many researchers investigate different aspects of the "5G-COVID 19" conspiracy theory and its spread as it epitomizes the potential harms of viral misinformation. Ahmed and colleagues [3] used Social Network Analysis to analyze the Twitter chatter during peak times of the chatter. Their analysis reveals that the number of people genuinely believing the conspiracy is rather low compared to the volume of the tweets. They conclude that apart from the believers, anti-conspiracy tweets, click baits or satiric tweets also contribute as much as believers, which further increases the dissemination of false information. Agley and Xiao [26] conducted an online survey on the believability of various conspiracy theories including the 5G narrative. In their work, they analyze the relationship between believability scores for different profiles

and their relationship with trust in science. Bruns, Harrington, and Hurcombe [23] use both quantitative and qualitative methods to understand how such misinformation escalated quickly up to violent attacks. They analyze the Facebook conversations from the start of the first rumor until the arson attempts by defining different phases and providing in-depth analysis for each phase. They discuss how pre-existing conspiracy networks or super-spreaders such as celebrities affected the propagation and the potential pitfalls that resulted in such virality. Finally, one of the most recent works regarding 5G-COVID 19 misinformation focuses on long-term diffusion across different countries and communication platforms [27]. They provide a detailed analysis of propagation on Twitter, and combine manual labeling and machine learning to obtain an estimate of the number of users sharing misinformation. In addition, factors that make this particular misinformation viral is thoroughly discussed. The authors conclude with key findings such as the involvement of financially motivated agents and specific policy recommendations such as the importance of ex-post observation and sharing results internationally to prevent reoccurrences of adversities in other countries.

2.2. Epidemiological Models of Misinformation Spread

Epidemiological models have wide application areas in social sciences as they can be used to model various processes from innovation adoption to rumor propagation [28]. Such models provide possible predictions on the course of events and allow policy makers to understand the behavior of the system. Models are highly simplified representations of the real system at hand. However, such general models might be useful to infer generic behavior patterns, provide a holistic approach of the problem, and reveal analogies and differences among a set of related problems [29].

Typical epidemiological models include various compartments or states such as Susceptible (S), Exposed (E), Infected (I), and Recovered (R) to represent different stages of an individual. Such models are utilized both for individual and aggregate level modeling practices. Typically transitions between states or compartments are determined by simple equations and parameters such as infection probabilities, infection duration, and contact frequencies. Since such equations involve causal relations

between different components of the system, these models can provide a causal analysis of the problem in addition to providing predictions.

The smallest types of compartmental epidemiological models are the Susceptible-Infected (SI) and Susceptible-Infected-Susceptible (SIS) models. For the SIS model people become susceptible again after a period of time as opposed to the SI model which assumes the transition from S to I is unidirectional. An extension to these models is the inclusion of "Recovered" or "Removed" component where the recovered people gain immunity to the infection (See Figure 3.1). Thus, Susceptible-Infected-Recovered (SIR) model assumes people never become susceptible again after gaining immunity whereas for the Susceptible-Infected-Recovered-Susceptible (SIRS) model there is a transition rate of recovered people to become susceptible again. A further extension is to include the "Exposed" state (SEIR models) to account for people who are exposed to the infection but the infection is first in incubation [28, 30–32].

By adapting the relevant definitions, such models are used to represent the diffusion of various types of misinformation. "Susceptible" usually represents the people who have not heard about the false information whereas "Infected" denotes people who have shared (or currently spreading) misinformation. Based on their applications, researchers include and adapt different definitions such as Recovered [20, 33], Skeptics [31, 34], Protesters [35], and Fact-checkers [36, 37].

For our specific case 5G-COVID 19 conspiracy theory, Kauk, Kreysa, and Schweinberger [20] use epidemiological modeling to build a SIR (Susceptible, Infected, and Recovered) model to simulate different mitigation strategies such as fact-checking and tweet deletion, and evaluate their effectiveness. The authors also point out a few shortcomings of the SIR model such as the inability to account for humorous or opposing tweets, reappearing incidence bursts, and extreme peaks thus calling for more complex models that account for such behavior.

2.3. Research Objectives

Given the severity and complexity of the problem at hand, the huge magnitude of the literature on misinformation is natural. However, the current attempts to combat misinformation are far from effective. The fundamental research in this domain is usually focused on one specific dimension of the problem such as propagation, detection, psychological factors, or network properties. Furthermore, the findings and conclusions are naturally sensitive to various factors such as the type of misinformation, the communication platform distributed, or psychological characteristics and dispositions of the investigated population. The modeling efforts, on the other hand, provide an overly generalized overview of the problem without benefiting from highly specific outputs generated by the fundamental research. The focal points of such studies are usually numerical analysis, prediction, and obtaining better fitting models to the observed data. Even in cases where such models are utilized for causal analysis and policy testing, the underlying simulation models rarely account for unique characteristics of the selected case, therefore interpretability of findings for other cases of misinformation is often questionable.

Thus we believe that more complex models that utilize the accumulated knowledge for the specific cases would be valuable to generate a holistic view of the problem. With adequate resolution, such models can serve as a platform to transfer highly specialized knowledge from one case to another and generate interpretability at the scale of the systems.

Undoubtedly, generating a one-size-fits-all model that can account for all types of misinformation spread is neither realistic nor feasible. However, we believe building a simulation model for a selected would pave the way for such efforts. System Dynamics methodology is an ideal fit for such a task as it allows the integration of main dynamic factors and provides a causal interpretation of the emerging dynamics. We believe the spread of the 5G-COVID 19 conspiracy theory is a perfect selection as a specific case due to the availability of quantitative and qualitative literature on the subject.

Therefore, the aim of the study to build a formal dynamic simulation model to analyze the spread of 5G-COVID-19 conspiracy theory in United Kingdom. We build a System Dynamics model for this specific case utilizing specific assumptions and simplifications regarding the 5G-COVID 19 misinformation. We then provide an analysis of the model to identify the causal feedback structure, evaluate the system behavior for manifold scenarios, and assess the effectiveness of potential structural mitigation strategies. We believe that from a methodological perspective, the constructed model will serve as an extension of currently adopted epidemiology models of false information diffusion to account for more complex cases, and from the perspective of the problem domain, it will contribute to a deeper understanding of the problem and provide a testing platform for different policies.

3. METHODOLOGY

In this study, the System Dynamics methodology is utilized to model the problem at hand. It is a widely used modeling methodology that is particularly used to address complex problems with many interacting components. System Dynamics models provide a dynamic hypothesis about the causal structure that generates the problem and allow the analysis of various policies and scenarios using simulation [38].

The problem addressed in our study is dynamic in nature, includes a human dimension, and contains complex interactions, feedback loops, and nonlinearities which makes System Dynamics particularly useful since it is advantageous in addressing such properties [39].

The main building blocks of System Dynamics models are "Stocks" and "Flows". Stocks represent the accumulated variables which can be either physical (Body Weight) or abstract (Stress Level) whereas flows represent the changes in stock variables. Inflows (Weight Gain) increase the stock level whereas outflows (Weight Loss) decrease it. An additional variable type is auxiliary variables or converters that are used to denote the intermediary parameters included in the model boundary. They can be either constant during the simulation or they can be dynamic either defined by an equation or a graphical function.

Stock-Flow Diagrams are an adequate form of representation of the system being analyzed. One such diagram for illustration purposes is presented in Figure 3.1. The model is the SIR model presented in Sterman [38]. Here Susceptible is a stock variable with the outflow of Infection Rate and Infected is a stock variable with inflow Infection Rate and outflow of Recovery Rate. The Auxiliary variables are Infectivity, Contact Rate, Total Population, and Average Duration of Infectivity. The mathematical relationship between stocks and flows is based on differential equations. Thus, the Infected level at ant time step is calculated by

$$Infected(t) = Infected(t - dt) + (Infection\ Rate - Recovery\ Rate) \times dt \qquad (3.1)$$

where

$$Infection\,Rate = Susceptible \times Infectivity \times Contact\,Rate \times \frac{Infected}{Total\ Population} \quad (3.2)$$

and

$$Recovery\ Rate = \frac{Infected}{Average\ Duration\ of\ Infectivity}. \quad (3.3)$$

In Figure 3.1, in addition to the causal links represented, the polarity information of such links is also presented. The polarity of a causal link indicates the possible effect of a change in the causing variable on the affected variable. More clearly, a positive sign indicates that an increase (decrease) in the causing variable would result in an increase (decrease) in the affected variable's value as compared to the value that it would take if there were no changes in the cause variable. Conversely, a negative sign indicates that an increase (decrease) in the causing variable would cause a decrease (increase) in the affected variable as compared to the value that it would take otherwise.

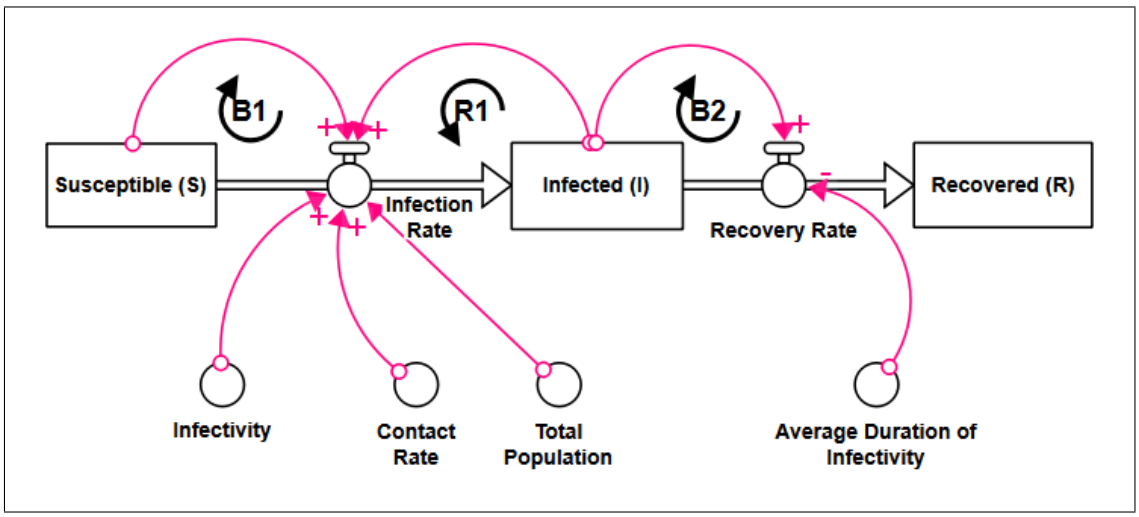


Figure 3.1. Stock-flow Diagram for the simple SIR model with causal loops.

Along with the causal information, incorporating polarities of causal links also reveals the underlying causal loops which can be either balancing or reinforcing. The polarity of a loop is determined by multiplying the polarities of each individual causal link along the loop. If the result of the calculation is 1 then the loop is a "reinforcing" loop or if it is -1 then the loop is "balancing". Identifying such loops in the

causal structure allows us to infer structural reasons for resulting behaviors, as we can define behavioral modes for reinforcing or balancing loops. Reinforcing loops usually generate either increasing growth or collapse, whereas balancing loops often generate goal-seeking behavior.

In Figure 3.1, R1 represents the reinforcing loop created by the Infected and Infection Rate. An increase in the Infected would result in more Infection Rate which also causes an increase in the Infected. Conversely, a decrease in Infected would cause less Infection Rate which results in a decrease in the level of Infected as compared to the value that it would take if there were no decrease in Infection Rate. On the other hand, B1 represents the balancing loop created by Susceptible and Infection Rate. As Susceptible decreases, it causes less Infection Rate which balances the Susceptible in turn.

In addition to the modeling tools mentioned above, the System Dynamics methodology will allow us to simulate the model at hand using differential equations. Such quantitative analysis is not a strict necessity as merely providing a dynamic hypothesis about the problem at hand, and identifying causal structures might also provide useful insights. However, as we applied in our study, using quantitative simulation for scenario and policy analysis is also useful in terms of generating quantitative comparisons and deductions. As a prerequisite of the credibility of such an analysis, both structure and behavior of the model should be tested, required parameters should be estimated using both quantitative and qualitative data in the literature prior to run simulation experiments. After the validity of the model is constructed, then analysis can be continued for the problem and further analysis can be applied to various scenarios and policy interventions.

Following the fundamental steps of System Dynamics methodology suggested by Barlas [39]; thus far we provided basic information about the issue and identified the problem. In the following sections, the formal model and a dynamic hypothesis is provided using differential equations, stock-flow & causal loop diagrams. Moreover, using the data at hand, parameters are estimated and the validity of the model is

discussed. Finally, the base behavior, various scenarios and policy interventions are analyzed, and conclusions are discussed in the last section.

4. OVERVİEW OF THE MODEL

Thus far, we provided a background on both our specific problem and also general structures used to model misinformation. From this point on, we will dive into the model structure that is specifically designed for the problem at hand, namely the spread of the 5G-COVID-19 conspiracy theory in the UK. Starting around February 2020, the misinformation swiftly spread over social media especially in UK. Since the dispute has quickly escalated, caused violent attacks on the 5G towers, and then swiftly fade out, we focus on short-term (6-7 months) dynamics of the propagation. The misinformation is highly rooted in conspiracist thinking and also affected by political polarization as there was a huge dispute about the involvement of the Chinese company in the construction of 5G towers prior to such an infodemic [23].

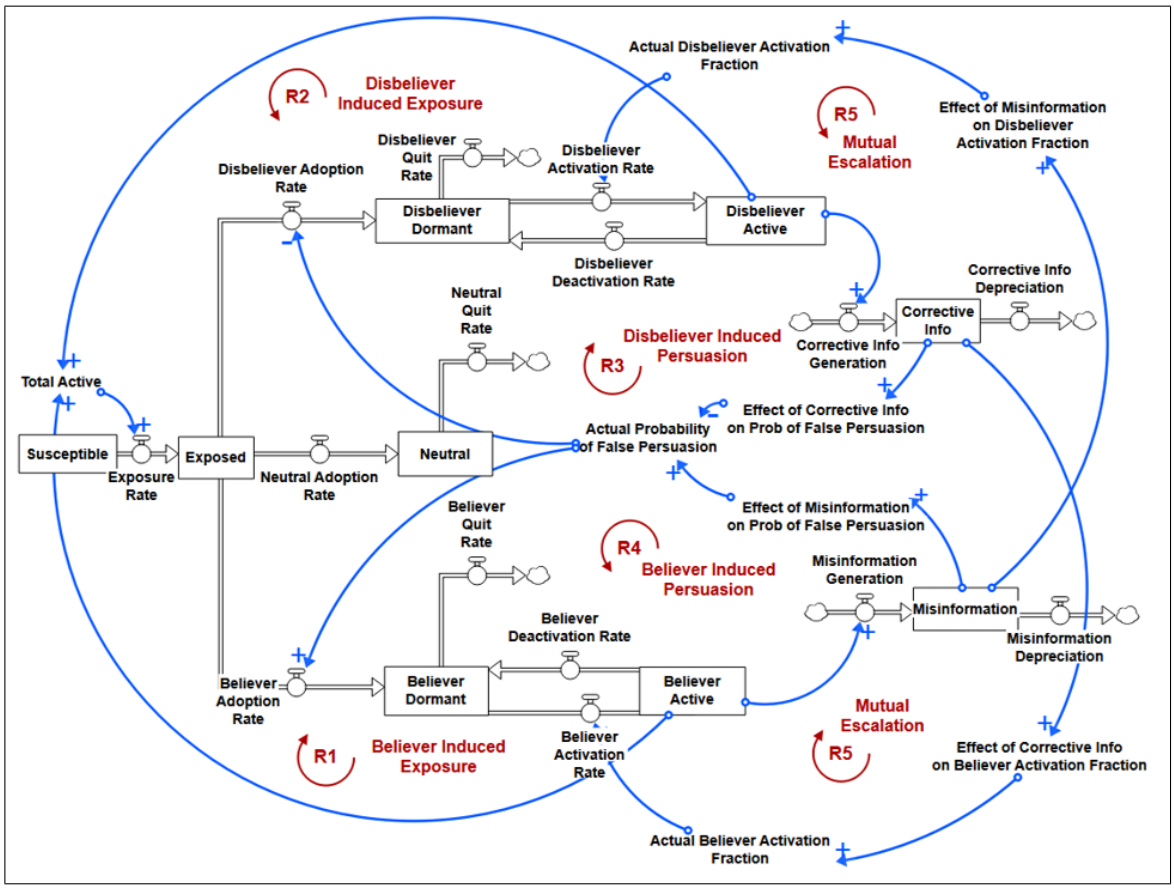


Figure 4.1. The simplified structure of the model and causal loops.

A simplified version of the stock-flow diagram is presented in Figure 4.1. Fundamentally model is an enriched version of the traditional SEIR model of information diffusion. Misinformation spread is often followed with corrective information either initiated by the disbelievers or by different authorities such as fact-checking organizations, scientific institutions, experts, etc. Accordingly, a typical component that is involved in the adaptations of epidemiologic models to misinformation spread is the people who are exposed to misinformation but do not believe it, although the naming (such as skeptics or fact-checkers) and the exact definition differs from one model to another. Similarly in our model, to differentiate between such a difference and reveal the competing dynamics between these groups, "Infected" stocks are separated as "Believers" and "Disbelievers". Believer stocks represent the people who think that the misinformation is true whereas Disbeliever stocks denote the people who are educated enough to know the rumor is false.

A possible interaction between Believer and Disbeliever stocks would be a flow from Believer to Disbeliever as believers can be convinced by the corrective information. However, we focus on short term dynamics of the propagation since the repercussions such as arson attacks are observed in a short period such as 3 to 4 months. Moreover, considering the highly polarized environment around the issue [40], conspiracy nature of the misinformation [23], and the possible resistance to change belief for the cases of conflicting corrective information with one's moral values and emotions [41], we assumed no transition between those two stocks.

Another distinction is made to differentiate between people who are actively spreading their views (either Believer Active or Disbeliever Active) on the issue or remain silent (Believer Dormant, Disbeliever Dormant, or Neutral). Therefore, Believer Dormant stock represents the people who believe the false information but remain silent on the issue whereas Believer Active is the people who believe and actively contribute to the spread of misinformation. Apart from Believer and Disbeliever stocks, people who are exposed can also remain neutral which is denoted as Neutral. The amount of information generated by Believers and Disbelievers is represented in Misinformation and Corrective Info stocks respectively.

The causal structure of the proposed model structure presents five main reinforcing loops:

- Believer - Induced Exposure (R1) & Disbeliever - Induced Exposure (R2): Exposure Rate is affected by amount of active people in the population. Thus, an increase in either Disbeliever Active or Believer Active stocks will result in an increased number of exposed people. Eventually, exposed people would proceed in this stock chain and increase the number of Disbeliever & Believer Active, closing the reinforcing loops.
- Believer - Induced Persuasion (R3) & Disbeliever - Induced Persuasion (R4): A constant fraction of Exposed becomes Neutral. The remaining fraction is split into Believer Dormant with Actual Probability of False Persuasion (p) and Disbeliever Dormant with the complementary probability (1-p). This probability is not constant as it is assumed that the available information would alter such fraction depending on the type of information. Thus, as the amount of Misinformation increases, the Actual Probability of False Persuasion also increases, resulting in more people adopting the false information. In turn, more believers would produce more misinformation closing the vicious cycle. A symmetric causal loop is present for the disbelievers as the increment in the Corrective Info would result in a smaller Actual Probability of False Persuasion thus increasing Disbeliever Adoption Rate.
- Mutual Escalation (R5): A more complex loop emerges from the competing dynamics between the opposing groups. As Misinformation increases, more dormant disbelievers are inclined to share their opinion followed by an increase in the Corrective info. Similarly, an increase in the Corrective info would result in more people becoming active for Believers. Such behavior is reported in experiments conducted on digital social networks [42] and also observed and analyzed by fact-checking organizations for similar cases such as "#NoMasks" movement on Twitter [43].

5. MODEL DESCRIPTION

In this section, exact equations utilized in the model is discussed. Figure 5.1 depicts the stock-flow structure for population stocks. The variables that are affected by Misinformation and Corrective Info are presented in a different color.

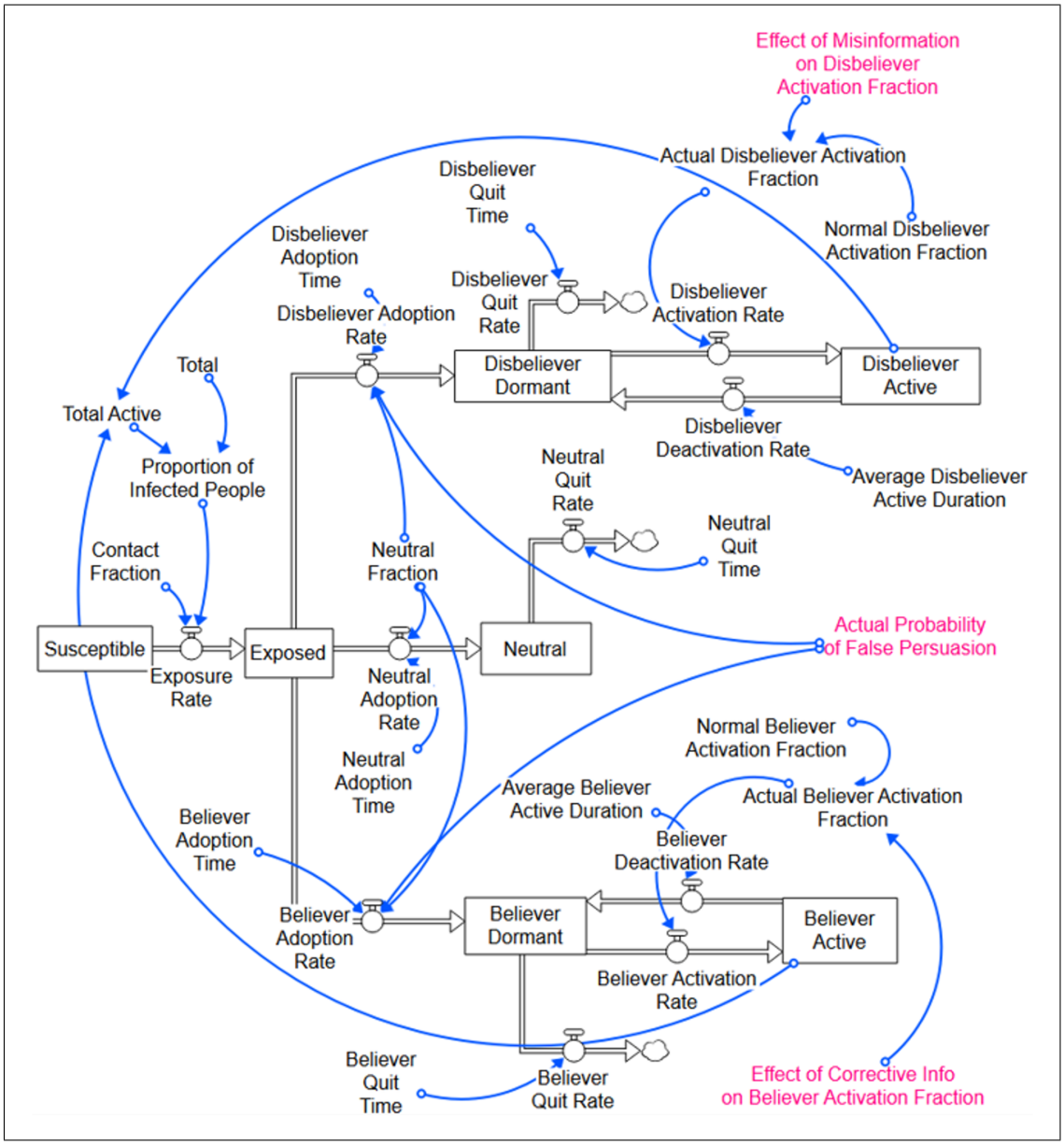


Figure 5.1. Population stocks.

Susceptible people contact other people with the Contact Fraction. The probability of such contact being with an active person is calculated by the ratio of Total Active to the Total number of people. Therefore, the resulting equation for Exposure Rate is:

$$Exposure\ Rate = Susceptible \times Contact\ Fraction \times Proportion\ of\ Infected\ People \tag{5.1}$$

which is analogous to traditional SIR models with "Infected" people including both Believer Active and Disbeliever Active. Thus the Proportion of Infected people is calculated as

$$Proportion\ of\ Infected\ People = \frac{Believer\ Active + Disbeliever\ Active}{Total\ Population}. \tag{5.2}$$

A constant fraction (Neutral Fraction) of Exposed remains neutral after the first exposure whereas the remaining is split between the opposing groups. The distribution is calculated using Actual Probability of False Persuasion. Adoption Rate, Quit Rate, and Deactivation Rate flows are modeled as typical delay formulations:

$$Outflow = Stock/Delay\ Time \tag{5.3}$$

using average times as delays. On the other hand, the Activation Rate flows are formulated as

$$Outflow = Stock \times Fraction \tag{5.4}$$

where Actual Activation Fractions are used as fractions. Actual Activation Fractions are calculated by multiplying graphical effect functions with baseline values.

Figure 5.2 represents the information stocks and related effect formulations. Misinformation Generation and Corrective Info Generation are calculated by the multiplication of active stocks with the Average Information Generation per person. Generated information depreciates with the depreciation delays (see Equation 5.3). All effect functions are standardized using Standard Misinformation per capita & Standard Corrective Info per capita meaning that they use the ratio of information stock per capita value and its standard info per capita value as input.

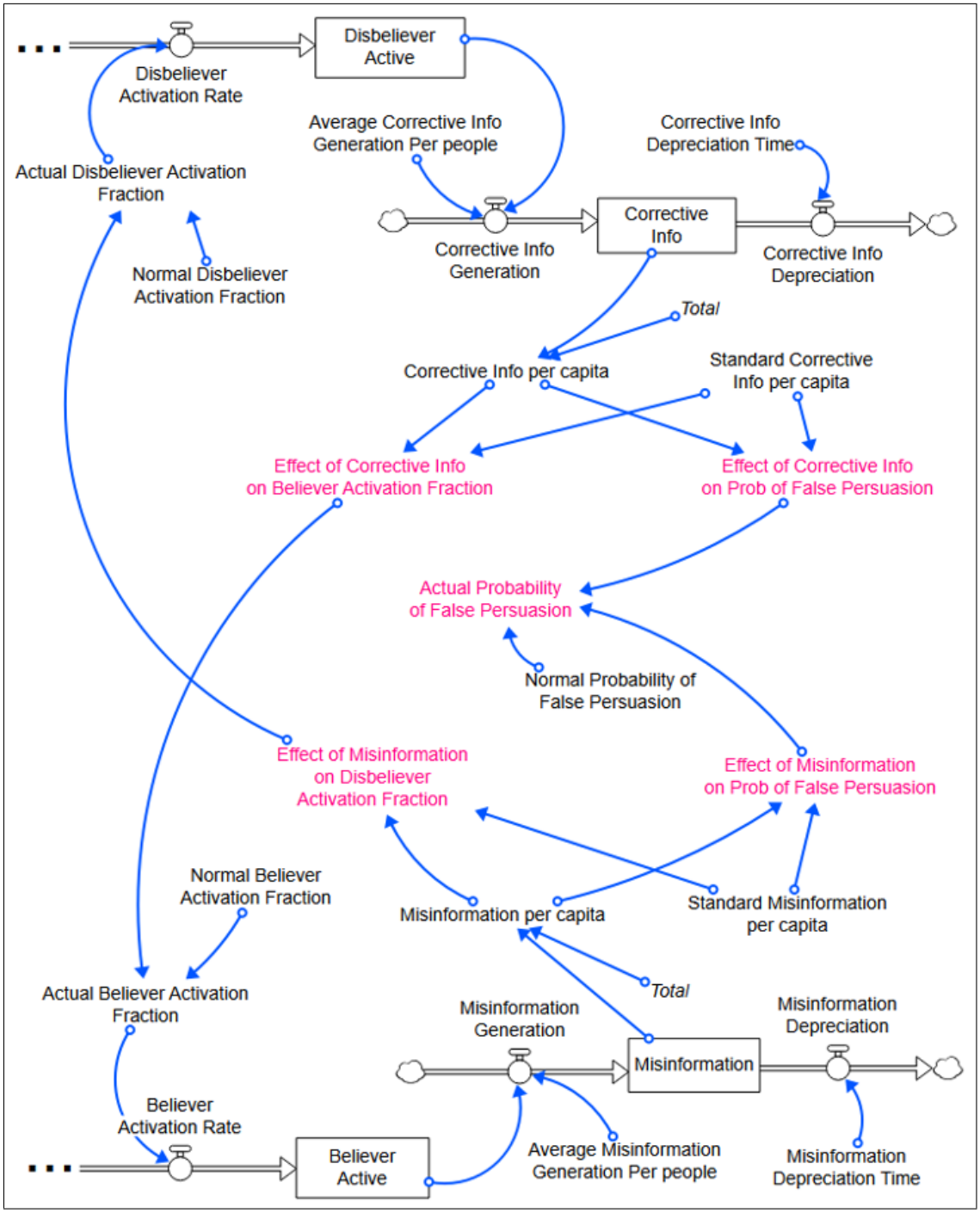


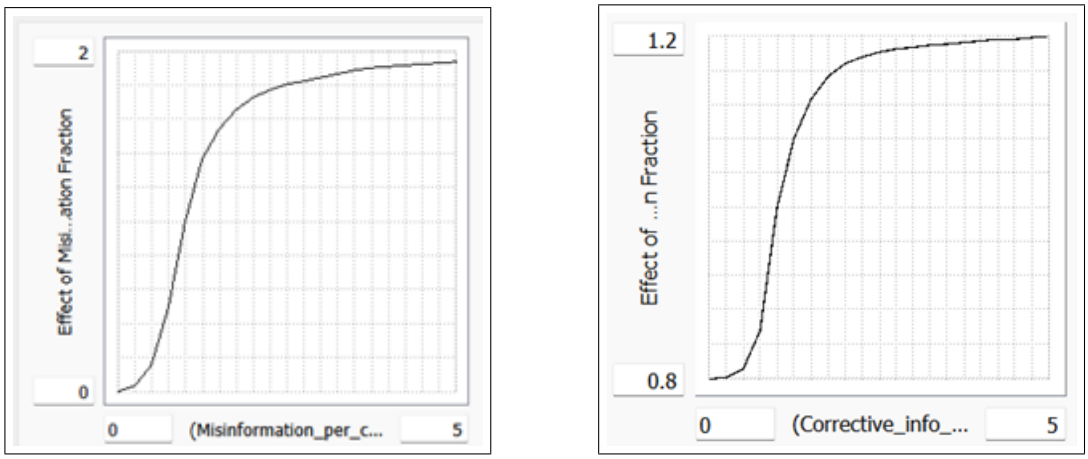
Figure 5.2. Information stocks and effect functions.

Actual values for Activation Fractions are derived by multiplying effects with the normal values of the parameters employing standard multiplicative effect formulation. For example, Actual Disveliever Activation fraction is calculated as

$$\begin{aligned} &Actual\ Disbeliever\ Activation\ Fr. = \\ &Normal\ Disbeliever\ Activation\ Fr. \times Eff.\ of\ Misinformation\ on \\ &\qquad\qquad\qquad\qquad Disbeliever\ Activation\ Fraction \end{aligned} \tag{5.5}$$

where

$$\begin{aligned} &Eff.\ of\ Misinformation\ on\ Disbeliever\ Activation\ Fraction = \\ &\qquad f(\frac{Misinformation\ per\ capita}{Standard\ Misinformation\ per\ capita}). \end{aligned} \tag{5.6}$$



(a) Effect of Misinformation on Disbeliever Activation

(b) Effect of Corrective Info on Believer Activation

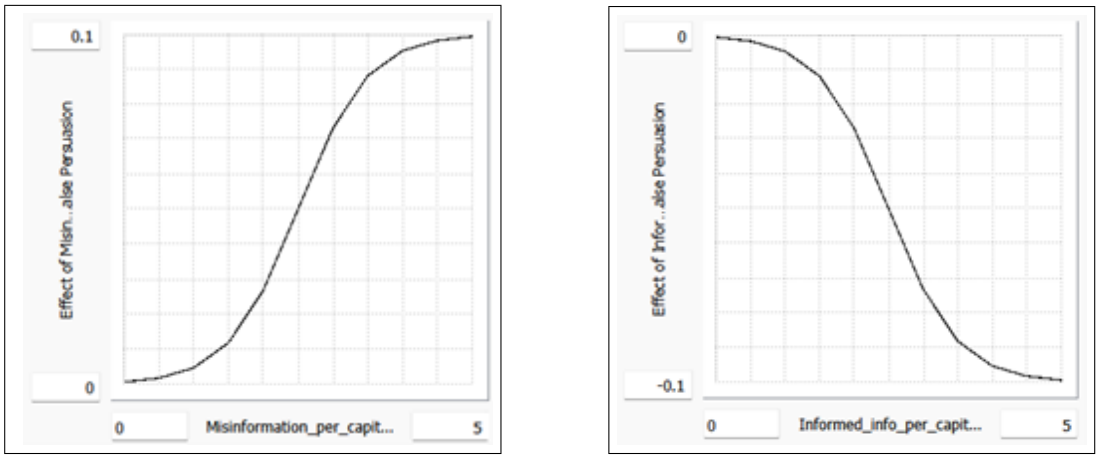
Figure 5.3. Graphical functions of effects of Misinformation and Corrective Information on Activation Fractions.

There are four graphical functions utilized in the model: two effects regarding the Actual Probability of False Persuasion and the other two regarding the activation fractions. Defined graphical effect functions of Effect of Misinformation on Disbeliever Activation Fraction and Effect of Corrective Info on Believer Activation Fraction are presented in Figure 5.3. For both, the input of the graphical function is the ratio of per capita values to their standardized values. A recent study reports that the decrease in perceived peer support increases opinion expression in the digital sphere, WhatsApp groups in particular [42]. Moreover, it is reported that sharing behavior of such misinformation is often affected by identity-based thinking [44]. Additionally

for a disbeliever to become active and start generating corrective information, encountering the misinformation is a necessity. Thus, it is fair to assume that the Effect of Misinformation on Disbeliever Activation Fraction should be an increasing function of Misinformation per capita. Therefore, it is assumed that initially if misinformation is not present, disbelievers should stay in the dormant state whereas as Misinformation per capita reaches a theoretical standard point then the Actual Disbeliever Activation Fraction assumes its normal value (at point (1,1)). As Misinformation per capita passes beyond that standard point; disbelievers are getting more active as they encounter Misinformation more frequently. Such an effect should reach saturation level as the remaining people in the dormant informed stock will be the least motivated ones to speak up. Thus, the increasing function is modeled as having S-shape.

The backfire effect of corrective action on misinformed people is somewhat contradictory. Some studies suggest that individuals might share information as an identity signaling mechanism especially people with conspiracy thinking [45]. Thus, corrective information might be perceived as an attack on self-identity which might increase the activation. However, some other studies discuss that such backfire effects might be evaded if the corrective actions are designed in an effective way (see [5, 46] for a detailed discussion). Nonetheless, for our case, we assume no transition between Believer & Disbeliever Stocks, thus only interested in whether such exposure to corrective information would cause any change in the sharing behavior of current believers. Given our specific case 5G-COVID 19, believers involve mostly conspiracy susceptible people. Consequently, we assumed that social group effects would be more prominent as the literature suggests for conspiracy thinkers [47]. Therefore, the Effect of Corrective Information on Believer Activation Fraction is modeled with the same logic as the misinformation effect, i.e. the effect is a logistic increasing function of Corrective Information per capita, only differing in minimum-maximum values. Considering the contradicting results in the literature, it is assumed that the effect should be less dominant compared to the Effect of Misinformation on Disbeliever Activation Fraction. Furthermore, since the motivation of believers' sharing behavior is not necessarily driven by the existence of Corrective Info the effect function takes a non-zero value when the Corrective Info was zero.

The effects regarding the information effects on the Probability of False Persuasion are formulated using additive graphical effect functions. Thus, the Actual Probability of False Persuasion is calculated as the sum of Normal Probability of False Persuasion, Effect of Misinformation on Prob of False Persuasion, and Effect of Corrective Info on Prob of False Persuasion. The fact that repeated exposure contributes to the belief is a constructed phenomenon [46, 48, 49]. Thus, the Effect of Misinformation on Prob of False Persuasion is an increasing function of Misinformation per capita since we would expect more people to be convinced by the misinformation on average as there is more misinformation available (Figure 5.4). Conversely, if competing Corrective Info is present, the believability of the false information should decrease, hence the corresponding effect is a decreasing function of Corrective Info per capita. The initial limits for these graphical functions are taken as 0.1 and -0.1 as a simplification.



(a) Effect of Misinformation on Prob. of False Persuasion

(b) Effect of Corrective Info on Prob. of False Persuasion

Figure 5.4. Graphical functions of effects of Misinformation and Corrective Information on Actual Probability of False Persuasion.

Finally, some additional clarifications on model assumptions & simplifications:

- Only active stocks (Believer Active and Disbeliever Active) contribute to the Exposure Rate for the base model. Thus, if there are no active people in both groups there wouldn't be any propagation even though there are dormant people.

- False persuasion probability (or believability of misinformation) is assumed to be independent of the exposure pathway (whether exposed by a disbeliever or believer). Thus, instead of using separate stocks for "Exposed by Believer" and "Exposed by Disbeliever", these two stocks are aggregated in the Exposed stock.
- Since the Exposure Rate formulation is based on the contacts between Susceptible and Active Stocks (and not affected by Misinformation or Corrective Information directly) in case of no misinformation, there can still be propagation in this specific social media platform. Therefore the corresponding assumption is that even though we silence all misinformation on a specific platform, there will still be some people that encounter the information from other sources or social media platforms thus still causing a propagation but not contributing to chatter on this platform.
- Normal Probability of False Persuasion is assumed to have some constant value. The actual value dynamically changes depending on existing information ecosystem which implicitly assumes that there should be some fraction of the total population who can either believe or disbelieve based on the availability of the competing information. Thus, the range of Actual Probability of False Persuasion presents an estimate of the fraction of people that can change their minds based on whether there is competing information or not.
- Believer Active Duration and Believer Activation Fraction are assumed to be larger than the Disbeliever Active Duration and Disbeliever Activation Fraction respectively, as people having a conspiracy mindset have a larger tendency to insist on their viewpoint since they are personally involved in the issue.
- Misinformation and Corrective Information have an artificial unit of "information" instead of tweets or posts. The reason is that since they are used in effect formulations, they should represent the effect of that type of information which should depreciate as time progress. Thus, rather than the physical unit of expression of that specific social media platform, these stocks are defined as soft variables and they correspond to sustained effects of their corresponding information type.
- The possibility of transition from believer to disbeliever (or vice versa) is not allowed, as well as the transition from Neutral Stock to any other stocks.

- The total population is assumed to be constant (or changes are negligible) during the simulation horizon.
- All other variables except the ones with effect function formulations (Activation Fractions and Probability of False Persuasion) are assumed to be constant during the simulation horizon in the base model, although the sensitivity results for each one are presented as supplementary material.

6. PARAMETER ESTIMATION AND VALIDITY OF THE MODEL

6.1. Parameter Estimation

The initial parametrization that constitutes "Base Run" is provided in Table 6.1. To deduce the model parameters, various research from the literature is utilized. Based on the believability scores obtained in the study of Agley and Xiao [26], the Normal Probability of False Persuasion is kept at around 0.2 during calibration. Initial stock values of Misinformation and Corrective Info are assumed zero as the spread is assumed to start at time t=0. Since we are mainly concerned with the dynamics of the system rather than acquiring a perfect fit to data, we select the total number of people in the system as 10000- a close number (approx. 9999) of the total estimated by Kauk, Kreysa, and Schweinberger [20]. Moreover, we assume the initial level of 10 for Believer Active to start the propagation, and the initial value of 0 for the other stocks.

The analysis conducted by Ahmed and colleagues [3] revealed that the prevalence of pro-conspiracy (34.8%) and anti-conspiracy (32.2%) tweets about the issue is quite close for the 1 week during the peak time of the debate. Thus, the calibration is made so that the Believer Active should be slightly larger than Disbeliever Active during the peak of the chatter.

To calibrate the remaining parameters two datasets from recent studies [20, 27] are utilized. In their work, Kauk, Kreysa, and Schweinberger [20] use Twitter hashtag data to approximate the level of "Infected" people for the traditional SIR model. The hashtag data consist of the number of daily tweets containing specific hashtags, mostly pro-conspiracy hashtags such as "stop5g", "5gkills", "5gcoranavirus", etc. Due to the volatility of daily hashtag data, the authors use the cumulative version of the data for the parameter calibration. Langguth and colleagues [27] analyze the long-term spread of the 5G-COVID 19 conspiracy with a specific emphasis on the spatial spread of misinformation. In their study, authors collect every tweet containing specific keywords

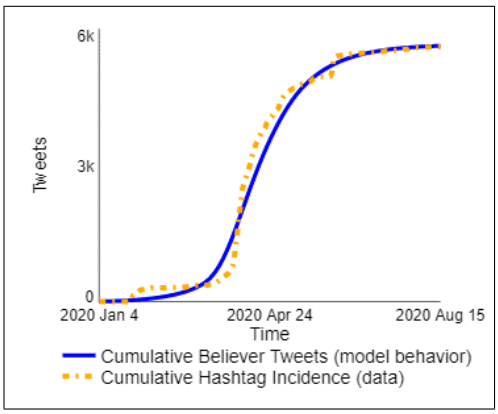
such as “5G”, “5g”, “coronavirus”, “COVID”, etc. for two years and then use manual labeling and machine learning to obtain an estimate of the exact number of tweets that contains misinformation. The authors provide the monthly aggregated version of the posted tweet that contains misinformation and is posted in the UK. Our simulation, however, spans the six months interval with time unit as days. Thus, the data in this study lack the necessary resolution to calibrate the model parameters. Therefore, we use the daily hashtag data from [20] for numerical calibration and normalized monthly posted tweets data from [27] to check the behavioral characteristics.

Table 6.1. Parameter values.

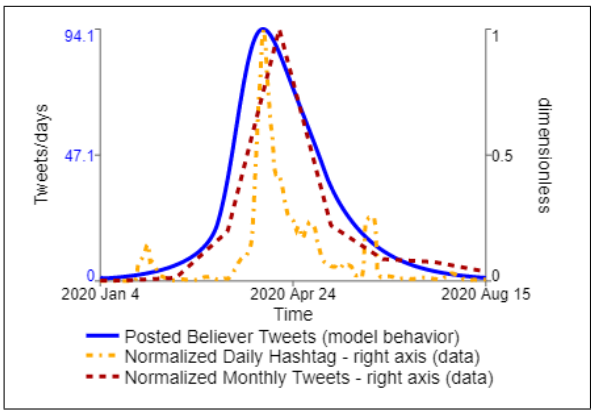
Parameter Name	**Unit**	**Value**
Normal Prob of False Persuasion	-	0.22
Neutral Fract	-	0.1
Contact Fraction	day -1	0.8
Believer Recovery Time	day	9.09
Disbeliever Recovery Time	day	9.09
Neutral Recovery Time	day	9.09
Average Believer Active Duration	day	3
Average Disbeliever Active Duration	day	1
Normal Believer Activation Fraction	day -1	0.68
Normal Disbeliever Activation Fraction	day -1	0.2
Average Corrective Info Generation Per people	information/(day*person)	1
Average Misinformation Generation per people	information/(day*person)	1
Corrective info Depreciation Delay	day	2
Misinformation Depreciation Delay	day	2
Standard corrective info per capita	information/person	0.02
Standard misinformation per capita	information/person	0.02

Parallel to the model fitting process in [20], the cumulative hashtag data is used for numerical calibration. The model at hand does not count the number of tweets explicitly. Therefore, to obtain a proxy for the cumulative incidence of tweets, it is

assumed that an average active believer posts one tweet per week (Average Time to Tweet = 7 days) under these hashtags. Thus, the Cumulative Believer Tweets are defined as a stock with a constant inflow equal to Believer Active /Average Time to Tweet. Although a tweet per week might seem like a small frequency, since the data at hand is the number of tweets under specific hashtags rather than all tweets about the issue, it is adequate to assume a longer period for Average Time to Tweet to avoid underestimating the number of believers.



(a) Cumulative Hashtag Incidence (data from [20])



(b) Posted Tweets per time (data from [20, 27])

Figure 6.1. Model behavior after calibration vs. various data sources: (a) Cumulative Hashtag Incidence, (b) Posted Tweets per time.

Figure 6.1. depicts the comparative graphs of data provided in the literature (dash-dot yellow lines) and the model behavior (solid blue lines). As provided in Figure 6.1a., the resulting Cumulative Believer Tweets provide a good fit with the cumulative hashtag incidence data from [20]. Posted Believer Tweets peak around the 8th of April, 2020 which coincides with the peak time observed in daily hashtag data (Figure 6.1b). Moreover, the Believer Active to Disbeliever Active ratio is close to 1 during the peak (Figure 6.2) which is in line with the findings provided by Ahmed and colleagues [3]. Although the current set of parameters does not explain the reoccurring peaks in hashtag data (Figure 6.1b), the base model is built with this parameter setting and whether such differences in the behavior can be obtained by further expanding the dynamic hypothesis or changing the parameter settings will be evaluated later.

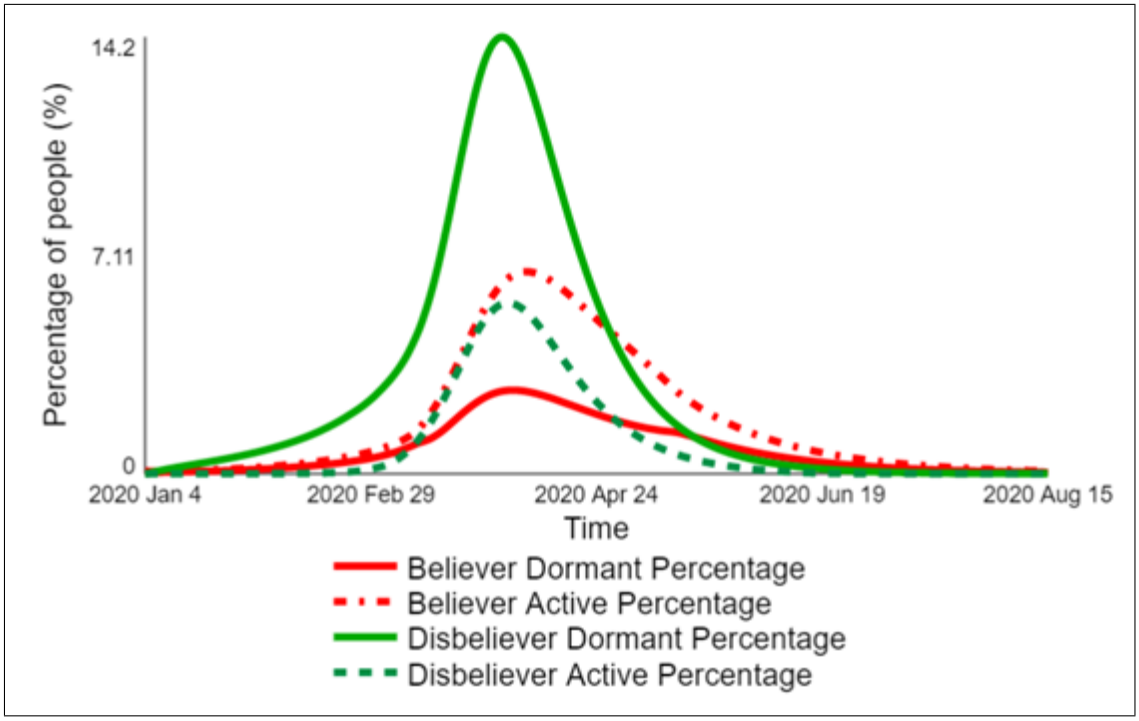


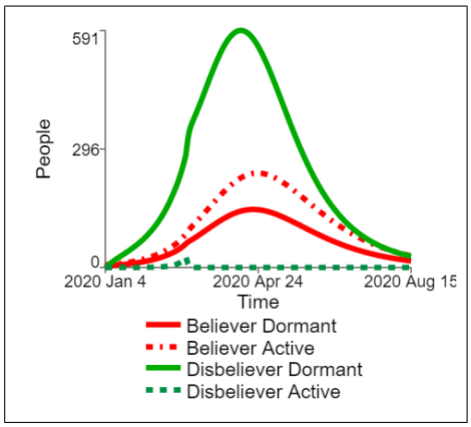
Figure 6.2. Dormant & Active Stocks after calibration.

6.2. Structural Credibility

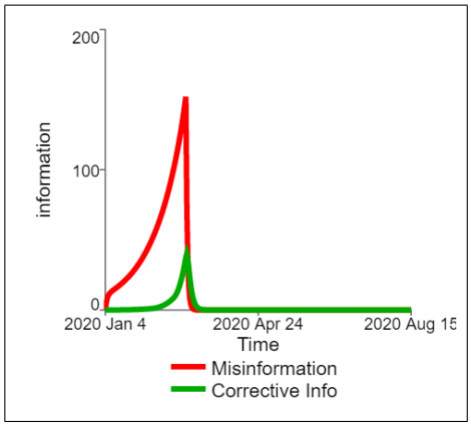
Structural validity is an essential part of the modeling process, especially for the causally descriptive simulation models. To reveal the underlying structural reasons causing the problem at hand or to provide policy analysis with real-life implications, the foremost condition is that the model equations and other causal relations embedded in the model should be consistent with the available knowledge about the real system [50]. Regarding this, the model structure is built upon the existing qualitative and quantitative literature, utilized mathematical equations are criticized with direct structure tests, and parameter consistency is ensured along the model building process. All model parameters have meaningful real-life counterparts, and assumptions or simplifications are clarified in the "Model Description" section.

6.3. Indirect Structure Tests

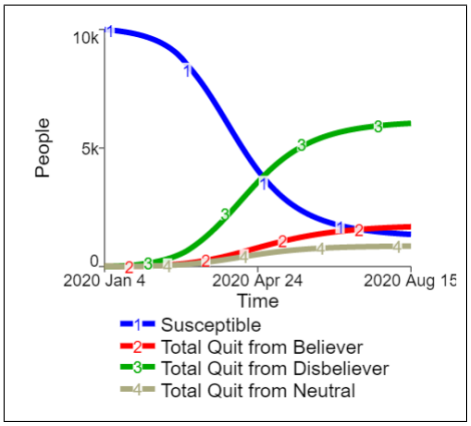
To further solidify the model validity, we applied several extreme condition tests as proposed by Barlas [50]. These tests aim to ensure that the simulation model would produce valid behaviors under extreme conditions as the real system would produce. Some examples of such conditions are having no initial active people, a very small probability of false persuasion, a very high probability of false persuasion, a sudden decrease in the Misinformation, or a sudden decrease in the Believer Active.



(a) Contact Fraction

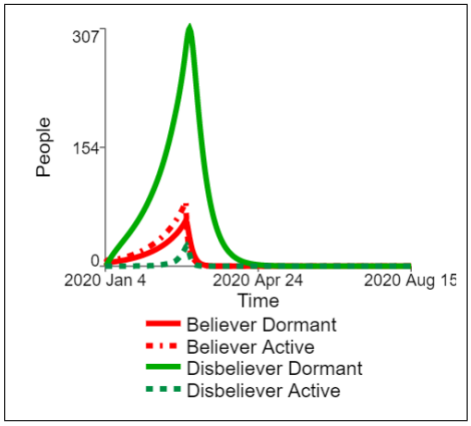


(b) Normal Believer Activation Fraction

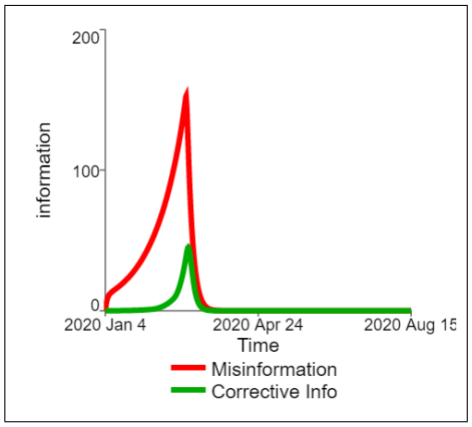


(c) Average Believer Active Duration

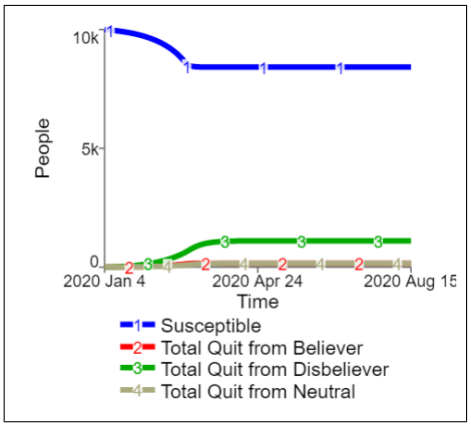
Figure 6.3. Model behavior for the extreme condition of eliminating Misinformation: (a) Dormant & Active Stocks, (b) Information Stocks, and (c) Susceptible and Total Quit Flows.



(a) Contact Fraction



(b) Normal Believer Activation Fraction



(c) Average Believer Active Duration

Figure 6.4. Model behavior for the extreme condition of eliminating Believer Active: (a) Dormant & Active Stocks, (b) Information Stocks, and (c) Susceptible and Total Quit Flows.

The resulting behaviors for the last two cases are presented in Figure 6.3 and Figure 6.4 whereas the results of the others are provided in Appendix B. If the believability of misinformation is too low or if there are no people knowing that misinformation, one should expect no spread of such information. Conversely, if the believability of misinformation is too high, we would expect that misinformation spreads to most of the population without many disbelievers. To test the model behavior under these extreme conditions, three simulation experiments are conducted with three different parameter settings: Normal Probability of False Persuasion = 0.001 (low believability), Normal Probability of False Persuasion = 0.999 (high believability), Initial Active = 0 (no people knowing misinformation). The resulting behavior of the model (see

Appendix B) is consistent with the real-life expectation for the aforementioned three cases.

The results of the other two tests are presented in Figures 6.3 and Figure 6.4. To test the behavior of the model, external outflows are activated on day 60 and implemented to Misinformation (Figure 6.3) and Believer Active (Figure 6.4) stocks. Elimination of Misinformation after day 60 (Figure 6.3b) results in no dispute on the subject for this platform whereas the propagation still exists (Figure 6.3a) with a lesser impact as the active people continue to spread the misinformation outside this social media platform. Elimination of Believer Active on day 60 (Figure 6.4) on the other hand, stops the propagation as observed in the stabilized levels in Susceptible (Figure 6.4c), thus also ending the dispute in this specific social media platform (Figure 6.4b).

7. BASE RUN AND SENSITIVITY ANALYSIS

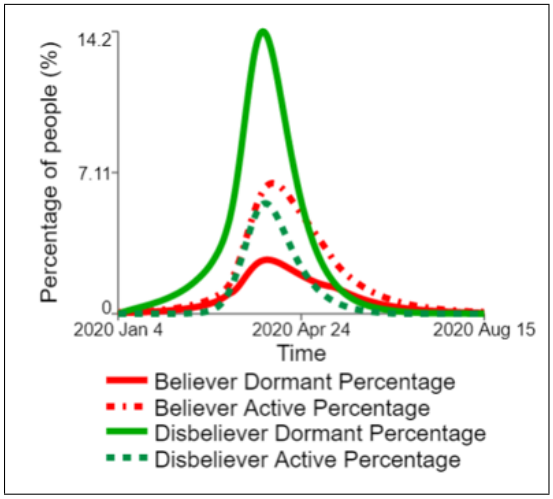
7.1. Base Run

To compare the effectiveness of different runs, three outcomes of interest are defined: Exposure Percentage, Total Believer Peak Percentage, and Believer Prevalence Percentage. Exposure Percentage denotes the percentage of the total population that have been exposed to this information, Total Believer Peak Percentage is the maximum number of total believers (both Dormant and Active) as a percentage of the whole population, and Believer Prevalence Percentage (or Believer Prevalence) is the percentage of the total population that has been a believer at least once in simulation horizon. For all three of them, we consider the final values at the end of the simulation run. The motivation behind defining different measures is that based on the nature of the misinformation, policymakers might target different outcomes. For instance, policymakers might try to minimize Total Believer Peak Percentage while countering misinformation that is expected to create violent responses in the audiences as it was in our specific case and numerous other ones [9,51]. On the other hand, considering the stickiness of misinformation and its sustained effects even after the debunking [15, 46], policymakers might prefer minimizing Believer Prevalence if the misinformation affects the long-term public health behavior, for instance.

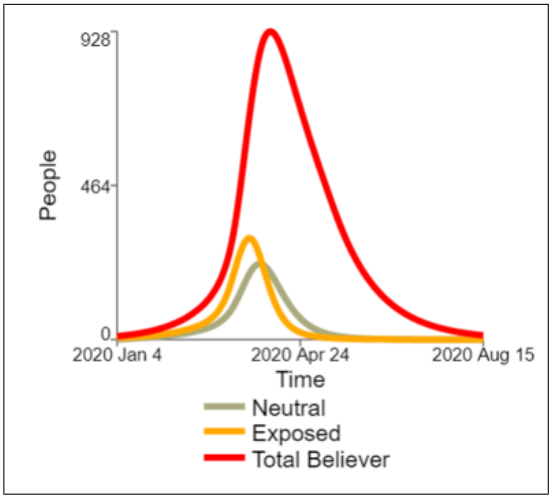
Table 7.1. Resulting outcomes of interest for the Base Run.

Exposure Percentage	99.22
Total Believer Peak Percentage	9.271
Believer Prevalence Percentage	20.13

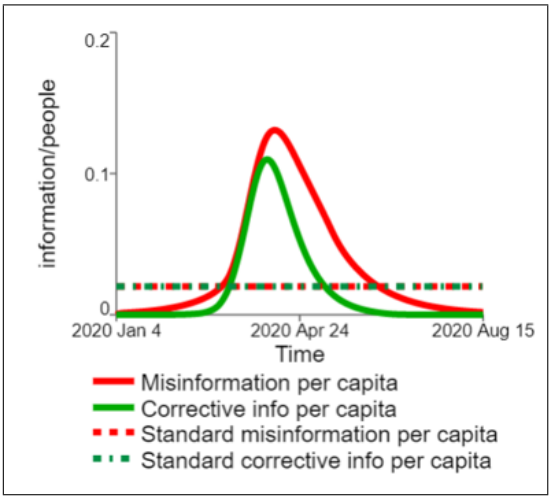
The base run represents 225 days (7.5 months) from the 4th of Jan to the 15th of Aug where the debate is prominent on Twitter as hashtags. The dynamics of the main stocks and variables are presented in Figure 7.1 and resulting outcome of interests are presented in Table 7.1.



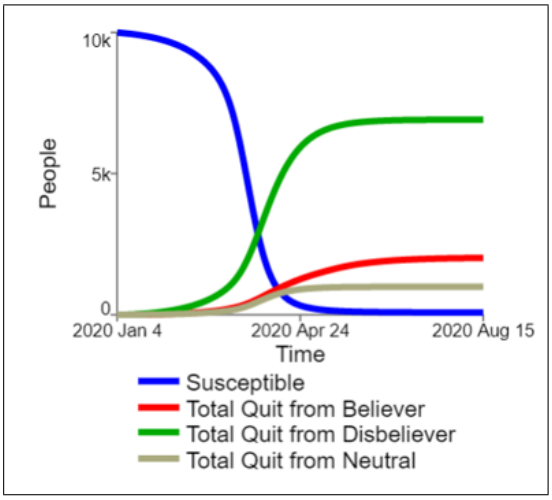
(a) Dormant & Active Stocks



(b) Neutral, Exposed, and Total Believer



(c) Information Stocks



(d) Susceptible and Cumulative Quit Flows

Figure 7.1. Resulting stock dynamics for the Base Run: (a) Dormant & Active Stocks, (b) Neutral, Exposed, and Total Believer, (c) Information Stocks, and (d) Susceptible and Cumulative Quit Flows.

The behavior of the four people stocks Figure 7.1a is similar, as the peak times and the shapes are nearly the same for all 4 stocks. Misinformation seems to exceed the Corrective Info for the whole period and nearly all of the Susceptible is depleted with Exposure Percentage equal to 99.22 %. Considering our assumption that Susceptible represents the people on the social media platform that has the potential to participate in the discussion, we can say that the maximum potential is reached for this case. It seems intuitively consistent, as the 5G narrative is one of the most viral in-

stances of misinformation involving distribution channels such as national TV, celebrity super-spreaders, and conspiracy theorists with preexisting social connections [23] thus resulting in a wider reach to various audiences.

The initial observation in Figure 7.1a is that although the magnitude difference between the Believer Dormant and Disbeliever Dormant is huge, the Active Stock levels are quite close for the two groups. Therefore, one simple insight is even when the pro-conspiracy people in the population are in minority, their presence in the digital sphere (i.e. Active Stocks) can dominate the educated people, as believers are more inclined to engage in social media. It should be noted that such an insight is provided by the enriched model whereas the traditional SIR models lack the necessary resolution for such an analysis.

7.2. Sensitivity of Model Behavior to Changes in the Model Parameters

Sensitivity analysis is conducted to analyze the model behavior for changes in model parameters. To avoid combinatorial complexity, the analysis is conducted one parameter at a time. Table 7.2 presents the selected parameters for sensitivity analysis and the minimum and maximum values included in the analysis. Full results of the sensitivity analysis are presented as a supplementary material whereas important runs are discussed in this section. The sensitivity runs are taken for 300 days (instead of 225 days in the Base Run) to be able to observe the long-term dynamics that might be caused by the changes in some of the parameters such as Contact Fraction.

Overall, the model shows consistent behavior in terms of model assumptions and real-life expectations. We observe no propagation of misinformation for the extreme parameter values such as Normal Believer Activation Fraction = 0 or Contact Fraction = 0.1. Conversely, we observe higher peak values and prevalence for higher levels of Normal Believer Active Duration and Normal Believer Activation Fraction, which is within our expectations.

Table 7.2. Selected Parameters for sensitivity analysis, base run values, analysis intervals, and increments.

Parameter Name	Value	Min-Max	Increment
Normal Prob of False Persuasion	0.22	0 - 0.5	0.1
Contact Fraction	0.8	0.1 - 0.9	0.1
Average Believer Active Duration	3	1 - 4	0.5
Average Disbeliever Active Duration	1	0.5 - 3	0.5
Normal Believer Activation Fraction	0.68	0 - 1	0.2
Normal Disbeliever Activation Fraction	0.2	0 - 0.5	0.1
Corrective info Depreciation Delay	2	1 - 3	0.5
Misinformation Depreciation Delay	2	1 - 3	0.5
Standard corrective info per capita	0.02	0.01 - 0.05	0.01
Standard misinformation per capita	0.02	0.01 - 0.05	0.01

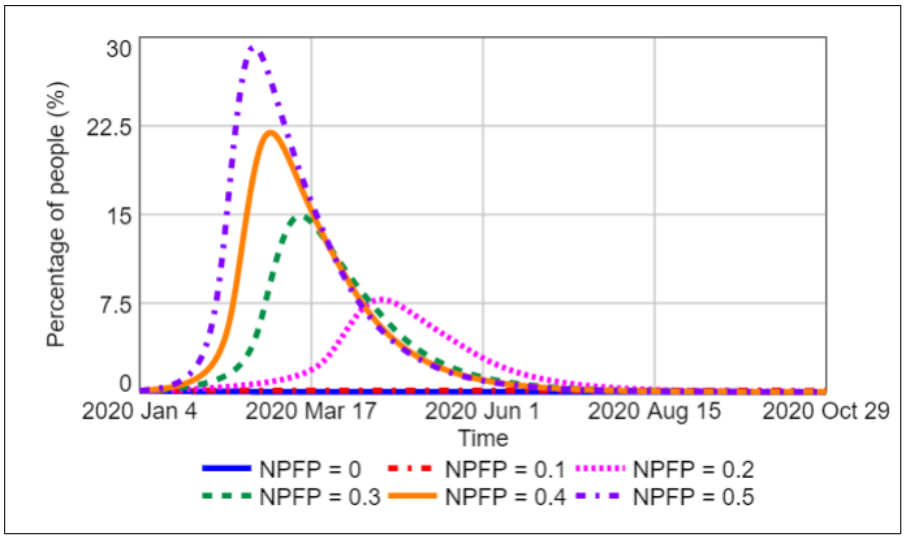


Figure 7.2. Sensitivity of Total Believer to Normal Probability of False Persuasion (NPFP).

The most influential variable in terms of changes in the outcome of interests is the Normal Probability of False Persuasion. The resulting behavior and outcome measures are presented in Figure 7.2 and Table 7.3 respectively. Since the parameter itself repre-

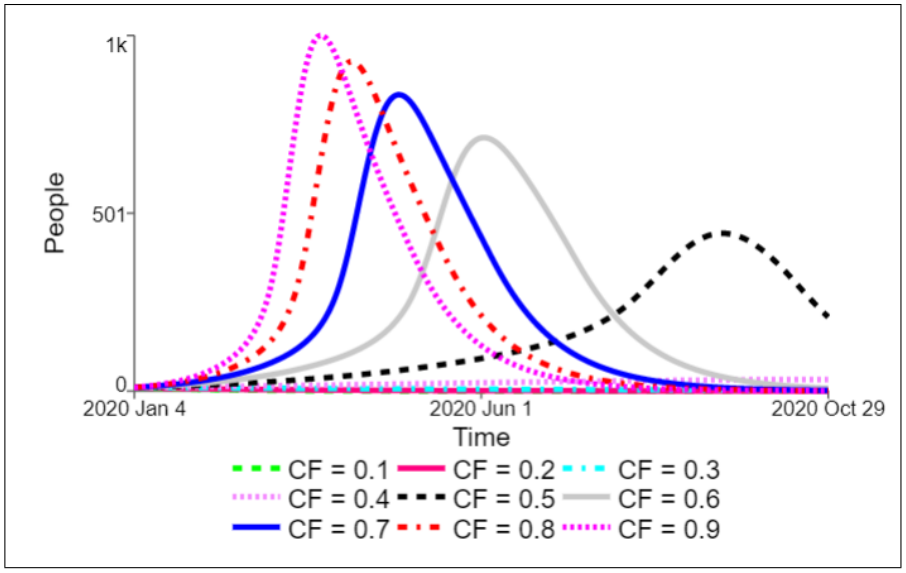
sents the believability of the misinformation, the model behavior is consistent with the expectation that the more believable the false information is the more people would fall for it. Another factor to notice is that, considering the changes in Exposure Percentage, there is a threshold value for Normal Probability of False Persuasion around 0.1 and 0.2 below which the epidemic does not start which is analogous to epidemiological models. If the believability of the information is above that threshold, Believer Prevalence Percentage and Total Believer Peak Percentage seem to linearly increase with the increasing probability of believing.

Table 7.3. Outcome of interests for different values of Normal Probability of False Persuasion.

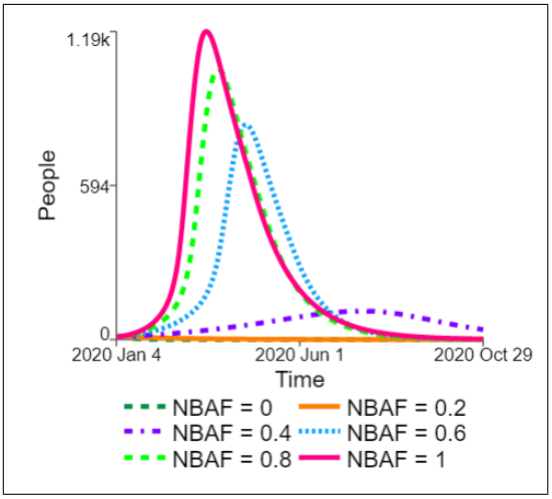
	Exposure Percentage	**Total Believer Peak Percentage**	**Believer Prevalence Percentage**
NPFP = 0:	1.52	0.1	0.1
NPFP = 0.1:	12.58	0.12	1.11
NPFP = 0.2:	98.85	7.8	17.98
NPFP = 0.3:	99.79	14.88	28.39
NPFP = 0.4:	99.95	21.92	38.6
NPFP = 0.5:	99.99	29.16	48.92

The Normal Probability of False Persuasion is not the only parameter with a threshold property. Figure 7.3 depicts the Total Believer levels for different values of Contact Fraction, Normal Believer Activation Fraction, and Average Believer Active Duration. We observe none to minuscule changes in Believer Active levels for Contact Fraction below 0.3 (Figure 7.3a), Normal Believer Activation Fraction (Figure 7.3b) below 0.2, and Average Believer Active Duration below 1 (Figure 7.3c). These four parameters are the only parameters that result in such behavior in the scope of this sensitivity analysis (see Table 7.2). Normal Probability of False Persuasion and Contact Fraction are related to the different characteristics of the misinformation and the structural properties in the social network of susceptible people respectively. On the other hand, Normal Believer Activation Fraction and Average Believer Active Duration

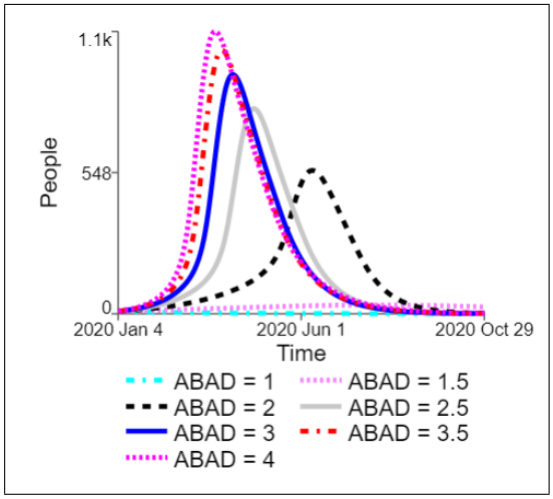
are related to the behavioral responses of people who believe misinformation. Thus, none of these parameters propose a leverage point that can be easily targeted by an intervention to eradicate the spread of false information. Interestingly, none of the changes in Disbeliever parameters can prevent the full spread in the current parameter settings although they are somewhat effective in restricting the spread which imply that those parameters might be promising to develop efficient mitigation strategies.



(a) Contact Fraction



(b) Normal Believer Activation Fraction



(c) Average Believer Active Duration

Figure 7.3. Sensitivity of Total Believer to (a) Contact Fraction, (b) Normal Believer Activation Fraction, and (c) Average Believer Active Duration.

Another observation is that changes in Believer parameters typically create a unidirectional change in the outcomes whereas responses to changes in Disbeliever parameters often have nonlinear outcomes. For example, increasing Believer Activation Fraction or Believer Active Duration always results in worse outcomes as observed by increased Believer Prevalence Percentage and Total Believer Peak Percentage (Figure 7.3b and c). However, increasing Disbeliever Activation Fraction (Figure 9.1 and Table 9.1) produces changed outcomes based on the trade-offs created by Disbeliever Induced Exposure (R2), Disbeliever Induced Persuasion (R3), and Mutual Escalation (R5) which will be further analyzed in policy analysis.

8. SCENARIO ANALYSIS

Thus far, the model structure and its credibility is discussed and the base behavior of the model has been analyzed. In this section we aim to expand our analysis by investigating what-if scenarios and analyzing changes in the outcome measures for these scenarios. The first two scenarios, "Neutral Sharing" and "Super-spreader", involve slight additions to the model structure which are explained in the corresponding section. The following scenario is obtained by slight changes in the base model parameters or by combining those changes with "Neutral Sharing" and "Super-spreader". Finally the last one investigates the model behavior under having a constant increase in Susceptible.

8.1. Base Run with Neutral Sharing

There is a growing body of evidence that people can share misinformation without essentially considering its veracity. Pennycook and colleagues [2] found that people's ability to distinguish between true and false information is significantly low when they are asked whether they would share that information compared to the case when they are asked about the veracity of the information. Thus, sharing behavior does not necessarily require belief in the misinformation but can have many other motives or reasons such as intuitive thinking, self-promotion, and even the thought that misinformation being true would be "interesting" [46]. Moreover, for our specific case 5G – COVID 19 conspiracy, there is also supporting evidence that financially motivated agents contribute to the spread [27]. Clearly, each of these types of sharing behaviors has different internal drivers and can be incorporated separately into the current model. However, such an investigation is beyond the scope of this work. Instead, we simplify the possible other reasons and just investigate whether the involvement of Neutral stock both in the exposure and misinformation generation will create a substantial behavior change in the current model.

To incorporate such a scenario into the model we define Neutral Engagement Fraction as the fraction of Neutral that generates Misinformation and also contributes to its spread. Thus, multiplication of Neutral and Neutral Engagement Fraction represents the number of people that acts as Believer Active even though they do not necessarily believe the misinformation. We denote this amount as Engaged Neutral. To generate such a behavior Engaged Neutral is added in Total Active for exposure and Misinformation Generation flow. A subsequent simplification of such formulation is contribution of an Engaged Neutral person to Misinformation spread is equal to the contribution of a Believer Active person.

Table 8.1. Comparative table of output measures for different values of Neutral Engagement Fraction.

	Exposure Percentage	**Total Believer Peak Percentage**	**Believer Prevalence Percentage**
Base Run:	99.22	9.27	20.13
NEF = 0.2:	99.4	9.88	20.35
NEF = 0.4:	99.52	10.42	20.55
NEF = 0.6:	99.62	10.91	20.73
NEF = 0.8:	99.7	11.35	20.91
NEF = 1:	99.75	11.76	21.08

Figure 8.1 depicts the comparative plot of Total Believer for different values of Neutral Engagement Fraction and resulting output measures are presented in Table 8.1. The main impact of increased Neutral Engagement Fraction is the earlier peak observed in the Total Believers. Naturally, sharing behavior of neutrals increase Exposure Rate as Susceptible people can also get exposed by getting in contact with random interactions with neutral people. The effect on Total Believer Peak Percentage seems more prominent compared to Believer Prevalence Percentage as for the former peak percentage has increased from 9.27% to 11.76% whereas for the latter the increase in believer prevalence (from 20.13% to 21.08%) is somehow limited.

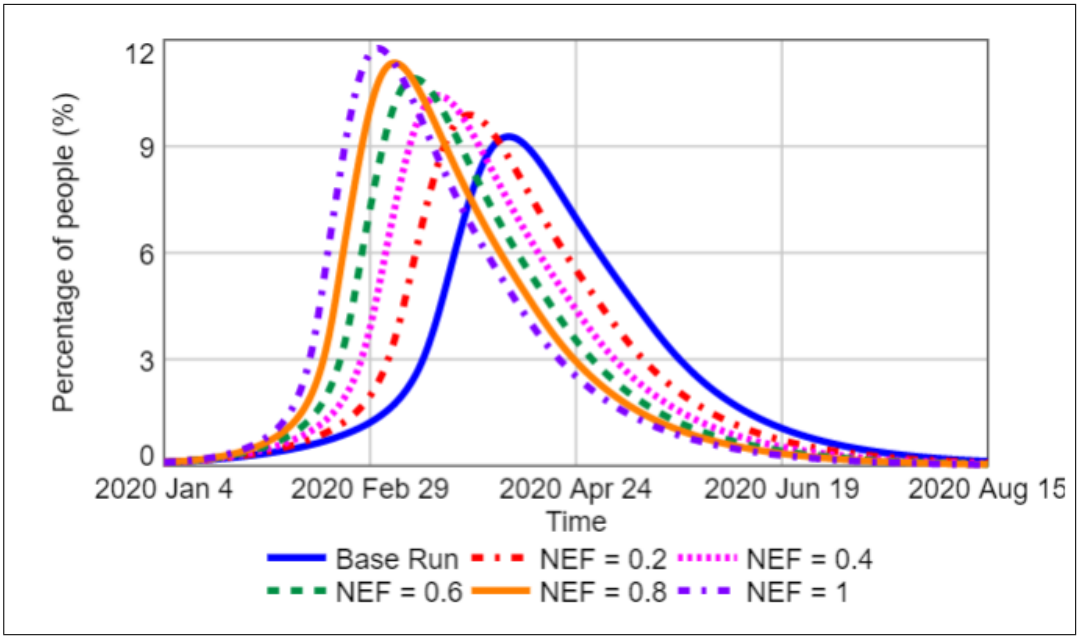


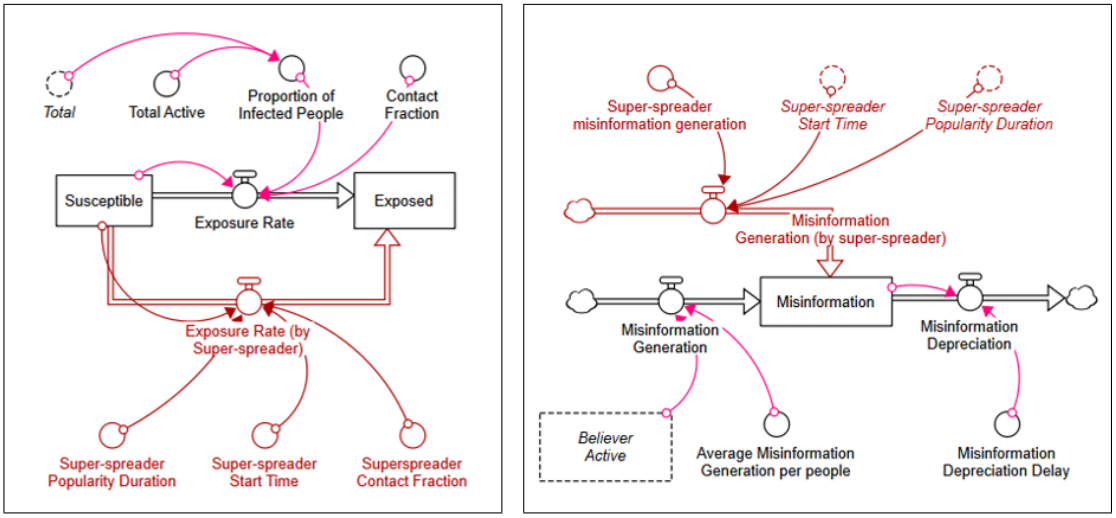
Figure 8.1. The comparative plot of Total Believer Percentage for different values of Neutral Engagement Fraction.

8.2. Base Run with Super-spreader

Bruns, Harrington, and Hurcombe [23] have analyzed how the 5G-COVID 19 conspiracy has spread in digital spaces by analyzing Facebook posts. One specific phase of propagation in their analysis includes the spread of conspiracy through celebrities. The amplitude of the exposure through these celebrity accounts reaches enormous numbers on the scale of millions. Rooting from this idea, we wanted to examine the impact of having such a "super-spreader" on the propagation dynamics and compare the measures with the base run.

Table 8.2. Parameters used for Super-spreader scenarios.

Parameter Name	**Unit**	**Value**
Super-spreader Start Time (SST)	day	10 - 130
Super-spreader Contact Fraction	day -1	0.01
Super-spreader Popularity Duration	day	3
Super-spreader Misinformation Generation	information/day	400



(a) Exposure Rate (b) Misinformation Generation

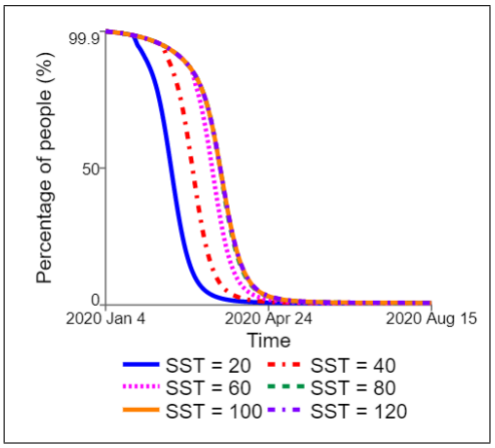
Figure 8.2. Modified Stock-Flow diagrams after incorporating super-spreader: (a) Exposure Rate, (b) Misinformation Generation.

The resulting dynamics of four stock variables are presented in Figure 8.3 and the outcome measures for the same run are provided in Table 8.3. The immediate effect of the super-spreader is the mobilization of people from Susceptible to Exposed (Figure 8.3a and b). When Super-spreader Start Time is larger than or equal to 100, there is almost no effect in terms of exposure since the Susceptible has been deployed by then. The effect on Misinformation is still present after 100 as can be observed by the sharp peaks in Figure 8.3.c. However, since the upstream stocks are already deployed, i.e. misinformation has already gone viral, having a super-spreader does not change the outcome of interests.

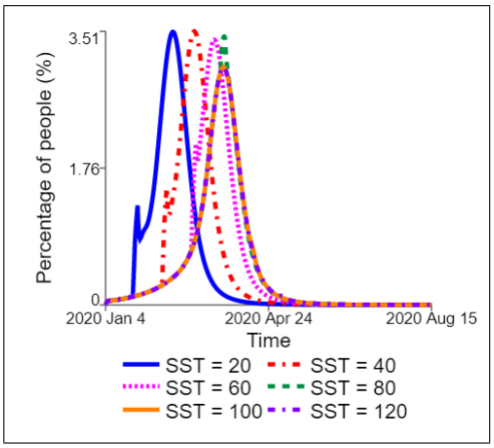
Considering Table 8.3, we can say that the earlier the super-spreader acts, the higher the damage it inflicts in terms of Total Believer Peak Percentage. If the super-spreader acts after misinformation have already spread, then the damage is nearly zero since outcomes after start time 100 are nearly the same as their base run values for all three outcomes of interest. Parallel to the observation made for the Neutral Sharing scenario, the adverse effect of having a super-spreader is relatively higher for Total Believer Peak Percentage as compared to the Believer Prevalence Percentage.

Table 8.3. Comparative tables of the outcome measures for different start times of super-spreader.

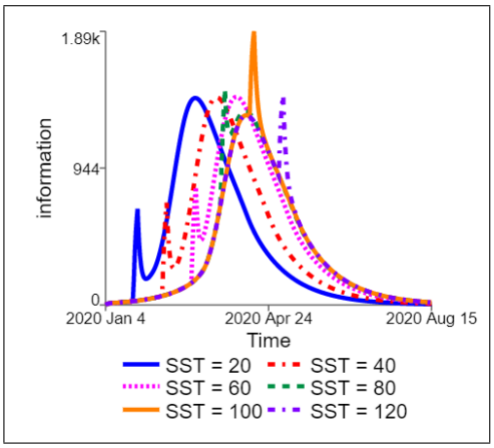
	Exposure Percentage	**Total Believer Peak Percentage**	**Believer Prevalence Percentage**
SST = 20:	99.4	10.11	20.28
SST = 40:	99.4	10.14	20.35
SST = 60:	99.4	10.16	20.92
SST = 80:	99.3	9.46	20.24
SST = 100:	99.2	9.27	20.14
SST = 120:	99.2	9.27	20.15



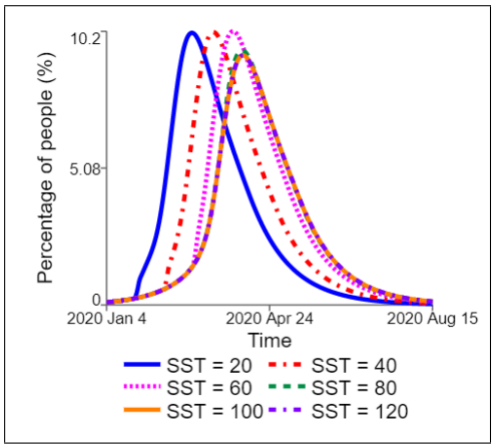
(a) Susceptible



(b) Exposed



(c) Misinformation



(d) Total Believer

Figure 8.3. Comparative plots of stocks for different start times of super-spreader: (a) Susceptible, (b) Exposed, (c) Misinformation, and (d) Total Believer.

8.3. Low Believability with Neutral Sharing and Super-spreader

As discussed in Section 7.2 for sufficiently low levels of Normal Probability of False Persuasion, we observe almost no spread of misinformation. Thus, this scenario aims to analyze whether the involvement of super-spreaders or neutral sharing can trigger the epidemic in such cases. To analyze such a scenario, we decrease the Normal Probability of False Persuasion from 0.22 to 0.1 and generate the base scenario for the "Low Believability" case. Then, we add super-spreader and normal sharing cases separately. For the Neutral Engagement Fraction (NEF), we experiment with values 0.5 and 0.9. For super-spreader cases, we try four scenarios by changing the Super-spreader Start Time (SST) (50 - 100) and changing the effectiveness of the super-spreader (High, Low).

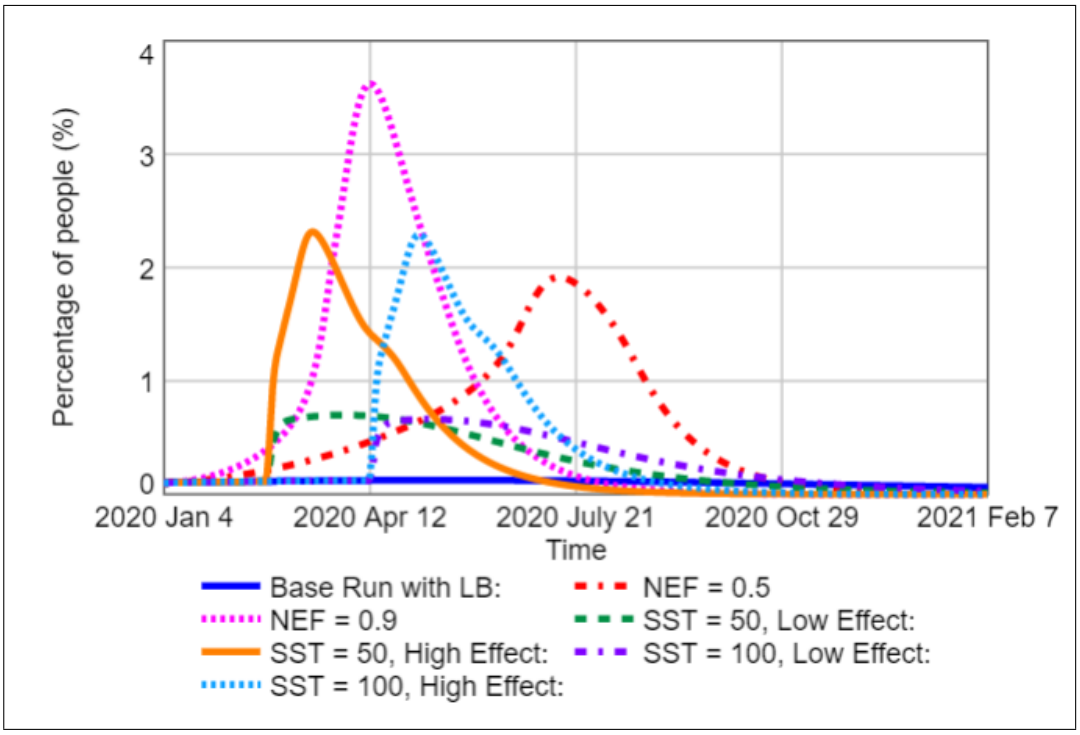


Figure 8.4. The comparative plot of Total Believer Percentage for low believability with neutral sharing and super-spreader scenarios.

The resulting dynamics of Total Believer for different cases are presented in Figure 8.4 and the corresponding outcomes of interest tables are presented in Table 8.4.

Considering the "Base Run with Low Believability", the misinformation spread is quite contained with Believer Prevalence Percentage at 1.75% and Exposure Rate is less than 20%. Total Believer is nearly constant during the simulation horizon and the Total Believer Peak Percentage is 0.12%.

The impact of Neutral Engagement is huge even when the engagement is small as 0.5. Considering Table 8.4, when NEF = 0.5, The Exposure Percentage jumps from 19.16% to 91.38% compared to the baseline. Therefore, we can say that the involvement of agents other than Believers or Disbelievers could alter the system behavior and help the epidemic to pass the threshold value. Comparing the two runs with varying NEF values we can say that difference in believer prevalence and exposure is small as compared to the increase in the peak percentage.

Table 8.4. Comparative table of the outcomes of interest for low believability with neutral sharing and super-spreader scenarios.

	Exposure Percentage	**Total Believer Peak Percentage**	**Believer Prevalence Percentage**
Base Run with LB:	19.16	0.12	1.75
NEF = 0.5:	91.38	1.91	8
NEF = 0.9:	97.63	3.62	8.36
SST = 50, Low Effect:	47.08	0.7	4.49
SST = 50, High Effect:	88.85	2.31	5.23
SST = 100, Low Effect:	45.61	0.66	4.34
SST = 100, High Effect:	88.7	2.29	5.59

Regarding super-spreader scenarios, looking at the Exposure Percentage (Table 8.4), similar to neutral engagement, super-spreaders can also trigger an infodemic of false information although the damage is limited for the "Low Effect" cases. The timing of the super-spreader seems to create a little impact on the outcomes of interest since the base scenario with low believability almost follows a stabilized spread.

With the current parametrization, the resulting Total Believer Peak Percentage values for "High Effect" super-spreaders are as high as the neutral engagement case with

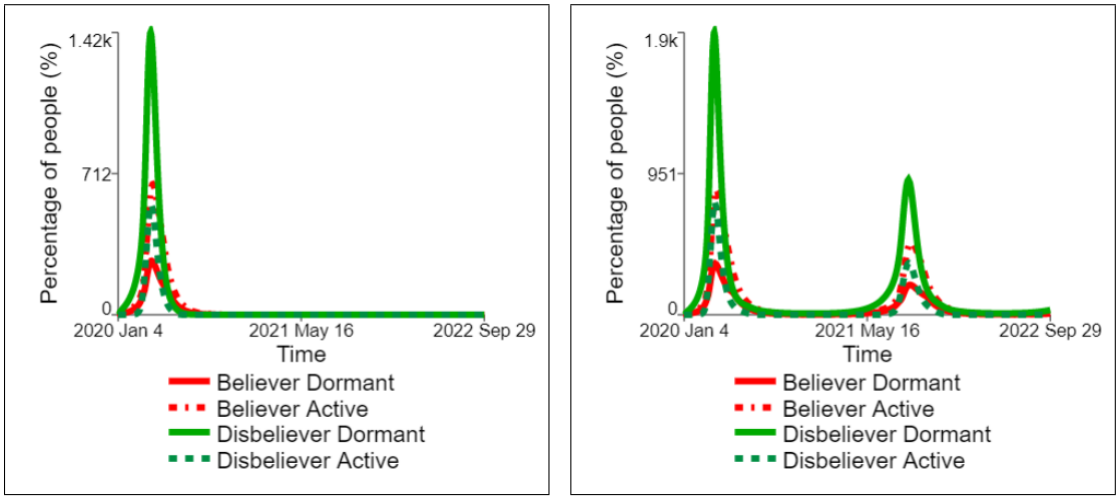
NEF= 0.5. On the other hand, contrasting the believer prevalence values for the same scenarios, neutral engagement produces higher values in comparison to super-spreaders. Moreover, the Exposure Percentages for these three runs are quite close. Thus, the lesser believer prevalence in super-spreader scenarios is not caused by less exposure, rather sustaining the effect of neutral engagement is more effective in increasing the Probability of False Persuasion thus resulting in less prevalence. Conversely, the sudden effect of super-spreaders is more impactful in increasing the peak value without affecting the Believer Prevalence Percentage much.

8.4. Increasing Susceptible Population

As it is already explained in the model building section, one assumption of the model is the constant population through the scope of the analysis since we mainly focus on the dynamics in the relatively short time horizon. However, we know that traditional SIR epidemic models with demographic changes, such as a constant increase in the susceptible population, might show rich behavior that can account for multiple waves of an epidemic [30]. Thus, to conduct a parallel analysis, we experiment with four different values of constant inflow to the Susceptible population. To see the effect of having such an inflow in the long run, we expand the simulation horizon to 1000 days for this scenario.

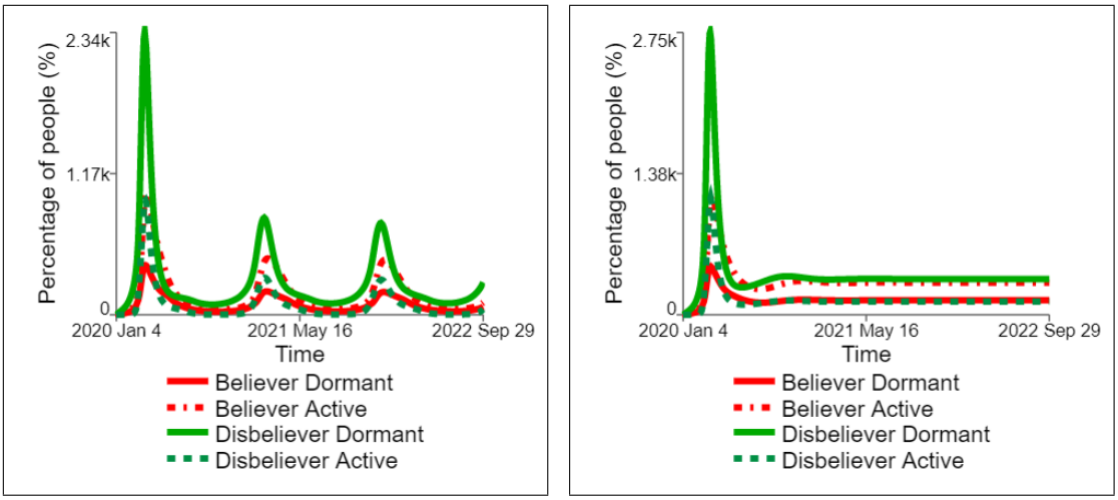
Figure 8.5 depicts the resulting dynamics in Believer and Disbeliever stocks for different levels of constant increase in the Susceptible population. A possible real-life interpretation of such an increase could be either an increase in the user population of that specific social media platform or an expansion in the network of the current Susceptible population. As we observe in Figure 8.5.a, without any increase in the Susceptible, we observe only a single wave for the infodemic. However, when there is an inflow to Susceptible, it allows necessary accumulation in the Susceptible stock to generate a second wave eventually (Figure 8.5b). As the level of this constant flow increases, the frequency of the following waves also increases (Figure 8.5c). If the level of such constant inflow to the susceptible is high enough, the system reaches an equilibrium where the dispute on the social media platform is sustained (Figure 8.5.d).

Interestingly, for such a case, the level of Believer and Disbeliever are quite close to each other despite the fact that Actual Probability False Persuasion is around 0.29. Even with the low tendency to believe such information, the over-represented dominance of Believers in the debate stems from the higher Activation Fraction and longer Active Duration in comparison to the disbelievers.



(a) No Increase

(b) 20 people/day



(c) 40 people/day

(d) 60 people/day

Figure 8.5. Comparative plots of Dormant and Active stock dynamics for different levels of constant inflow to the Susceptible: (a) No Increase, (b) 20 people/day, (c) 40 people/day, and (d) 60 people/day.

9. POLICY INTERVENTIONS

9.1. Decreased Disbeliever Activation Fraction

In their work, Ahmed and colleagues [3] suggest the lesser interaction of the disbeliever group with the believers would be a better option in terms of isolation of the contagion. Thus, one policy can be designed to target the involvement of disbelievers in the debate by either endorsing or discouraging it. To analyze the effectiveness of such a strategy we experiment with different values of Normal Disbeliever Activation within the interval [0, 0.4].

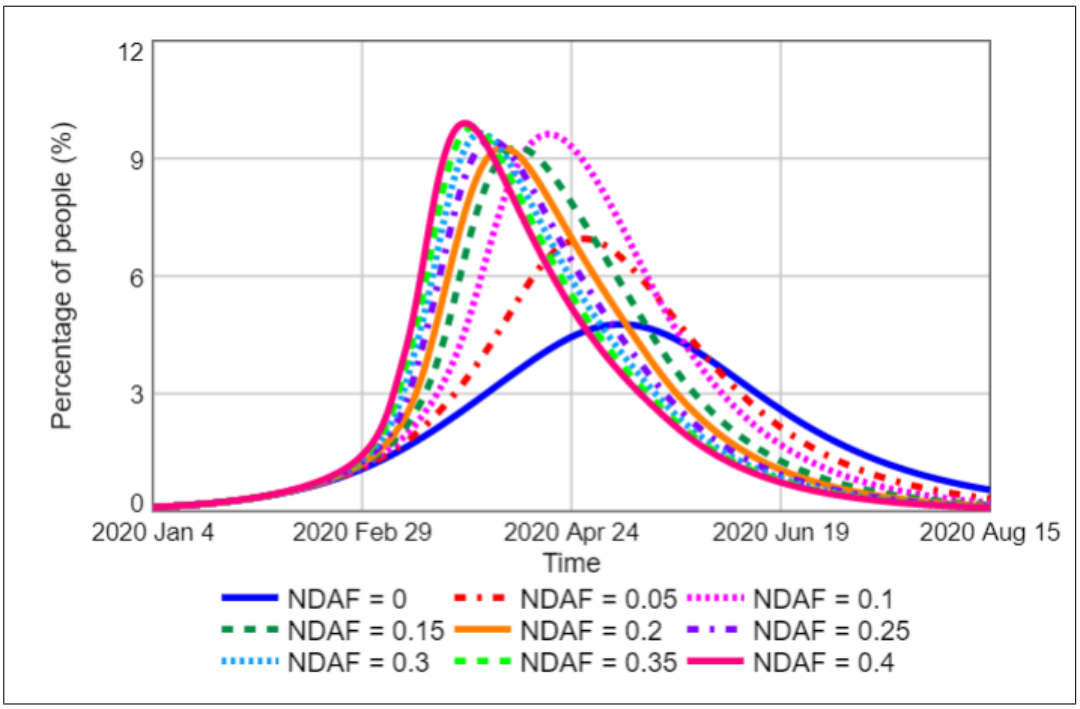


Figure 9.1. The comparative plot of Total Believer Percentage for different values of Normal Disbeliever Activation Fraction (NDAF).

Simulation results are presented in Figure 9.1 and subsequent measures are presented in Table 9.1. The percentage values in Table 9.1 are colored with respect to the given "High" and "Low" values in the lower part of the table. Considering the

Total Believer Peak Percentage the optimal outcome is obtained when the activation of disbelievers is zero i.e. when none of the disbelievers share their opinion or generate corrective information on the subject. Thus, decreasing the Disbeliever Activation seems favorable intervention while trying to minimize the peak value of believers. On the other hand, increasing activation further seems to have little to no effect on Total Believer Peak Percentage although the peak value shows a small increase with the increasing disbeliever activation. Exposure Percentage follows the same pattern as the peak value, since the minimum exposure is obtained for the case of no activation of disbelievers and the maximum is obtained when Normal Disbeliever Activation Fraction is at maximum.

Table 9.1. Comparative table of three output measures for different values of Normal Disbeliever Activation Fraction (NDAF).

	Exposure Percentage	**Total Believer Peak Percentage**	**Believer Prevalence Percentage**
NDAF (0):	90.66	4.77	20.02
NDAF (0.05):	95.64	6.95	22.82
NDAF (0.1):	98.51	9.62	23.6
NDAF (0.15):	98.88	9.28	21.34
NDAF (0.2):	99.22	9.27	20.13
NDAF (0.25):	99.49	9.45	19.51
NDAF (0.3):	99.67	9.63	19.06
NDAF (0.35):	99.79	9.78	18.69
NDAF (0.4):	99.87	9.9	18.37
No Intervention	**99.22**	**9.27**	**20.13**

As opposed to the other two measures Believer Prevalence Percentage does not improve as the Normal Disbeliever Activation Fraction decreases. Furthermore, the outcome is far away from being linear. Starting from zero, increasing Normal Disbeliever Activation Fraction (NDAF) results in a higher Believer Prevalence Percentage initially, whereas after 0.1 we observe better outcomes with the best value of 18.5% is achieved when the Normal Disbeliever Activation Fraction is 0.4.

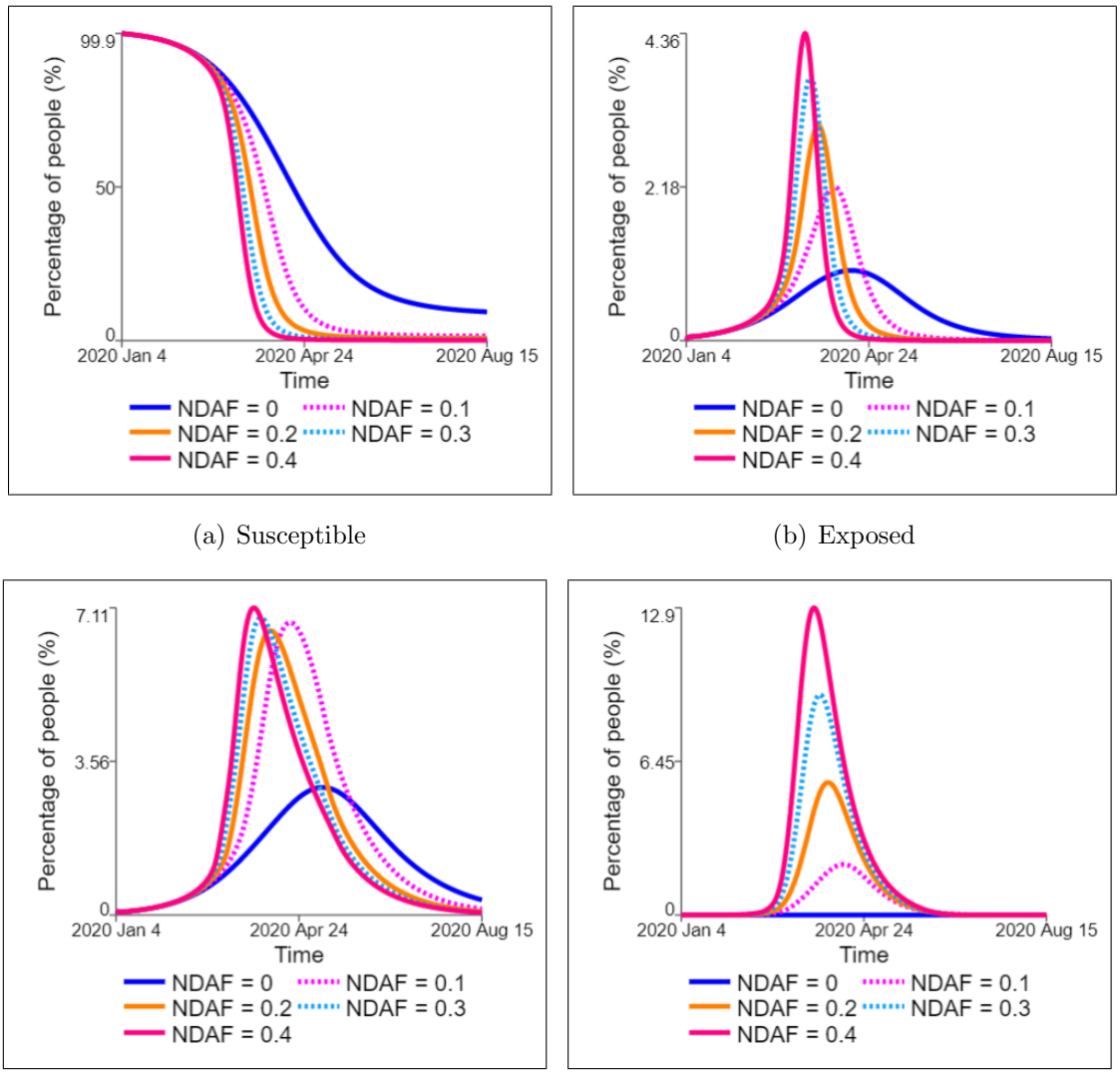
Considering Table 9.1, it is evident that optimal outcomes for two measures, Total Believer Peak Percentage and Believer Prevalence Percentage, are obtained in different parameter settings. Furthermore, the strategy to improve the Base Run (NDAF = 0.2), either by promoting or discouraging the activation of disbelievers, also depends on the selected outcome of interest.

To understand the multifold effects of such intervention, comparative plots of four stocks are presented in Figure 9.2. To simplify the representation, only 5 of the runs are presented.

One direct effect of increased Normal Disbeliever Activation Fraction is the increased Exposure Percentage (Table 9.1). The susceptible people become exposed to misinformation as they get in contact with either Believer Active or Disbeliever Active people. Therefore, if Normal Disbeliever Activation Fraction is constrained, it results in less Disbeliever Active (Figure 9.2d) and a smaller number of people getting exposed to misinformation. Conversely, if Normal Disbeliever Activation Fraction is increased, a higher number of Disbeliever Active would directly increase the flow from Susceptible to Exposed (Figure 9.2a, 9.2b).

An indirect but parallel effect is observed through the Mutual Escalation Loop (R5). As mentioned earlier, as the number of people actively contributing to the debate increases, it also encourages more people from the opposing side to become active. Therefore, increased Normal Disbeliever Activation Fraction also increases the Believer Active (Figure 9.2c) which again directly increases the number of people mobilized from Susceptible to Exposed (Figure 9.2a, 9.2b).

Considering only the Exposure Percentage, these two pathways synergically work which is evident in a parallel increase in Exposure Percentage in response to an increase in Normal Disbeliever Activation Fraction (Table 9.1). The simplicity in this relation is rooted in the fact that Exposure Percentage is not affected by the information dominance in the digital sphere rather it is only a measure of the virality of the misinformation independent of whether it is believed or disbelieved.



(a) Susceptible

(b) Exposed

(c) Believer Active

(d) Disbeliever Active

Figure 9.2. Comparative plots of dynamics of four stocks for different values of Normal Disbeliever Activation Fraction (NDAF): (a) Susceptible, (b) Exposed, (c) Believer Active, and (d) Disbeliever Active.

Same direct (increase in Disbeliever Active) and indirect (increase in Believer Active through Mutual Escalation (R5)) pathways, however, become antagonists when we consider the effect of changing Normal Disbeliever Activation Fraction on Actual Probability of False Persuasion. Considering the direct pathway, the more disbelievers involved in the discussion the more Corrective Info is generated by disbelievers. Conversely, through Mutual Escalation Loop (R5), increased Believer Active will result in

increased levels of Misinformation. Thus, two information types compete in their effect on Probability of False Persuasion. The resulting Actual Probability of False Persuasion is presented in Figure 9.3. It can be seen that for the smaller values of NDAF such as 0.1 the Actual Probability of False Persuasion is greater than no activation case. Thus, for the small activation case, the effect through Mutual Escalation Loop (R5) dominates the direct increase in Disbeliever Active. However, when the NDAF is high enough the increase in Disbeliever Active dominates its indirect effect on Believer Active resulting in less probability of believing misinformation.

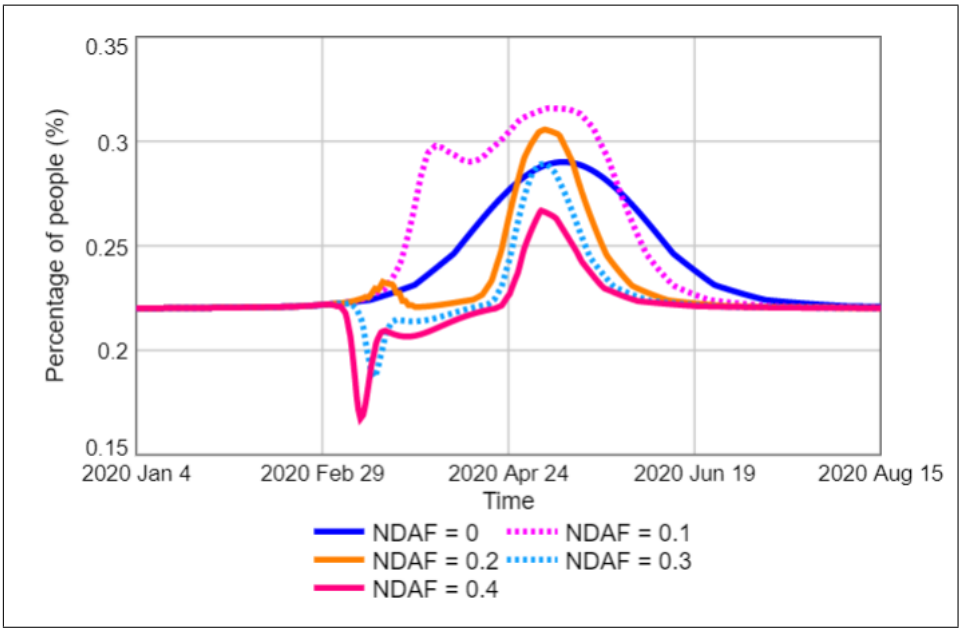


Figure 9.3. The comparative plot of Total Believer Percentage for different values of Normal Disbeliever Activation Fraction (NDAF).

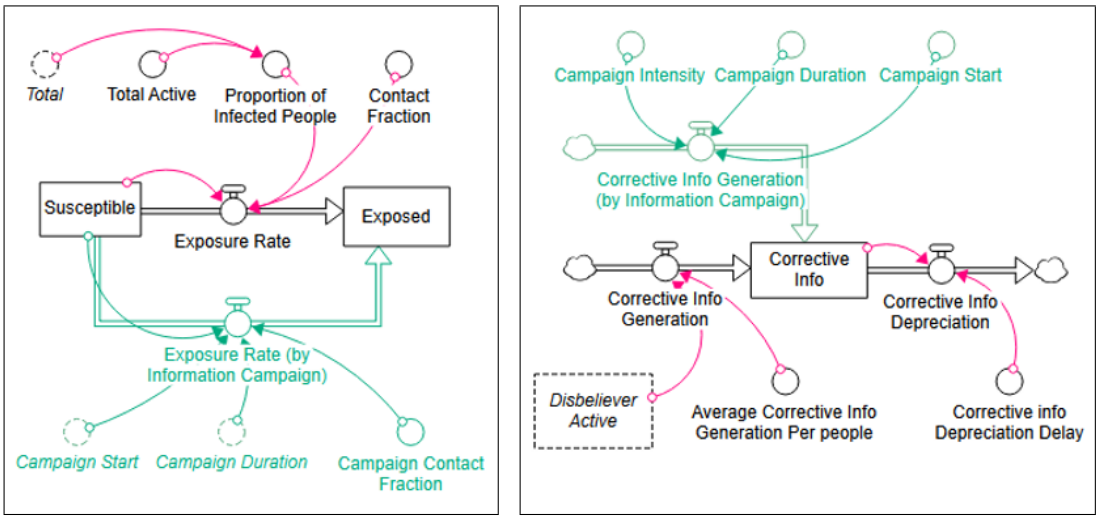
9.2. Corrective Information Campaign

One of the main criticisms of the management of the 5G-COVID 19 conspiracy spread was the lack of explanations and denouncements by the authority figures [3]. Bruns, Hurcombe and, Harrington [52] suggest that one reason that such claims are not falsified by mainstream news and journalist was to avoid generating unnecessary attention on the issue and prevent a possible "backfire" effect by recognizing the con-

spiracy theorists. Using this idea, an "Information Campaign" on the social media platform is tested as a mitigation strategy to assess the effectiveness of intervention and existence of such a backfire effect.

Table 9.2. Parameters used for Super-spreader scenarios.

Parameter Name	Unit	Value
Campaign Start	day	20 - 100
Campaign Contact Fraction	day -1	0.01
Campaign Duration	day	10 - 50
Campaign Intensity	information/day	300



(a) Exposure Rate (b) Misinformation Generation

Figure 9.4. Modified Stock-Flow diagrams after incorporating super-spreader: (a) Exposure Rate, (b) Misinformation Generation.

The information campaign is implemented as an inflow to the Corrective Information starting at Campaign Start for the duration of Campaign Duration. The amount of inflow is denoted as Campaign Intensity and assumed to be constant during the campaign. The analysis is conducted for different values of Campaign Start and Campaign Duration, whereas Campaign Intensity is kept constant (equal to 300) for the analysis to simplify the comparisons. Similar to the super-spreader case, the campaign also

increases exposure which is modeled as a separate flow from Susceptible to Exposed which is denoted as Exposure Rate (by Information Campaign) based on Campaign Contact Fraction for the duration of Campaign Duration.

Table 9.3. Comparative table of Believer Prevalence for different values of Campaign Start and Campaign Duration.

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	18.87	16.58	16.35	16.26	16.22
	30	18.62	16.64	16.43	16.35	16.31
	40	18.17	16.75	16.56	16.5	16.46
	50	17.63	16.94	16.79	16.74	16.7
	60	17.61	17.34	17.22	17.18	17.13
	70	18.49	18.32	18.26	18.21	18.17
	80	19.89	19.81	19.77	19.72	19.7

When we consider the Believer Prevalence, given a fixed start time, increasing the duration of the campaign seems to be effective for all starting dates which presents a consistent scenario with intuitive thinking. However, although the change of direction remains the same, the observed improvement is insignificant for the interventions after 70, given that the Believer Prevalence Percentage was 20.25% for the no intervention case.

Regarding the start time of the campaign, unidirectional thinking would suggest intervening as early as possible would produce better outcomes. However, results indicate that late interventions might generate better outcomes if the duration of the intervention is not lasting. For example, if the campaign duration is determined as 10 days, the best outcome is achieved when we implement such a policy on day 60. Given that the misinformation peaks between days 70 – 80, we can deduce that for such a campaign to be used at maximum effectiveness, it should be sustained until the peak in total believers is achieved.

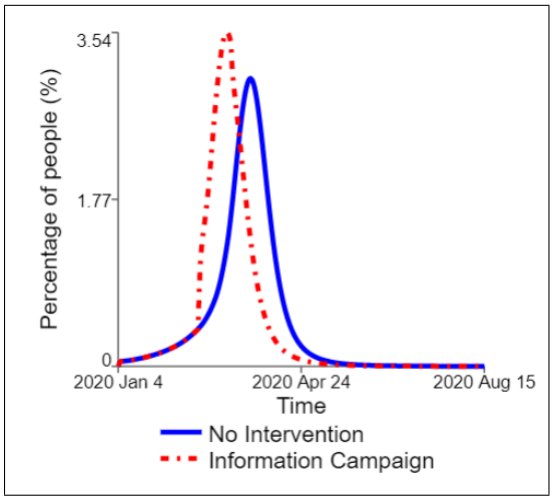
When we evaluated the effect of the information campaign on peak percentage, we see minuscule changes in the Total Believer Peak Percentage (Table 9.4) with respect to the no intervention case. Since conducting such a campaign generates higher exposure one can expect to see higher peak values in response. However, the resulting peak values between the runs are not significantly different suggesting that the improvement provided by the corrective information somewhat balances the possible backlash that can be caused by higher exposure. Thus, we conclude that such interventions do not create a substantial improvement for the peak value in the scope of this analysis.

Table 9.4. Comparative table of Believer Peak Percentage for different values of Campaign Start and Campaign Duration.

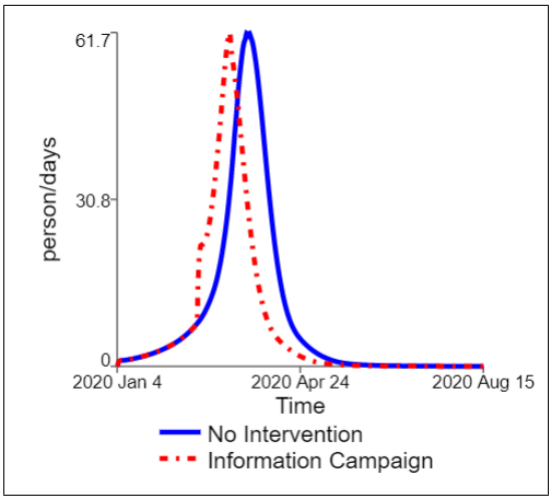
		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	9.47	8.26	8.38	8.46	8.46
	30	9.25	8.24	8.37	8.44	8.44
	40	8.86	8.21	8.36	8.41	8.41
	50	8.37	8.2	8.36	8.37	8.37
	60	8.15	8.26	8.39	8.39	8.39
	70	8.51	8.68	8.71	8.71	8.71
	80	9.42	9.47	9.47	9.47	9.47

Considering the possible back-fire effect mentioned in the literature, we observe no such case for Believer Prevalence in our analysis. Possible pathways to be effective in such adversity could be either increased exposure of the susceptible population to misinformation (Disbeliever Induced Exposure (R2)) or provocation of believers (increase in Actual Believer Activation Fraction through Mutual Escalation (R5)). Figure 9.5 presents the dynamics of Exposed and Believer Dormant stocks for the cases of no intervention and information campaign. We see that the campaign, in fact, causes a higher exposure. However, the flow from Exposed to Believer Dormant seems to change little to none in response to the change in the Exposed. The underlying reason is the effect of Corrective Information on Actual Probability of False Persuasion as evident in Figure 9.5.c. Thus the possibility of such a backfire effect is actually determined

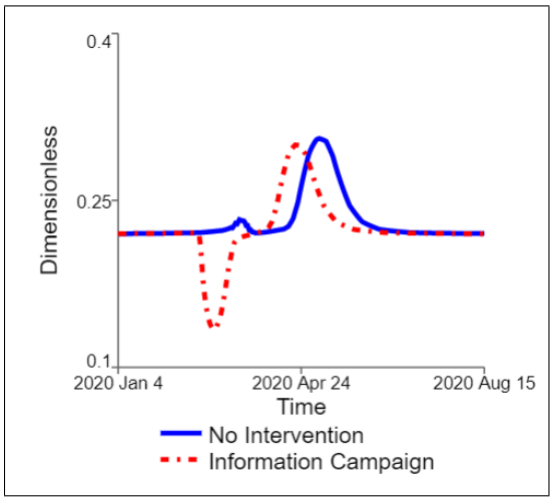
by the trade-off between the possible effect of corrective information on exposure and its effect on the believability of such information. In our parametrization, the results suggest that the benefit of such a campaign outweighs the possible adversity generated by exposure, which supports the qualitative observations made in the literature [52].



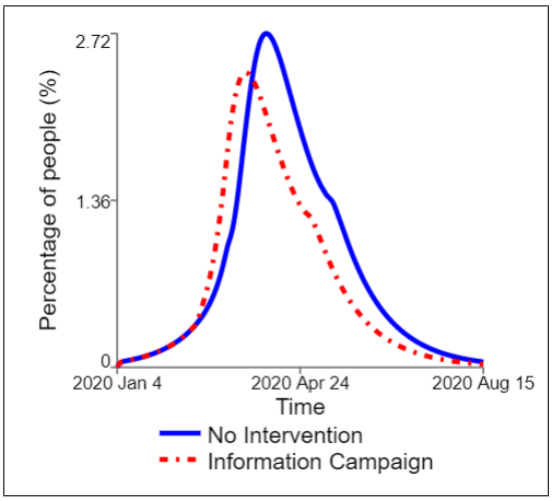
(a) Exposed



(b) Believer Adoption Rate



(c) Actual Probability False Persuasion



(d) Believer Dormant

Figure 9.5. Comparative plots for the Information Campaign and no intervention cases: (a) Exposed, (b) Believer Adoption Rate, (c) Actual Probability of False Persuasion, and (d) Believer Dormant.

10. ANALYSIS OF POLICY INTERVENTIONS FOR VARIOUS SCENARIOS

Thus far we analyzed the effectiveness of two policy interventions for the base case. To expand the analysis and evaluate the consistency of findings in the policy analysis under different scenarios, we conducted experiments on four scenarios: Neutral Sharing, Super-spreader, Lower Believability with Neutral Sharing, and Lower Believability with Super-spreader. The results of the complete analysis are presented in Appendix C whereas a partial summary of the results is presented in this section.

Table 10.1 presents the resulting outcomes of interest for the intervention of decreasing disbeliever activation. In the base case scenario, we observed that the optimal policy was dependent on the selected outcome of interest. For that case, we conclude that decreasing Normal Disbeliever Activation is effective in limiting the Total Believer Peak Percentage whereas it can create higher believer prevalence (Table 9.1). When we consider the results presented in Table 10.1, we observe a similar pattern under all scenarios except the Lower Believability with Neutral Sharing scenario. For that specific case, we see that the optimal outcome of both peak and prevalence is obtained when there is no involvement of the disbelieving population. The scenarios with low believability were the cases where the infodemic is triggered by the involvement of either super-spreaders or neutral sharing. Therefore for these cases, a decrease in the Normal Disbeliever Activation Fraction generates a substantial difference in the Exposure Rate, i.e. prevention of a possible outbreak, which partly explains the improvement observed in the Believer Prevalence.

However, even in the low believability cases, when we consider the Believer Prevalence for the highest and lowest levels of the Normal Disbeliever Activation Fraction, the difference is not large. Moreover, decreasing Normal Disbeliever Activation Fraction to 0 is nearly impossible to be applied as a policy intervention in real life. Even though the policymakers recommend informed people not to engage in disputes on

Table 10.1. Comparative table of three outcomes of interests for different values of Normal Disbeliever Activation Fraction under different scenarios.

		Exposure Percentage	**Total Believer Peak Percentage**	**Believer Prevalence Percentage**
Neutral Sharing	**NDAF (0):**	96.5	8.09	24.68
	NDAF (0.1):	99.2	11.42	23.9
	NDAF (0.2):	99.6	10.91	20.73
	NDAF (0.3):	99.9	11.15	19.75
	NDAF (0.4):	99.9	11.33	19.11
Super-spreader	**NDAF (0):**	93	5.83	22.18
	NDAF (0.1):	98.8	10.52	24.23
	NDAF (0.2):	99.4	10.16	20.92
	NDAF (0.3):	99.8	10.6	20.14
	NDAF (0.4):	99.9	10.96	19.72
Lower Believability with Neutral Sharing	**NDAF (0):**	67	1.48	6.79
	NDAF (0.1):	95.7	4.8	12.09
	NDAF (0.2):	97.6	3.89	8.33
	NDAF (0.3):	99.1	3.91	7.53
	NDAF (0.4):	99.7	3.96	7.12
Lower Believability with Super-spreader	**NDAF (0):**	44.7	0.98	4.48
	NDAF (0.1):	68.7	1.57	6.84
	NDAF (0.2):	88.8	2.31	5.2
	NDAF (0.3):	89.3	1.94	3.69
	NDAF (0.4):	90.4	1.79	3.06

digital space when they encounter such misinformation, it is not realistic for every person to obey such a suggestion. Thus the overall conclusion about the implication of decreasing interaction with believers in the social media platform is that; it is mostly effective in decreasing the Total Believer Peak while possibly creating worse outcomes in terms of Believer Prevelance.

For the corrective information campaign policy, the conclusion for the base case scenario was; applying such a campaign as early as possible and sustaining it as long as possible generates better outcomes in terms of Believer Prevalence. The finding seems to be consistent among all scenarios with the exception of Lower Believability

with the Super-spreader scenario which was presented in Table 10.2. The optimal time of intervention seems to be day 50 which is also the activation time of the super-spreader. Therefore, a conclusion is that immediate action in cases where an infodemic is triggered by a super-spreader is as effective as pre-intervention.

Table 10.2. Comparative table of Believer Prevalence for different values of Campaign Start and Campaign Duration for the Lower believability with a super-spreader scenario.

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	5.52	5.39	4.62	3.78	3.07
	30	5.12	4.35	3.81	3.18	2.79
	40	4.37	3.96	3.24	2.75	2.58
	50	4.88	3.83	2.88	2.6	2.55
	60	4.96	3.81	3.23	3.05	3.01
	70	4.72	4.49	4.28	4.17	4.14
	80	4.97	4.76	4.61	4.56	4.56

In the base run, we also concluded that a back-fire effect due to the exposure generated by the corrective information campaign is not observed within the selected range of parameters. The results from scenarios also support that since such an effect is not observed either in any of the scenarios the policy analysis is applied. In general, the observation that the corrective information campaign is much more effective in targeting the prevalence whereas decreasing disbeliever activation is beneficial in terms of decreasing the peak value holds for other scenarios as well.

11. CONCLUSION

We have built a simulation model for the propagation of 5G-COVID 19 misinformation on social media using the System Dynamics methodology. Starting from more generic models, modeling steps, assumptions and simplifications are discussed regarding the unique characteristics of the problem. The model is constructed using both quantitative and qualitative literature, validated with real data, and challenged with extreme condition tests. Scenario analysis is presented for super-spreader, neutral engagement, and low believability cases. In addition, using the proposed interventions from the literature, policy analysis is conducted for two different interventions: Decreasing Disbeliever Activation and Corrective Information Campaign. Finally, the effectiveness of policies and consistency of findings for different scenarios is discussed.

Scenario analysis presents that for misinformation that would go viral without neutral engagement or super-spreaders, the addition of such cases does not create considerable differences in the outcome of interests. However, the experiment with misinformation with lower believability revealed that the incidence of such scenarios might generate substantial changes by triggering the reinforcing loops and resulting in an epidemic. Super-spreaders constitute a higher risk in terms of increasing peak percentage whereas they are somewhat less effective in believer prevalence as compared to neutral sharing.

The results from the policy analysis indicate that the nonlinear relationships in the system present several trade-offs. Firstly, the outcomes of interest do not always follow the same direction of change for the policy interventions. As an example, regarding the Believer Prevalence Percentage, increasing Disbeliever Activation Fraction unidirectionally produces better outcomes (Table 9.1). On the other hand, the same policy would produce worse outcomes in terms of Total Believer Peak Percentage as evident in the increased level of peak values in Figure 9.1. Such a difference is important as one outcome might be more important for specific types of misinformation. For instance, in 5G-COVID 19 case, looking at the reported violence in the comments on social me-

dia platforms even in the early stages [23] policymakers can focus on minimizing the maximum number of believers to avoid violent protests but, for a case about information that can change long-term behaviors such as climate action, decision-makers may prioritize minimizing the total number of believers during the spread.

Another observation is that decreasing Disbeliever Activation which is commonly cited as a prevention method in the literature does not follow a unidirectional pattern in terms of effectiveness. Our analysis indicates that limiting the involvement of the disbelievers in the online discussion does not necessarily produce better outcomes for Believer Prevalence but rather presents a trade-off between exposing more people or generating more corrective information to inform people. The model at hand is calibrated for a specific case thus generalizability of numerical results to other cases can be questionable.

Results from the policy analysis of the information campaign indicate that such an intervention should be planned carefully to maximize the improvements in the outcomes. If the duration of the campaign is limited then it should be timed carefully, while if such a limitation is not present the impact will be maximum when the intervention starts as early as possible and is sustained until the peak value of the exposed people is achieved. Furthermore, the possibility of a backfire effect through exposure is also considered. Such an effect is neither observed for the base case nor for the experimented scenarios, although a possible occurrence pathway is identified.

The model is built upon one of the simplest information diffusion model, SEIR, with rather small changes in the model structure, utilizing the specific characteristics of the problem considered. Despite such simplicity and very few causal effects included, it still presents several trade-offs that indicate the complexity of the problem at hand. In addition, the causally descriptive nature of the System Dynamics model allowed us to analyze the reasons for the observed behavior which paves the way for extending the analysis by modifying and calibrating the model for other cases of misinformation. Such extension would also provide structural analogies between different cases of misinformation which is a promising path for future work.

To sum up, the analysis of the model represents several trade-offs that can result in unintended consequences for the proposed mitigation strategies in the literature and the designed scenario analysis indicates that the inclusion of neutral sharing or super-spreaders can create significant change in the behavior. The model generates substantial policy implications for a specific case of misinformation while contributing to the modeling practices utilized in the literature.

The future research for this research would be testing the model assumptions, outputs, and robustness of the insights further by making use of richer cross-sectional and dynamic data. Moreover, other possible mitigation strategies can be incorporated into the model to assess the effectiveness for different scenarios. Another agenda is to expand the model by including more user profiles such as "like-seekers" instead of just two opposing sides as there are different motivations for other groups to engage. Different susceptibility of these groups may also be incorporated as such a classification and differentiation is presented in the current literature. Finally, a discussion on similarities and disparities between our specific case and other types of misinformation spread can be useful to infer the harmony of interventions for different cases of false information spread.

REFERENCES

1. Moon, A., "Two-Thirds of American Adults Get News from Social Media: Survey", 2017, www.reuters.com/article/us-usa-internet-socialmedia/two-thirds-of-american-adults-get-news-from-social-media-survey-idUSKCN1BJ2A8, accessed on September 29, 2022.

2. Pennycook, G., J. McPhetres, Y. Zhang, J. G. Lu and D. G. Rand, "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention", *Psychological Science*, Vol. 31, No. 7, pp. 770–780, 2020.

3. Ahmed, W., J. Vidal-Alaball, J. Downing and F. López Seguí, "COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data", *Journal of Medical Internet Research*, Vol. 22, No. 5, pp. 1–9, 2020.

4. Ingram, M., "Facebook's Fact Checking Can Make Fake News Spread Even Faster", 2017, https://fortune.com/2017/05/16/facebook-fact-checking/, accessed on September 29, 2022.

5. Pennycook, G. and D. G. Rand, "The Psychology of Fake News", *Trends in Cognitive Sciences*, Vol. 25, No. 5, pp. 388–402, 2021.

6. Ammara, U., H. Bukhari and J. Qadir, "Analyzing Misinformation Through The Lens of Systems Thinking", *Proceedings of the Truth and Trust Online (TTO)*, pp. 55–63, Virtual, 2020.

7. Kumar, S. and N. Shah, "False Information on Web and Social Media: A Survey", ArXiv:1804.08559v1 [cs.SI], 2018.

8. Varol, O., E. Ferrara, . C. A. Davis, F. Menczer and A. Flammini, "Online Human-Bot Interactions: Detection, Estimation, and Characterization", *Proceedings of the*

Eleventh International Association for the Advancement of Artificial Intelligence (AAAI) Conference on Web and Social Media, pp. 280–289, Montréal, Canada, 2017.

9. Wu, L., F. Morstatter, K. M. Carley and H. Liu, "Misinformation in Social Media: Definition, Manipulation, and Detection", *Special Interest Group on Knowledge Discovery in Data Explorations Newsletter*, Vol. 21, No. 2, p. 80–90, 2019.

10. Lazer, D. M. J., M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild and et al., "The Science of Fake News", *Science*, Vol. 359, No. 6380, pp. 1094–1096, 2018.

11. Allcott, H. and M. Gentzkow, "Social Media and Fake News in the 2016 Election", *Journal of Economic Perspectives*, Vol. 31, No. 2, pp. 211–36, 2017.

12. World Economic Forum, "Digital Wildfires in a Hyperconnected World", 2013, `https://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/`, accessed on September 29, 2022.

13. Caled, D. and M. J. Silva, "Digital Media and Misinformation: An Outlook on Multidisciplinary Strategies Against Manipulation", *Journal of Computational Social Science*, Vol. 5, No. 1, pp. 123–159, 2021.

14. Zhou, X. and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities", *Association for Computing Machinery Computing Surveys*, Vol. 53, No. 5, pp. 1–40, 2020.

15. Lewandowsky, S., U. K. H. Ecker, C. M. Seifert, N. Schwarz and J. Cook, "Misinformation and Its Correction: Continued Influence and Successful Debiasing", *Psychological Science in the Public Interest*, Vol. 13, No. 3, pp. 106–131, 2012.

16. Chan, M. S., C. R. Jones, K. H. Jamieson and D. Albarracín, "Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation",

Psychological Science, Vol. 28, No. 11, pp. 1531–1546, 2017.

17. Vosoughi, S., D. Roy and S. Aral, "The Spread of True and False News Online", *Science*, Vol. 359, No. 6380, pp. 1146–1151, 2018.

18. Vicario, M. D., A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley and W. Quattrociocchi, "The Spreading of Misinformation Online", *Proceedings of the National Academy of Sciences*, Vol. 113, No. 3, pp. 554–559, 2016.

19. Zhao, Z., J. Zhao, Y. Sano, O. Levy, H. Takayasu, M. Takayasu, L. Daqing and S. Havlin, "Fake News Propagate Differently from Real News even at Early Stages of Spreading", *European Physical Journal of Data Science*, Vol. 9, No. 1, pp. 1–14, 2020.

20. Kauk, J., H. Kreysa and S. R. Schweinberger, "Understanding and Countering the Spread of Conspiracy Theories in Social Networks: Evidence from Epidemiological Models of Twitter Data", *Public Library of Science One*, Vol. 16, No. 8, pp. 1–20, 2021.

21. Lotito, Q. F., D. Zanella and P. Casari, "Realistic Aspects of Simulation Models for Fake News Epidemics over Social Networks", *Future Internet*, Vol. 13, No. 3, pp. 1–20, 2021.

22. Bridgman, A., E. Merkley, O. Zhilin, P. J. Loewen, T. Owen and D. Ruths, "Infodemic Pathways: Evaluating the Role That Traditional and Social Media Play in Cross-National Information Transfer", *Frontiers in Political Science*, Vol. 3, 2021.

23. Bruns, A., S. Harrington and E. Hurcombe, "'Corona? 5G? or both?': The Dynamics of COVID-19/5G Conspiracy Theories on Facebook", *Media International Australia*, Vol. 177, No. 1, pp. 12–29, 2020.

24. Brewis, H., "Nightingale Hospital Phone Mast Attacked as 5G Conspiracy Theory Rages", 2020, `https://www.standard.co.uk/news/uk/nhs-nightingale-phone`

-mast-arson-attack-5g-conspiracy-a4414351.html, accessed on September 29, 2022.

25. BBC News, "Mast Fire Probe Amid 5G Coronavirus Claims", 2020, https://www.bbc.com/news/uk-england-52164358#, accessed on September 29, 2022.

26. Agley, J. and Y. Xiao, "Misinformation about COVID-19: Evidence for Differential Latent Profiles and a Strong Association with Trust in Science", *BMC Public Health*, Vol. 21, No. 1, pp. 1–12, 2021.

27. Langguth, J., P. Filkuková, S. Brenner, D. Schroeder and K. Pogorelov, "COVID-19 and 5G Conspiracy Theories: Long Term Observation of a Digital Wildfire", *International Journal of Data Science and Analytics*, pp. 1–18, 2022.

28. Li, M., X. Wang, K. Gao and S. Zhang, "A Survey on Information Diffusion in Online Social Networks: Models and Methods", *Information*, Vol. 8, No. 4, pp. 1–21, 2017.

29. Huppert, A. and G. Katriel, "Mathematical Modelling and Prediction in Infectious Disease Epidemiology", *Clinical Microbiology and Infection*, Vol. 19, No. 11, pp. 999–1005, 2013.

30. Keeling, M. and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*, pp. 26–31, Princeton University Press, Princeton, 2008.

31. Jin, F., W. Wang, L. Zhao, E. Dougherty, Y. Cao, C.-T. Lu and N. Ramakrishnan, "Misinformation Propagation in the Age of Twitter", *Computer*, Vol. 47, No. 12, pp. 90–94, 2014.

32. Mathur, A. and C. P. Gupta, "A Review on Epidemiological Methods to Detect Untrue Information", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 10, No. 11, pp. 16–19, 2021.

33. Khurana, P. and D. Kumar, "Sir Model for Fake News Spreading Through Whatsapp", *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies*, pp. 423–427, Jaipur, India, 2018.

34. Maleki, M., E. Mead, M. Arani and N. Agarwal, "Using an Epidemiological Model to Study the Spread of Misinformation During the Black Lives Matter Movement", ArXiv:2103.12191 [cs.SI], 2021.

35. Paul, A. K. and M. H. A. Biswas, "Modeling the Dynamics of Spreading Rumors and Fake News through Online and Social Media", *Proceedings of the 2nd International Conference on Industrial and Mechanical Engineering and Operations Management*, pp. 214–223, Dhaka, Bangladesh, 2019.

36. Tambuscio, M., G. Ruffo, A. Flammini and F. Menczer, "Fact-checking Effect on Viral Hoaxes", *Proceedings of the 24th International Conference on World Wide Web*, pp. 977–982, New York, USA, 2015.

37. Tambuscio, M., D. F. Oliveira, G. L. Ciampaglia and G. Ruffo, "Network Segregation in a Model of Misinformation and Fact-checking", *Journal of Computational Social Science*, Vol. 1, No. 2, pp. 261–275, 2018.

38. Sterman, J. D., *Business Dynamics: Systems Thinking and Modeling in a Complex World*, McGraw-Hill, Boston, 2000.

39. Barlas, Y., *System Dynamics: Systemic Feedback Modeling for Policy Analysis*, pp. 1131–1175, United Nations Educational, Scientific and Cultural Organization-Encyclopedia of Life Support Systems (UNESCO-EOLSS) Publishers, Paris, 2002.

40. Meese, J., J. Frith and R. Wilken, "COVID-19, 5G Conspiracies and Infrastructural Futures", *Media International Australia*, Vol. 177, No. 1, pp. 30–46, 2020.

41. Trevors, G. and M. C. Duffy, "Correcting COVID-19 Misconceptions Requires Caution", *Educational Researcher*, Vol. 49, No. 7, pp. 538–542, 2020.

42. Ma, S. and H. Zhang, "Opinion Expression Dynamics in Social Media Chat Groups: An Integrated Quasi-Experimental and Agent-Based Model Approach", *Complexity*, Vol. 2021, pp. 1–14, 2021.

43. Dotto, C. and L. Morrish, "Coronavirus: How Pro-mask Posts Boost the Anti-mask Movement", 2020, `https://firstdraftnews.org/articles/coronavirus-how-pro-mask-posts-boost-the-anti-mask-movement/`, accessed on September 29, 2022.

44. Oyserman, D. and A. Dawson, "Your Fake News, Our Facts: Identity-based Motivation Shapes What We Believe, Share, and Accept", R. Greifeneder, M. Jaffe, E. Newman and N. Schwarz (Editors), *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation*, pp. 173–195, Routledge, London, 2020.

45. Ackland, R. and K. Gwynn, "Truth and the Dynamics of News Diffusion on Twitter", R. Greifeneder, M. Jaffe, E. Newman and N. Schwarz (Editors), *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation*, pp. 27–46, Routledge, London, 2020.

46. Ecker, U. K. H., S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga and M. A. Amazeen, "The Psychological Drivers of Misinformation Belief and Its Resistance to Correction", *Nature Reviews Psychology*, Vol. 1, No. 1, pp. 13–29, 2010.

47. van Prooijen, J.-W. and K. M. Douglas, "Belief in Conspiracy Theories: Basic Principles of an Emerging Research Domain", *European Journal of Social Psychology*, Vol. 48, No. 7, pp. 897–908, 2018.

48. Dechêne, A., C. Stahl, J. Hansen and M. Wänke, "The Truth About the Truth: A Meta-Analytic Review of the Truth Effect", *Personality and Social Psychology Review*, Vol. 14, No. 2, pp. 238–257, 2010.

49. Hasher, L., D. Goldstein and T. Toppino, "Frequency and the Conference of Ref-

erential Validity", *Journal of Verbal Learning & Verbal Behavior*, Vol. 16, No. 1, pp. 107–112, 1977.

50. Barlas, Y., "Formal Aspects of Model Validity and Validation in System Dynamics", *System Dynamics Review*, Vol. 12, No. 3, pp. 183–210, 1996.

51. Duvar English, "Turkish Doctor at Target of Anti-vaxxers Finds Calf Tongues in Front of Office", 2022, `https://www.duvarenglish.com/turkish-doctor-professin-senol-at-target-of-anti-vaxxers-finds-calf-tongues-in-front-of-office-news-61091`, accessed on September 29, 2022.

52. Bruns, A., E. Hurcombe and S. Harrington, "Covering Conspiracy: Approaches to Reporting the COVID/5G Conspiracy Theory", *Digital Journalism*, Vol. 10, No. 6, pp. 930–951, 2022.

APPENDIX A: MODEL EQUATIONS

Top-Level Model:

Believer Active(t) = Believer Active(t - dt) + (Believer Activation Rate
- Believer Deactivation Rate) * dt NON-NEGATIVE

INIT Believer Active = Active Initial

UNITS: People

Believer Dormant(t) = Believer Dormant(t - dt) + (Believer Deactivation Rate +
Believer Adoption Rate - Believer Activation Rate - Believer Quit Rate) *
dt NON-NEGATIVE

INIT Believer Dormant = 0

UNITS: People

Corrective Info(t) = Corrective Info(t - dt) + (Corrective Info Generation +
"Corrective Info Generation (by Information Campaign)" -
Corrective Info Depreciation) * dt NON-NEGATIVE

INIT Corrective Info = 0

UNITS: information

Cumulative Believer Tweets(t) = Cumulative Believer Tweets(t - dt) +
(Posted Believer Tweets) * dt NON-NEGATIVE

INIT Cumulative Believer Tweets = 0

UNITS: Tweets

Disbeliever Active(t) = Disbeliever Active(t - dt) +
(Disbeliever Activation Rate - Disbeliever Deactivation Rate) * dt
NON-NEGATIVE

INIT Disbeliever Active = 0

UNITS: People

Disbeliever Dormant(t) = Disbeliever Dormant(t - dt) +
(Disbeliever Deactivation Rate + Disbeliever Adoption Rate -
Disbeliever Activation Rate - Disbeliever Quit Rate) * dt NON-NEGATIVE

INIT Disbeliever Dormant = 0

UNITS: People

Exposed(t) = Exposed(t - dt) + (Exposure Rate +
"Exposure Rate (by Super-spreaders)" +
"Exposure Rate (by Corrective Info Campaign)" - Disbeliever Adoption Rate -
Believer Adoption Rate - Neutral Adoption Rate) * dt NON-NEGATIVE
INIT Exposed = 0
UNITS: People
Misinformation(t) = Misinformation(t - dt) + (Misinformation Generation +
"Misinformation Generation (by super-spreader)" -
Misinformation Depreciation) * dt NON-NEGATIVE
INIT Misinformation = 0
UNITS: information
Neutral(t) = Neutral(t - dt) + (Neutral Adoption Rate - Neutral Quit Rate)
* dt NON-NEGATIVE
INIT Neutral = 0
UNITS: People
Susceptible(t) = Susceptible(t - dt) + (- Exposure Rate -
"Exposure Rate (by Super-spreaders)" -
"Exposure Rate (by Corrective Info Campaign)") * dt NON-NEGATIVE
INIT Susceptible = S Initial
UNITS: People
Total Quit from Believer(t) = Total Quit from Believer(t - dt) +
(Total Quit from Believer Increase) * dt NON-NEGATIVE
INIT Total Quit from Believer = 0
UNITS: People
Total Quit from Disbeliever(t) = Total Quit from Disbeliever(t - dt) +
(Total Quit from Disbeliever Increase) * dt NON-NEGATIVE
INIT Total Quit from Disbeliever = 0
UNITS: People
Total Quit from Neutral(t) = Total Quit from Neutral(t - dt) +
(Total Quit from Neutral Increase) * dt NON-NEGATIVE
INIT Total Quit from Neutral = 0
UNITS: People

Believer Activation Rate =

Believer Dormant*Actual Believer Activation Fraction UNIFLOW

OUTFLOW PRIORITY: 1

UNITS: person/days

Believer Adoption Rate =

Exposed*(1-Neutral Fract)*Actual Probability of False Persuasion/Believer

Adoption Time UNIFLOW

OUTFLOW PRIORITY: 2

UNITS: person/days

Believer Deactivation Rate = Believer Active/Average Believer Active

Duration UNIFLOW

UNITS: person/days

Believer Quit Rate = Believer Dormant/Believer Quit Time UNIFLOW

OUTFLOW PRIORITY: 2

UNITS: People/days

Corrective Info Depreciation =

Corrective Info/Corrective info Depreciation Time UNIFLOW

UNITS: information/days

Corrective Info Generation =

Disbeliever Active*Average Corrective Info Generation Per people UNIFLOW

UNITS: information/days

"Corrective Info Generation (by Information Campaign)" = STEP(1,

Campaign Start)*Campaign Intensity*"IsCampaign (bool)" -

STEP(1, Campaign Start+Campaign Duration)*Campaign Intensity*

"IsCampaign (bool)" UNIFLOW

UNITS: information/days

Disbeliever Activation Rate =

Disbeliever Dormant*Actual Disbeliever Activation Fraction UNIFLOW

OUTFLOW PRIORITY: 1

UNITS: person/days

Disbeliever Adoption Rate =

Exposed*(1-Neutral Fract)*(1-Actual Probability of False Persuasion)/

Disbeliever Adoption Time UNIFLOW

OUTFLOW PRIORITY: 1

UNITS: person/days

Disbeliever Deactivation Rate =

Disbeliever Active/Average Disbeliever Active Duration UNIFLOW

UNITS: person/days

Disbeliever Quit Rate = Disbeliever Dormant/Disbeliever Quit Time UNIFLOW

OUTFLOW PRIORITY: 2

UNITS: People/days

Exposure Rate = Susceptible*(Contact Fraction*Proportion of Infected People)

UNIFLOW

OUTFLOW PRIORITY: 1

UNITS: People/days

"Exposure Rate (by Corrective Info Campaign)" = (STEP(1,

Campaign Start)-STEP(1,

Campaign Start+Campaign Duration))*Susceptible*Campaign Contact Fraction*

"IsCampaign (bool)" UNIFLOW

OUTFLOW PRIORITY: 3

UNITS: People/days

"Exposure Rate (by Super-spreaders)" = (STEP(1,

"Super-spreader start time")-STEP(1,

"Super-spreader start time"+"Super-spreader popularity duration"))

Susceptible*Superspreader Contact fraction*"IsSuperspreader (bool)" UNIFLOW

OUTFLOW PRIORITY: 2

UNITS: People/days

Misinformation Depreciation = Misinformation/Misinformation Depreciation Time

UNIFLOW

UNITS: information/days

Misinformation Generation =

Believer Active*Average Misinformation Generation per people+Neutral*Neutral

Misinformation Generation Per people*Neutral Engagement Fraction UNIFLOW

UNITS: information/days

"Misinformation Generation (by super-spreader)" = (STEP(1,
"Super-spreader start time")-STEP(1,
"Super-spreader start time"+"Super-spreader popularity duration"))*
"Super-spreader misinformation generation"*
"IsSuperspreader (bool)" UNIFLOW
UNITS: information/days
Neutral Adoption Rate = Exposed*(Neutral Fract)/Neutral Adoption Time
UNIFLOW
OUTFLOW PRIORITY: 3
UNITS: person/days
Neutral Quit Rate = Neutral/Neutral Quit Time UNIFLOW
UNITS: People/days
Posted Believer Tweets = Believer Active/Average Time To Tweet UNIFLOW
UNITS: Tweets/days
Total Quit from Believer Increase = Believer Quit Rate UNIFLOW
UNITS: People/days
Total Quit from Disbeliever Increase = Disbeliever Quit Rate UNIFLOW
UNITS: People/days
Total Quit from Neutral Increase = Neutral Quit Rate UNIFLOW
UNITS: People/days
Active Initial = 10
UNITS: People
Actual Believer Activation Fraction =
Normal Believer Activation Fraction*Effect of Corrective Info on Believer
Activation Fraction
UNITS: Per Day
Actual Disbeliever Activation Fraction =
Normal Disbeliever Activation Fraction*Effect of Misinformation on
Disbeliever Activation Fraction
UNITS: Per Day
Actual Probability of False Persuasion =
Normal Prob of False Persuasion+Prob of False Persuasion Effect

Multiplier*Effect of Misinformation on Prob of False Persuasion+

Effect of Corrective Info on Prob of False Persuasion*

Prob of False Persuasion Effect Multiplier

UNITS: Dimensionless

Average Believer Active Duration = 3

UNITS: Days

Average Corrective Info Generation Per people = 1

UNITS: Information/(Day*People)

Average Disbeliever Active Duration = 1

UNITS: Days

Average Misinformation Generation per people = 1

UNITS: Information/(Day*People)

Average Time To Tweet = 7

UNITS: People*days/Tweets

Believer Active Percentage = 100*Believer Active/Total

UNITS: fraction

Believer Adoption Time = 1

UNITS: Days

Believer Dormant Percentage = 100*Believer Dormant/Total

UNITS: fraction

Believer Prevalence Percentage = 100*Total Quit from Believer/Total

UNITS: fraction

Believer Quit Time = 1/0.11

UNITS: Days

Campaign Contact Fraction = 0.01

UNITS: Per Day

Campaign Duration = 20

UNITS: Days

Campaign Intensity = 300

UNITS: information/days

Campaign Start = 70

UNITS: Day

Contact Fraction = 0.8

UNITS: Per Day

Corrective info Depreciation Time = 2

UNITS: Days

Corrective info per capita = Corrective Info/Total

UNITS: information/people

Cumulative Incidence Data = GRAPH(TIME)

Points: (0.0, 0), (1.0, 0), (2.0, 0), (3.0, 0), (4.0, 0), (5.0, 0), (6.0, 0),
(7.0, 0), (8.0, 0), (9.0, 0), (10.0, 0), (11.0, 0), (12.0, 0), (13.0, 0),
(14.0, 0), (15.0, 0), (16.0, 0), (17.0, 0), (18.0, 0), (19.0, 0), (20.0, 0),
(21.0, 0), (22.0, 0), (23.0, 131), (24.0, 136), (25.0, 155), (26.0, 179),
(27.0, 209), (28.0, 253), (29.0, 262), (30.0, 267), (31.0, 275), (32.0, 279),
(33.0, 281), (34.0, 282), (35.0, 298), (36.0, 298), (37.0, 300), (38.0, 301),
(39.0, 301), (40.0, 301), (41.0, 301), (42.0, 302), (43.0, 302), (44.0, 305),
(45.0, 305), (46.0, 309), (47.0, 310), (48.0, 312), (49.0, 314), (50.0, 314),
(51.0, 314), (52.0, 315), (53.0, 315), (54.0, 317), (55.0, 318), (56.0, 321),
(57.0, 321), (58.0, 321), (59.0, 324), (60.0, 335), (61.0, 343), (62.0, 347),
(63.0, 349), (64.0, 349), (65.0, 353), (66.0, 354), (67.0, 358), (68.0, 359),
(69.0, 359), (70.0, 363), (71.0, 365), (72.0, 365), (73.0, 372), (74.0, 375),
(75.0, 375), (76.0, 382), (77.0, 404), (78.0, 417), (79.0, 444), (80.0, 464),
(81.0, 501), (82.0, 517), (83.0, 542), (84.0, 570), (85.0, 599), (86.0, 625),
(87.0, 671), (88.0, 735), (89.0, 800), (90.0, 1023), (91.0, 1333), (92.0, 1592),
(93.0, 1911), (94.0, 2113), (95.0, 2453), (96.0, 2592), (97.0, 2709), (98.0, 2774),
(99.0, 2888), (100.0, 3017), (101.0, 3163), (102.0, 3257), (103.0, 3352),
(104.0, 3475), (105.0, 3533), (106.0, 3597), (107.0, 3675), (108.0, 3740),
(109.0, 3780), (110.0, 3822), (111.0, 3925), (112.0, 3973), (113.0, 4031),
(114.0, 4068), (115.0, 4100), (116.0, 4172), (117.0, 4206), (118.0, 4229),
(119.0, 4277), (120.0, 4380), (121.0, 4444), (122.0, 4507), (123.0, 4559),
(124.0, 4590), (125.0, 4621), (126.0, 4642), (127.0, 4661), (128.0, 4686),
(129.0, 4703), (130.0, 4710), (131.0, 4728), (132.0, 4751), (133.0, 4767),
(134.0, 4786), (135.0, 4791), (136.0, 4805), (137.0, 4813), (138.0, 4827),
(139.0, 4834), (140.0, 4849), (141.0, 4881), (142.0, 4912), (143.0, 4918),

(144.0, 4922), (145.0, 4941), (146.0, 4944), (147.0, 4951), (148.0, 4951), (149.0, 4953), (150.0, 4957), (151.0, 4963), (152.0, 4966), (153.0, 4967), (154.0, 5049), (155.0, 5379), (156.0, 5388), (157.0, 5407), (158.0, 5412), (159.0, 5416), (160.0, 5423), (161.0, 5427), (162.0, 5428), (163.0, 5429), (164.0, 5435), (165.0, 5445), (166.0, 5448), (167.0, 5456), (168.0, 5460), (169.0, 5461), (170.0, 5462), (171.0, 5464), (172.0, 5467), (173.0, 5467), (174.0, 5473), (175.0, 5474), (176.0, 5479), (177.0, 5481), (178.0, 5481), (179.0, 5483), (180.0, 5485), (181.0, 5487), (182.0, 5489), (183.0, 5494), (184.0, 5498), (185.0, 5499), (186.0, 5503), (187.0, 5507), (188.0, 5513), (189.0, 5514), (190.0, 5519), (191.0, 5519), (192.0, 5523), (193.0, 5524), (194.0, 5525), (195.0, 5527), (196.0, 5529), (197.0, 5530), (198.0, 5531), (199.0, 5533), (200.0, 5538), (201.0, 5551), (202.0, 5555), (203.0, 5557), (204.0, 5562), (205.0, 5586), (206.0, 5587), (207.0, 5588), (208.0, 5589), (209.0, 5590), (210.0, 5592), (211.0, 5595), (212.0, 5598), (213.0, 5600), (214.0, 5600), (215.0, 5600), (216.0, 5600), (217.0, 5600), (218.0, 5605), (219.0, 5607), (220.0, 5608), (221.0, 5609), (222.0, 5610), (223.0, 5610), (224.0, 5611)

UNITS: Tweets

Daily Hashtag Data = GRAPH(TIME)

Points: (0.0, 0.0), (1.0, 0.0), (2.0, 0.0), (3.0, 0.0), (4.0, 0.0), (5.0, 0.0), (6.0, 0.0), (7.0, 0.0), (8.0, 0.0), (9.0, 0.0), (10.0, 0.0), (11.0, 0.0), (12.0, 0.0), (13.0, 0.0), (14.0, 0.0), (15.0, 0.0), (16.0, 0.0), (17.0, 0.0), (18.0, 0.0), (19.0, 0.0), (20.0, 0.0), (21.0, 0.0), (22.0, 0.0), (23.0, 131.0), (24.0, 5.0), (25.0, 19.0), (26.0, 24.0), (27.0, 30.0), (28.0, 44.0), (29.0, 9.0), (30.0, 5.0), (31.0, 8.0), (32.0, 4.0), (33.0, 2.0), (34.0, 1.0), (35.0, 16.0), (36.0, 0.0), (37.0, 2.0), (38.0, 1.0), (39.0, 0.0), (40.0, 0.0), (41.0, 0.0), (42.0, 1.0), (43.0, 0.0), (44.0, 3.0), (45.0, 0.0), (46.0, 4.0), (47.0, 1.0), (48.0, 2.0), (49.0, 2.0), (50.0, 0.0), (51.0, 0.0), (52.0, 1.0), (53.0, 0.0), (54.0, 2.0), (55.0, 1.0), (56.0, 3.0), (57.0, 0.0), (58.0, 0.0), (59.0, 3.0), (60.0, 11.0), (61.0, 8.0), (62.0, 4.0), (63.0, 2.0), (64.0, 0.0), (65.0, 4.0), (66.0, 1.0), (67.0, 4.0), (68.0, 1.0), (69.0, 0.0), (70.0, 4.0), (71.0, 2.0), (72.0, 0.0), (73.0, 7.0), (74.0, 3.0), (75.0, 0.0), (76.0, 7.0), (77.0, 22.0), (78.0, 13.0), (79.0, 27.0), (80.0, 20.0), (81.0, 37.0), (82.0, 16.0), (83.0, 25.0), (84.0, 28.0), (85.0, 29.0), (86.0, 26.0), (87.0, 46.0), (88.0, 64.0),

(89.0, 65.0), (90.0, 223.0), (91.0, 310.0), (92.0, 259.0), (93.0, 319.0), (94.0, 202.0), (95.0, 340.0), (96.0, 139.0), (97.0, 117.0), (98.0, 65.0), (99.0, 114.0), (100.0, 129.0), (101.0, 146.0), (102.0, 94.0), (103.0, 95.0), (104.0, 123.0), (105.0, 58.0), (106.0, 64.0), (107.0, 78.0), (108.0, 65.0), (109.0, 40.0), (110.0, 42.0), (111.0, 103.0), (112.0, 48.0), (113.0, 58.0), (114.0, 37.0), (115.0, 32.0), (116.0, 72.0), (117.0, 34.0), (118.0, 23.0), (119.0, 48.0), (120.0, 103.0), (121.0, 64.0), (122.0, 63.0), (123.0, 52.0), (124.0, 31.0), (125.0, 31.0), (126.0, 21.0), (127.0, 19.0), (128.0, 25.0), (129.0, 17.0), (130.0, 7.0), (131.0, 18.0), (132.0, 23.0), (133.0, 16.0), (134.0, 19.0), (135.0, 5.0), (136.0, 14.0), (137.0, 8.0), (138.0, 14.0), (139.0, 7.0), (140.0, 15.0), (141.0, 32.0), (142.0, 31.0), (143.0, 6.0), (144.0, 4.0), (145.0, 19.0), (146.0, 3.0), (147.0, 7.0), (148.0, 0.0), (149.0, 2.0), (150.0, 4.0), (151.0, 6.0), (152.0, 3.0), (153.0, 1.0), (154.0, 82.0), (155.0, 330.0), (156.0, 9.0), (157.0, 19.0), (158.0, 5.0), (159.0, 4.0), (160.0, 7.0), (161.0, 4.0), (162.0, 1.0), (163.0, 1.0), (164.0, 6.0), (165.0, 10.0), (166.0, 3.0), (167.0, 8.0), (168.0, 4.0), (169.0, 1.0), (170.0, 1.0), (171.0, 2.0), (172.0, 3.0), (173.0, 0.0), (174.0, 6.0), (175.0, 1.0), (176.0, 5.0), (177.0, 2.0), (178.0, 0.0), (179.0, 2.0), (180.0, 2.0), (181.0, 2.0), (182.0, 2.0), (183.0, 5.0), (184.0, 4.0), (185.0, 1.0), (186.0, 4.0), (187.0, 4.0), (188.0, 6.0), (189.0, 1.0), (190.0, 5.0), (191.0, 0.0), (192.0, 4.0), (193.0, 1.0), (194.0, 1.0), (195.0, 2.0), (196.0, 2.0), (197.0, 1.0), (198.0, 1.0), (199.0, 2.0), (200.0, 5.0), (201.0, 13.0), (202.0, 4.0), (203.0, 2.0), (204.0, 5.0), (205.0, 24.0), (206.0, 1.0), (207.0, 1.0), (208.0, 1.0), (209.0, 1.0), (210.0, 2.0), (211.0, 3.0), (212.0, 3.0), (213.0, 2.0), (214.0, 0.0), (215.0, 0.0), (216.0, 0.0), (217.0, 0.0), (218.0, 5.0), (219.0, 2.0), (220.0, 1.0), (221.0, 1.0), (222.0, 1.0), (223.0, 0.0), (224.0, 1.0)

UNITS: Tweets

"Daily Hashtag with Moving Average (7 days)" = GRAPH(TIME)

Points: (0.0, 0.0), (1.0, 0.0), (2.0, 0.0), (3.0, 0.0), (4.0, 0.0), (5.0, 0.0), (6.0, 0.0), (7.0, 0.0), (8.0, 0.0), (9.0, 0.0), (10.0, 0.0), (11.0, 0.0), (12.0, 0.0), (13.0, 0.0), (14.0, 0.0), (15.0, 0.0), (16.0, 0.0), (17.0, 0.0), (18.0, 0.0), (19.0, 0.0), (20.0, 0.0), (21.0, 0.0), (22.0, 0.0), (23.0, 18.7), (24.0, 19.4), (25.0, 22.1), (26.0, 25.6), (27.0, 29.9), (28.0, 36.1), (29.0, 37.4), (30.0, 19.4), (31.0, 19.9), (32.0, 17.7), (33.0, 14.6), (34.0, 10.4), (35.0, 6.4), (36.0, 5.1), (37.0, 4.7), (38.0, 3.7), (39.0, 3.1), (40.0, 2.9), (41.0, 2.7), (42.0, 0.6), (43.0, 0.6), (44.0, 0.7), (45.0, 0.6),

(46.0, 1.1), (47.0, 1.3), (48.0, 1.6), (49.0, 1.7), (50.0, 1.7), (51.0, 1.3), (52.0, 1.4),
(53.0, 0.9), (54.0, 1.0), (55.0, 0.9), (56.0, 1.0), (57.0, 1.0), (58.0, 1.0), (59.0, 1.3),
(60.0, 2.9), (61.0, 3.7), (62.0, 4.1), (63.0, 4.0), (64.0, 4.0), (65.0, 4.6), (66.0, 4.3),
(67.0, 3.3), (68.0, 2.3), (69.0, 1.7), (70.0, 2.0), (71.0, 2.3), (72.0, 1.7), (73.0, 2.6),
(74.0, 2.4), (75.0, 2.3), (76.0, 3.3), (77.0, 5.9), (78.0, 7.4), (79.0, 11.3),
(80.0, 13.1), (81.0, 18.0), (82.0, 20.3), (83.0, 22.9), (84.0, 23.7), (85.0, 26.0),
(86.0, 25.9), (87.0, 29.6), (88.0, 33.4), (89.0, 40.4), (90.0, 68.7), (91.0, 109.0),
(92.0, 141.9), (93.0, 183.7), (94.0, 206.0), (95.0, 245.4), (96.0, 256.0),
(97.0, 240.9), (98.0, 205.9),
(99.0, 185.1), (100.0, 158.0), (101.0, 150.0),
(102.0, 114.9), (103.0, 108.6), (104.0, 109.4),
(105.0, 108.4), (106.0, 101.3),
(107.0, 94.0), (108.0, 82.4), (109.0, 74.7), (110.0, 67.1),
(111.0, 64.3), (112.0, 62.9), (113.0, 62.0), (114.0, 56.1), (115.0, 51.4), (116.0, 56.0),
(117.0, 54.9), (118.0, 43.4), (119.0, 43.4), (120.0, 49.9), (121.0, 53.7), (122.0, 58.1),
(123.0, 55.3), (124.0, 54.9), (125.0, 56.0), (126.0, 52.1), (127.0, 40.1), (128.0, 34.6),
(129.0, 28.0), (130.0, 21.6), (131.0, 19.7), (132.0, 18.6), (133.0, 17.9), (134.0, 17.9),
(135.0, 15.0), (136.0, 14.6), (137.0, 14.7), (138.0, 14.1), (139.0, 11.9), (140.0, 11.7),
(141.0, 13.6), (142.0, 17.3), (143.0, 16.1), (144.0, 15.6), (145.0, 16.3), (146.0, 15.7),
(147.0, 14.6), (148.0, 10.0), (149.0, 5.9), (150.0, 5.6), (151.0, 5.9), (152.0, 3.6),
(153.0, 3.3), (154.0, 14.0), (155.0, 61.1), (156.0, 62.1), (157.0, 64.3), (158.0, 64.1),
(159.0, 64.3), (160.0, 65.1), (161.0, 54.0), (162.0, 7.0), (163.0, 5.9), (164.0, 4.0),
(165.0, 4.7), (166.0, 4.6), (167.0, 4.7), (168.0, 4.7), (169.0, 4.7), (170.0, 4.7),
(171.0, 4.1), (172.0, 3.1), (173.0, 2.7), (174.0, 2.4), (175.0, 2.0), (176.0, 2.6),
(177.0, 2.7), (178.0, 2.4), (179.0, 2.3), (180.0, 2.6), (181.0, 2.0), (182.0, 2.1),
(183.0, 2.1), (184.0, 2.4), (185.0, 2.6), (186.0, 2.9), (187.0, 3.1), (188.0, 3.7),
(189.0, 3.6), (190.0, 3.6), (191.0, 3.0), (192.0, 3.4), (193.0, 3.0), (194.0, 2.6),
(195.0, 2.0), (196.0, 2.1), (197.0, 1.6), (198.0, 1.7), (199.0, 1.4), (200.0, 2.0),
(201.0, 3.7), (202.0, 4.0), (203.0, 4.0), (204.0, 4.6), (205.0, 7.9), (206.0, 7.7),
(207.0, 7.1), (208.0, 5.4), (209.0, 5.0), (210.0, 5.0), (211.0, 4.7), (212.0, 1.7),
(213.0, 1.9), (214.0, 1.7), (215.0, 1.6), (216.0, 1.4), (217.0, 1.1), (218.0, 1.4),
(219.0, 1.3), (220.0, 1.1), (221.0, 1.3), (222.0, 1.4), (223.0, 1.4), (224.0, 1.6)

UNITS: Tweets/day

Disbeliever Active Percentage = Disbeliever Active*100/Total

UNITS: fraction

Disbeliever Adoption Time = 1

UNITS: Days

Disbeliever Dormant Percentage = Disbeliever Dormant*100/Total

UNITS: fraction

Disbeliever Quit Time = 1/0.11

UNITS: Days

Effect of Corrective Info on Believer Activation Fraction =

GRAPH((Corrective info per capita)/Standard corrective info per capita)

Points: (0.000, 0.8000), (0.250, 0.8019), (0.500, 0.8114), (0.750, 0.8564),

(1.000, 1.0000), (1.250, 1.0806), (1.500, 1.1261), (1.750, 1.1526), (2.000, 1.1678),

(2.250, 1.1754), (2.500, 1.1810), (2.750, 1.1848), (3.000, 1.1867), (3.250, 1.1894),

(3.500, 1.1905), (3.750, 1.1924), (4.000, 1.1943), (4.250, 1.1962), (4.500, 1.1962),

(4.750, 1.1981), (5.000, 1.2000) GF EXTRAPOLATED

UNITS: Dimensionless

Effect of Corrective Info on Prob of False Persuasion =

GRAPH(Corrective info per capita/Standard corrective info per capita)

Points: (0.000, -0.000669285092428), (0.500, -0.00179862099621),

(1.000, -0.00474258731776), (1.500, -0.0119202922022), (2.000, -0.026894142137),

(2.500, -0.05), (3.000, -0.073105857863), (3.500, -0.0880797077978),

(4.000, -0.0952574126822), (4.500, -0.0982013790038), (5.000, -0.0993307149076)

GF EXTRAPOLATED

UNITS: Dimensionless

Effect of Misinformation on Disbeliever Activation Fraction =

GRAPH((Misinformation per capita/Standard misinformation per capita))

Points: (0.000, 0.000), (0.250, 0.035), (0.500, 0.159), (0.750, 0.493),

(1.000, 1.000), (1.250, 1.366), (1.500, 1.542), (1.750, 1.656), (2.000, 1.727),

(2.250, 1.771), (2.500, 1.806), (2.750, 1.822), (3.000, 1.844), (3.250, 1.865),

(3.500, 1.886), (3.750, 1.900), (4.000, 1.908), (4.250, 1.915), (4.500, 1.922),

(4.750, 1.929), (5.000, 1.936) GF EXTRAPOLATED

UNITS: Dimensionless

Effect of Misinformation on Prob of False Persuasion =

GRAPH(Misinformation per capita/Standard misinformation per capita)

Points: (0.000, 0.000669285092428), (0.500, 0.00179862099621),

(1.000, 0.00474258731776), (1.500, 0.0119202922022), (2.000, 0.026894142137),

(2.500, 0.05), (3.000, 0.073105857863), (3.500, 0.0880797077978),

(4.000, 0.0952574126822), (4.500, 0.0982013790038), (5.000, 0.0993307149076)

GF EXTRAPOLATED

UNITS: Dimensionless

Exposed Percentage = Exposed*100/Total

UNITS: fraction

Exposure Percentage = 100*(Total-Susceptible)/Total

UNITS: fraction

Informed Prevalence Percentage = Total Quit from Disbeliever*100/Total

UNITS: fraction

"IsCampaign (bool)" = 0

UNITS: Dimensionless

"IsSuperspreader (bool)" = 0

UNITS: Dimensionless

Labeled Cumulative Monthy Tweets = GRAPH(TIME)

Points: (15.0, 2), (45.0, 14), (75.0, 199), (105.0, 1141), (135.0, 1348),

(165.0, 1430), (195.0, 1501), (225.0, 1536), (255.0, 1595), (285.0, 1653),

(315.0, 1711), (345.0, 1752), (375.0, 1799), (405.0, 1819), (435.0, 1842),

(465.0, 1864), (495.0, 1882), (525.0, 1902), (555.0, 1918), (585.0, 1926),

(615.0, 1933), (645.0, 1946), (675.0, 1957)

UNITS: Tweets/day

Labeled Monthy Tweet Counts = GRAPH(TIME)

Points: (15.0, 2), (45.0, 12), (75.0, 186), (105.0, 942), (135.0, 208),

(165.0, 82), (195.0, 71), (225.0, 35), (255.0, 59), (285.0, 57), (315.0, 58),

(345.0, 41), (375.0, 47), (405.0, 20), (435.0, 23), (465.0, 23), (495.0, 17),

(525.0, 21), (555.0, 16), (585.0, 8), (615.0, 7), (645.0, 13), (675.0, 11)

UNITS: Tweets/day

Max Tweet Count = 942

UNITS: Tweets/day

Misinformation Depreciation Time = 2

UNITS: Days

Misinformation per capita = Misinformation/Total

UNITS: information/people

Neutral Adoption Time = 1

UNITS: Days

Neutral Dormant Percentage = Neutral*100/Total

UNITS: fraction

Neutral Engagement Fraction = 0

UNITS: fraction

Neutral Fract = 0.1

UNITS: Dimensionless

Neutral Misinformation Generation Per people = 1

UNITS: Information/(Day*People)

Neutral Prevalence Percentage = Total Quit from Neutral*100/Total

UNITS: fraction

Neutral Quit Time = 1/0.11

UNITS: Days

Normal Believer Activation Fraction = 0.68

UNITS: Per Day

Normal Disbeliever Activation Fraction = 0.2

UNITS: Per Day

Normal Prob of False Persuasion = 0.22

UNITS: Dimensionless

“Normalized Daily Hashtag with Moving Average (7 days)” =

“Daily Hashtag with Moving Average (7 days)”/256

UNITS: Tweets/day

Normalized Labeled Monthly Tweet Data =

Labeled Monthy Tweet Counts/Max Tweet Count

UNITS: Dimensionless

Prob of False Persuasion Effect Multiplier = 1

UNITS: Dimensionless

Proportion of Infected People = Total Active/Total

UNITS: unitless

S Initial = 10000

UNITS: People

Standard corrective info per capita = 0.02

UNITS: information/people

Standard misinformation per capita = 0.02

UNITS: information/people

“Super-spreader misinformation generation” = 800

UNITS: Information/Day

“Super-spreader popularity duration” = 3

UNITS: days

“Super-spreader start time” = 50

UNITS: days

Superspreader Contact fraction = 0.02

UNITS: per day

Susceptible Percentage = 100*Susceptible/Total

UNITS: fraction

Total = S Initial+Active Initial

UNITS: People

Total Active = Disbeliever Active+Believer Active+Neutral*

Neutral Engagement Fraction

UNITS: People

Total Active Peak = MAXPEAK(Total Active)

UNITS: People

Total Active Percentage = 100*Total Active/Total

UNITS: fraction

Total Believer = Believer Active + Believer Dormant SUMMING CONVERTER

UNITS: People

Total Believer Peak = MAXPEAK(Total Believer)

UNITS: People

Total Believer Peak Percentage = 100*Total Believer Peak/Total

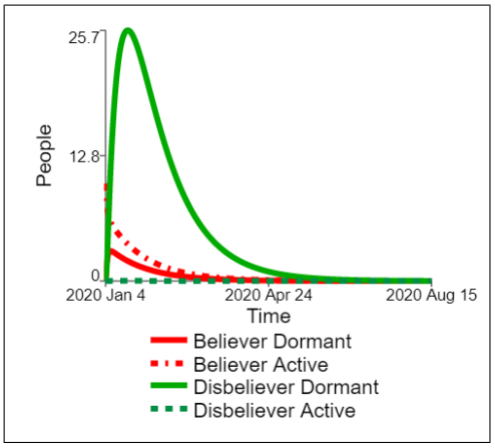
UNITS: fraction

Total Believer Percentage = 100*Total Believer/Total

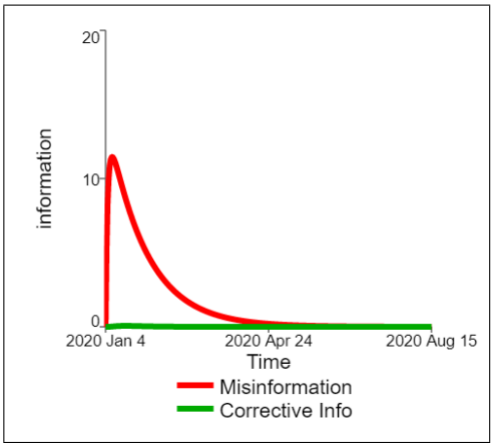
UNITS: fraction

The model has 108 (108) variables (array expansion in parens). In root model and 0 additional modules with 3 sectors. Stocks: 13 (13) Flows: 23 (23) Converters: 72 (72) Constants: 36 (36) Equations: 59 (59) Graphicals: 9 (9) There are also 4 expanded macro variables.

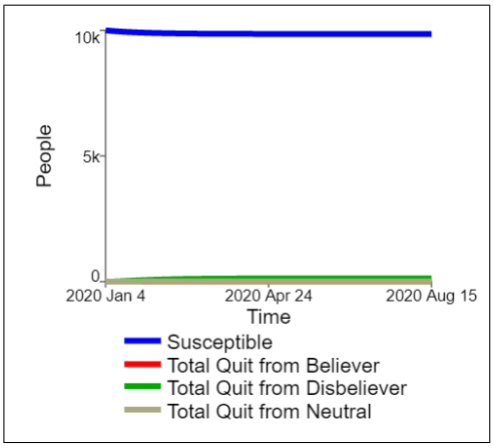
APPENDIX B: ADDITIONAL EXTREME CONDITION TEST RESULTS



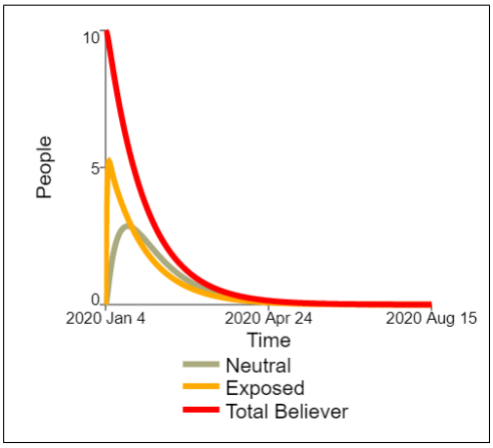
(a) Dormant & Active Stocks



(b) Information Stocks

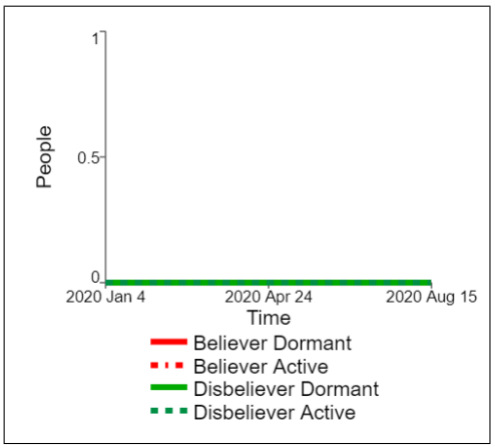


(c) Susceptible and Cumulative Quit Flows

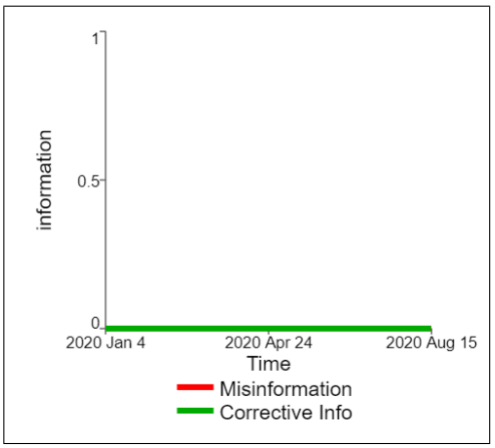


(d) Neutral, Exposed, and Total Believer

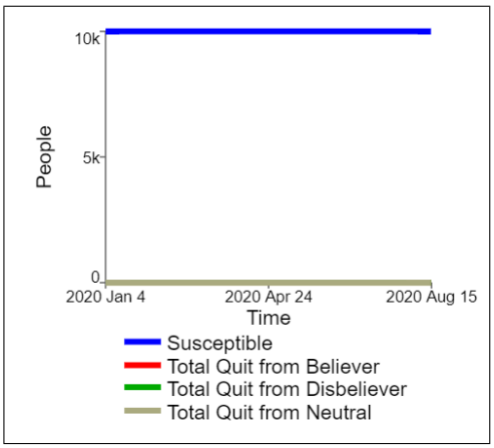
Figure B.1. Extreme condition test results of having Normal Probability of False Persuasion = 0.001: (a) Dormant & Active Stocks, (b) Information Stocks, (c) Susceptible and Cumulative Quit Flows, and (d) Neutral, Exposed, and Total Believer.



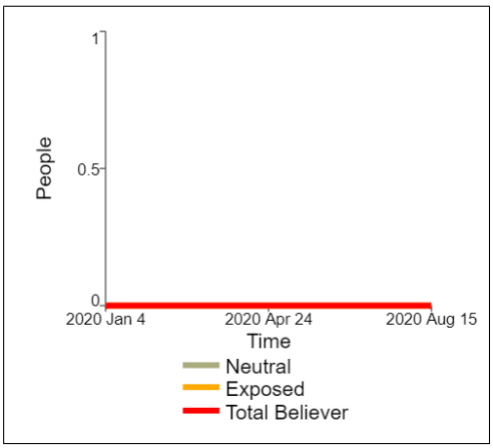
(a) Dormant & Active Stocks



(b) Information Stocks

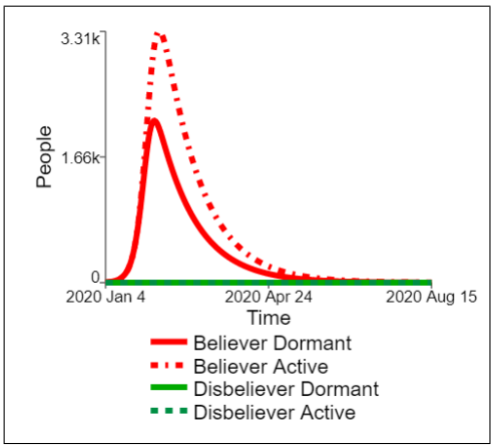


(c) Susceptible and Cumulative Quit Flows

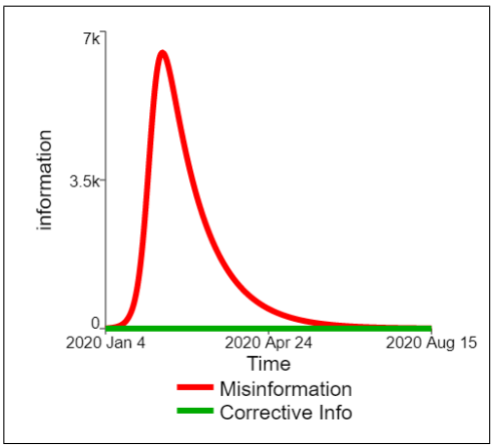


(d) Neutral, Exposed, and Total Believer

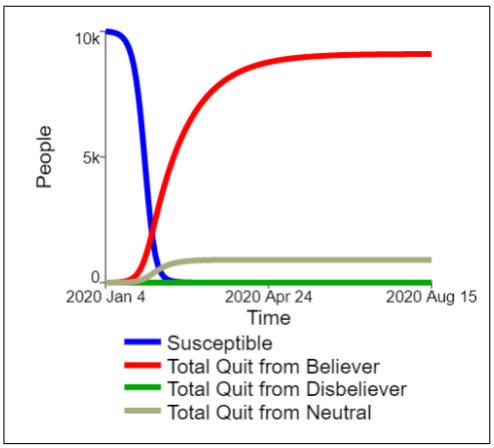
Figure B.2. Extreme condition test results of having no Active Initial: (a) Dormant & Active Stocks, (b) Information Stocks, (c) Susceptible and Cumulative Quit Flows, and (d) Neutral, Exposed, and Total Believer.



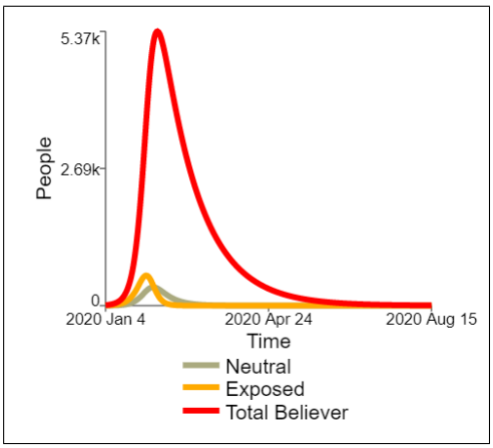
(a) Dormant & Active Stocks



(b) Information Stocks



(c) Susceptible and Cumulative Quit Flows



(d) Neutral, Exposed, and Total Believer

Figure B.3. Extreme condition test results of having Normal Probability of False Persuasion = 0.999: (a) Dormant & Active Stocks, (b) Information Stocks, (c) Susceptible and Cumulative Quit Flows, and (d) Neutral, Exposed, and Total Believer.

APPENDIX C: COMPLETE POLICY INTERVENTION RESULTS FOR ALL SCENARIOS

Table C.1. Policy analysis results of Decreasing Disbeliever Activation for Neutral Sharing Scenario (NEF = 0.6).

	Exposure Percentage	**Total Believer Peak Percentage**	**Believer Incidence Percentage**
NDAF (0):	96.5	8.09	24.68
NDAF (0.05):	98.5	10.62	25.48
NDAF (0.1):	99.2	11.42	23.9
NDAF (0.15):	99.4	10.89	21.66
NDAF (0.2):	99.6	10.91	20.73
NDAF (0.25):	99.8	11.03	20.17
NDAF (0.3):	99.9	11.15	19.75
NDAF (0.35):	99.9	11.25	19.41
NDAF (0.4):	99.9	11.33	19.11
No policy	99.6	10.91	20.73

Table C.2. Comparative table of Exposure Percentage for different values of Campaign Start and Campaign Duration for the Neutral Sharing scenario.

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	99.6	99.6	99.6	99.7	99.7
	30	99.6	99.6	99.6	99.7	99.7
	40	99.6	99.6	99.6	99.7	99.7
	50	99.6	99.6	99.7	99.7	99.7
	60	99.7	99.7	99.7	99.8	99.8
	70	99.7	99.7	99.7	99.8	99.8
	80	99.7	99.7	99.7	99.8	99.8
	90	99.7	99.7	99.8	99.8	99.8
	100	99.7	99.7	99.8	99.8	99.8

Table C.3. Comparative table of Total Believer Peak Percentage for different values of Campaign Start and Campaign Duration for the Neutral Sharing scenario.

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	10.17	9.95	10.05	10.05	10.05
	30	9.94	9.95	10.04	10.04	10.04
	40	9.93	10.02	10.08	10.08	10.08
	50	10.52	10.62	10.62	10.62	10.62
	60	11.02	11.02	11.02	11.02	11.02
	70	10.91	10.91	10.91	10.91	10.91
	80	10.91	10.91	10.91	10.91	10.91
	90	10.91	10.91	10.91	10.91	10.91
	100	10.91	10.91	10.91	10.91	10.91

Table C.4. Comparative table of Believer Prevalence Percentage for different values of Campaign Start and Campaign Duration for the Neutral Sharing scenario.

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	18.53	17.8	17.64	17.6	17.58
	30	18.35	17.96	17.83	17.8	17.78
	40	18.59	18.35	18.26	18.24	18.21
	50	19.74	19.61	19.57	19.55	19.53
	60	20.58	20.54	20.51	20.49	20.49
	70	20.69	20.66	20.63	20.62	20.62
	80	20.69	20.67	20.66	20.65	20.65
	90	20.69	20.68	20.68	20.67	20.67
	100	20.72	20.72	20.72	20.71	20.71

Table C.5. Policy analysis results of Decreasing Disbeliever Activation for Super-spreader Scenario (SST = 60).

	Exposure Percentage	**Total Believer Peak Percentage**	**Believer Incidence Percentage**
NDAF (0):	93	5.83	22.18
NDAF (0.05):	96.8	8.33	24.31
NDAF (0.1):	98.8	10.52	24.23
NDAF (0.15):	99.1	10.08	21.87
NDAF (0.2):	99.4	10.16	20.92
NDAF (0.25):	99.6	10.38	20.45
NDAF (0.3):	99.8	10.6	20.14
NDAF (0.35):	99.8	10.8	19.91
NDAF (0.4):	99.9	10.96	19.72
No policy	99.39	10.16	20.92

Table C.6. Comparative table of Exposure Percentage for different values of Campaign Start and Campaign Duration for Super-spreader Scenario (SST = 60).

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	99.3	99.2	99.2	99.3	99.4
	30	99.3	99.2	99.2	99.3	99.4
	40	99.3	99.2	99.2	99.3	99.4
	50	99.3	99.3	99.4	99.4	99.5
	60	99.3	99.4	99.4	99.5	99.5
	70	99.4	99.5	99.5	99.6	99.6
	80	99.4	99.5	99.6	99.6	99.7
	90	99.4	99.5	99.6	99.6	99.7
	100	99.4	99.5	99.6	99.6	99.7

Table C.7. Comparative table of Total Believer Peak Percentage for different values of Campaign Start and Campaign Duration for Super-spreader Scenario (SST = 60).

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	9.47	8.26	8.38	8.46	8.46
	30	9.39	8.36	8.47	8.54	8.54
	40	9.12	8.47	8.62	8.65	8.65
	50	9.3	9.33	9.48	9.48	9.48
	60	9.44	9.59	9.65	9.65	9.65
	70	10.06	10.17	10.17	10.17	10.17
	80	10.3	10.3	10.3	10.3	10.3
	90	10.16	10.16	10.16	10.16	10.16
	100	10.16	10.16	10.16	10.16	10.16

Table C.8. Comparative table of Believer Prevalence Percentage for different values of Campaign Start and Campaign Duration for the Super-spreader Scenario (SST = 60).

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	18.89	16.6	16.36	16.27	16.24
	30	18.67	16.68	16.47	16.39	16.35
	40	18.38	16.96	16.78	16.72	16.68
	50	18.79	18.38	18.24	18.19	18.15
	60	19.2	18.98	18.9	18.87	18.83
	70	20.31	20.2	20.16	20.12	20.1
	80	20.8	20.75	20.71	20.68	20.67
	90	20.87	20.82	20.79	20.78	20.77
	100	20.85	20.82	20.81	20.8	20.79

Table C.9. Policy analysis results of Decreasing Disbeliever Activation for low believability with Neutral Sharing Scenario (NEF = 0.5 for 400 days).

	Exposure Percentage	**Total Believer Peak Percentage**	**Believer Incidence Percentage**
NDAF (0):	67	1.48	6.79
NDAF (0.05):	84.5	2.65	9.39
NDAF (0.1):	95.7	4.8	12.09
NDAF (0.15):	96.4	4.01	9.29
NDAF (0.2):	97.6	3.89	8.33
NDAF (0.25):	98.6	3.89	7.85
NDAF (0.3):	99.1	3.91	7.53
NDAF (0.35):	99.5	3.94	7.3
NDAF (0.4):	99.7	3.96	7.12
No policy	97.6	3.89	8.33

Table C.10. Comparative table of Exposure Percentage for different values of Campaign Start and Campaign Duration for low believability with Neutral Sharing Scenario (NEF = 0.5 for 400 days).

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	97.2	94.4	93.2	94.2	94.8
	30	97.6	95.6	96.3	96.7	97.1
	40	97.6	97.3	97.5	97.8	98
	50	96.9	97.2	97.5	97.7	98
	60	97.4	97.7	97.9	98.1	98.3
	70	97.9	98.1	98.3	98.4	98.6
	80	97.9	98.1	98.3	98.5	98.7
	90	97.9	98.1	98.3	98.5	98.7
	100	97.9	98.1	98.3	98.5	98.7

Table C.11. Comparative table of Total Believer Peak Percentage for different values of Campaign Start and Campaign Duration for low believability with Neutral Sharing Scenario (NEF = 0.5 for 400 days).

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	3.82	2.13	1.57	1.66	1.66
	30	4.15	2.39	2.6	2.63	2.63
	40	3.85	3.26	3.28	3.28	3.28
	50	2.94	2.95	2.97	2.97	2.97
	60	3.39	3.41	3.41	3.41	3.41
	70	3.93	3.93	3.93	3.93	3.93
	80	3.9	3.9	3.9	3.9	3.9
	90	3.89	3.89	3.89	3.89	3.89
	100	3.89	3.89	3.89	3.89	3.89

Table C.12. Comparative table of Believer Prevalence Percentage for different values of Campaign Start and Campaign Duration for low believability with Neutral Sharing Scenario (NEF = 0.5 for 400 days).

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	8.12	4.4	3.2	3.24	3.17
	30	8.64	4.95	5.16	5.12	5.08
	40	7.84	6.51	6.39	6.36	6.33
	50	6.18	6.03	5.98	5.95	5.9
	60	7.2	7.1	7.07	7.04	6.99
	70	8.21	8.17	8.14	8.09	8.06
	80	8.29	8.26	8.2	8.17	8.15
	90	8.29	8.23	8.19	8.17	8.16
	100	8.24	8.2	8.18	8.17	8.17

Table C.13. Policy analysis results of Decreasing Disbeliever Activation for low believability with Super-spreader Scenario (SST = 50).

	Exposure Percentage	**Total Believer Peak Percentage**	**Believer Incidence Percentage**
NDAF (0):	44.7	0.98	4.48
NDAF (0.05):	52.2	1.15	5.23
NDAF (0.1):	68.7	1.57	6.84
NDAF (0.15):	90.1	2.63	7.6
NDAF (0.2):	88.8	2.31	5.2
NDAF (0.25):	89	2.1	4.25
NDAF (0.3):	89.3	1.94	3.69
NDAF (0.35):	89.6	1.84	3.31
NDAF (0.4):	90.4	1.79	3.06
No policy	88.85	2.31	5.23

Table C.14. Comparative table of Exposure Percentage for different values of Campaign Start and Campaign Duration for low believability with Super-spreader Scenario (SST = 50).

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	89	88.7	86.3	81.6	80.1
	30	88.8	86.1	82.3	80.9	81.7
	40	87.2	82	79.5	79.9	81.6
	50	85.5	76.3	73.7	75.6	78.1
	60	84	77.3	77.3	79.3	81.5
	70	88	88.7	89.5	90.5	91.5
	80	89.5	90.2	91.1	92	92.8
	90	89.3	90.2	91.2	92.1	93
	100	89.5	90.5	91.5	92.4	93.2

Table C.15. Comparative table of Total Believer Peak Percentage for different values of Campaign Start and Campaign Duration for low believability with Super-spreader Scenario (SST = 50).

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	2.31	2.19	1.77	1.46	1.46
	30	2.31	1.93	1.56	1.56	1.56
	40	2.13	1.55	1.55	1.55	1.55
	50	1.55	1.44	1.44	1.44	1.44
	60	1.64	1.64	1.64	1.64	1.64
	70	2.27	2.27	2.27	2.27	2.27
	80	2.31	2.31	2.31	2.31	2.31
	90	2.31	2.31	2.31	2.31	2.31
	100	2.31	2.31	2.31	2.31	2.31

Table C.16. Comparative table of Believer Prevalence Percentage for different values of Campaign Start and Campaign Duration for low believability with Super-spreader Scenario (SST = 50).

		Campaign Duration				
		10	**20**	**30**	**40**	**50**
Campaign Start	**20**	5.52	5.39	4.62	3.78	3.07
	30	5.12	4.35	3.81	3.18	2.79
	40	4.37	3.96	3.24	2.75	2.58
	50	4.88	3.83	2.88	2.6	2.55
	60	4.96	3.81	3.23	3.05	3.01
	70	4.72	4.49	4.28	4.17	4.14
	80	4.97	4.76	4.61	4.56	4.56
	90	4.91	4.74	4.67	4.66	4.66
	100	4.91	4.82	4.8	4.8	4.8