# How We Change Our Minds Matters

## Misinformation, ABMs, and Deep Uncertainty

EPA2942: EPA Master Thesis

Felicitas Reddel

Delft University of Technology

TU Delft

# How We Change Our Minds Matters

## Misinformation, ABMs, and Deep Uncertainty

by

Felicitas Reddel

Student Number: 5360293

| | |
|---|---|
| Chair & First supervisor: | Jan Kwakkel |
| Second supervisor: | Amineh Ghorbani |
| | |
| Institution: | Delft University of Technology |
| Faculty: | Faculty of Technology, Policy, and Management (TPM) |
| Degree: | M.Sc. Engineering and Policy Analysis |
| Specialization: | Modeling, Simulation and Gaming |
| Project Duration: | Feb, 2022 - Aug, 2022 |
| Date of Defense: | 29 August 2022 |

The associated code and the model are available at
https://github.com/felicity-reddel/MisinfoPy

TU Delft

# Preface

The Masters degree of 'Engineering and Policy Analysis' provided me with a lot of tools for thinking through grand challenges in complex socio-technical systems. Furthermore, it gave me the opportunity to learn about powerful tools to model these systems. With these tools, we can acknowledge what we do not know and we can benefit from the modern age of computing to understand deeply uncertain systems better. And we might even find solutions that could work well in spite of that uncertainty. I have always been motivated to help others and to aim at leaving a positive mark on the world. But learning about these tools has empowered me to pursue this goal in a more effective fashion because the map is not the territory. And when we acknowledge where our map is fuzzy, when we pay attention to see in what way the map is blurry and to what extent it is so, then we can think clearly without tricking ourselves into the tempting valleys of false certainty.

I am grateful that I had the opportunity to explore the societally relevant challenge of misinformation on social media during the previous internship with the Dutch Ministry of Interior under Haye Hazenberg. The work of that internship laid the foundation for this thesis project. Without Haye and his support, the previous internship, and by implication also this thesis, would have not taken the shape it did. The internship taught me a lot about designing and implementing agent-based models that are at the interface of society and technology. Within this thesis project, I could not only further hone my modeling skills, but also apply some of the methods of the fascinating field of Decision-Making under Deep Uncertainty. I could apply many skills that I had acquired and practiced over the last two years. And I could do so in an area that is impacting people's lives. The previous research that I could find on this grand challenge had virtually ignored that there is structural uncertainty at the heart of the models. Within this project, I had the chance to start exploring this deep uncertainty. The results of this project indicate that this uncertainty does make a difference when deciding which strategy might be a good way forward to tackle this grand challenge. There is a lot left to explore, but I am happy that I had the chance to start the exploration.

I am incredibly grateful for the mentorship I received along the way. Amineh has supported me throughout the years – be it within the TPM AI-lab, in educational projects such as the previous internship and this thesis, or in helping me to go after my career aspirations. I am extremely grateful for her support in all these facets of my life at the TU Delft. And regarding Jan, I am not even sure where to start. Despite being a key figure in the field of Decision-Making under Deep Uncertainty and Exploratory Modeling, and despite having a schedule that would not be manageable for most people, Jan has been incredibly caring and always quick in providing support. What I have learned from him has inspired me and has shaped my overall path. I am incredibly grateful for having had the opportunity to be supervised by Jan.

Lastly, I want to thank my husband Max. When I was outside of my comfort zone, Max has been an amazing pillar of strength to me. His emotional support has been nurturing my growth. This, and him celebrating the wins with me, made this project far more fulfilling and more empowering than it otherwise would have been.

*Felicitas Reddel*
*The Hague, August 2022*

# Summary

Misinformation on social media is an urgent grand challenge. Misinformation has caused excess deaths because people abstain from getting vaccinated and other evidence-based prevention behaviors. Misinformation also influences various other important topics such as climate change. And it has the potential to influence countless other areas. In order to exemplify the broader point, this project focuses on simple example of beliefs around the safety of COVID-19 vaccinations. To find ways to successfully tackle this grand challenge, it is crucial to have thorough understanding of the system. Modeling and simulation can be a powerful tool to support our reasoning about big and complex systems such as misinformation on social media. Therefore, I choose to look at how modeling and simulation can be useful for the study of misinformation on social media and of potential counter-measures.

Agent-based models (ABMs) are one of the useful modeling paradigms for this grand challenge. And while there is a body of literature on ABMs in the field of misinformation research, there is structural uncertainty about how to represent the way that people change their minds on social media. Different types of representations of this updating process are used. It is unclear which of them is the most suitable representation of the real-world process and also to what extent it makes a difference for the choice of counter-measures. Moreover, the choice between these different belief update functions is usually not discussed. And to the best of my knowledge, nobody has explored the issue of whether the choice between belief update functions makes a substantial difference in the conclusions from the studies.

Because of the significance of this grand challenge and the lack of exploring a key structural uncertainty, I choose to apply a method for exploring uncertainty in the context of ABMs. More specifically, because the structural uncertainty about the belief update function is a central component of models in this field, I explore a method for handling this structural uncertainty. This project is a show case of the value that methods from the field of Decision-Making Under Deep Uncertainty (DMDU) have for the field of misinformation focused ABMs. Yet, applying a DMDU approach is not only useful for enabling exploration of uncertainties. With many DMDU methods, it is possible to evaluate policies based on not only a single, but on multiple objectives. As far as I know, also the evaluation of multiple objectives has not previously been done in the field of opinion dynamic models such as ABMs which focus on misinformation on social media. However, policies that aim at tackling the misinformation challenge do not only impact one single stakeholder, but a multitude of diverse stakeholders who care about various aspects of the system. If we pick policies by only optimizing for one objective, we run the risk of merely shifting the problem. To find solutions that are sustainable and work for the whole system, it is helpful to consider multiple metrics that stakeholders care about. The ranking and filtering by multiple objectives is not trivial. But there is a method called non-dominated ranking which can be applied to do exactly that. This results in so-called Pareto-optimal policies. It is in this specific niche that I pursue the following methodological question within the field of agent-based misinformation modeling:

**Main Research Question**

How does the consideration of **structural uncertainty** with respect to the choice between different **belief update functions** influence the resulting **Pareto-optimal policies** and their performance?

I look at three alternative belief update functions, where each belief update function is represented by

one model. I show that the choice of the belief update function makes a significant difference for what kind of policies are Pareto-optimal and for the outcomes that stem from these policies. To investigate how the choice of the belief update function influences which policies are Pareto-optimal and what kind of outcomes result, I apply the DMDU-method of Many-Objective Robust Decision-Making (MORDM) approach. With DMDU methods, modellers can acknowledge the uncomfortable situation in which we know that we have uncertainties, ruining the possibility of using models as reliable prediction machines. These uncertainties can be about the real world's states (i.e., parametric uncertainties) or its processes (i.e., structural uncertainty). When applying DMDU methods, modellers can aim to find policies that perform robustly over a large number of possible instantiations of parametric or structural uncertainties. In this project, I first evaluate more than 26'000 candidate policies with each of the three belief update functions. Then, I select a set of Pareto-optimal policies for each belief update function. Additionally, I select a set of policies that seem optimal when only considering a single metric. Subsequently, I re-evaluate Pareto-optimal policies of each belief update function under deep uncertainty to gain a better impression of their performance. Finally, I compare the commonalities and differences between the selected policies and their performances. This, I do for either method of selection and for all three belief update functions.

To explore the structural uncertainty, I use a model which can be instantiated with either of the alternative belief update functions. I refer to these three possible instantiations as the three different models. The first model uses the commonly used function based on the research by Deffuant (hereafter 'DEFFUANT model'). In it, beliefs are always updated by a fix percentage towards the newly incoming information. In this project, this newly incoming information is the belief that is represented in a seen post. The second model samples whether a belief update happens or not. If an update happens, the new belief is the average between the previous belief and the newly incoming information. We call this the 'SAMPLE model'. Unfortunately, neither of these two models includes well-established phenomena from social psychology. Examples of such phenomena include for instance that we are more willing to update towards beliefs that are more similar to ours, that we have limited attention capacity, and that it takes more to change someone's mind when they are very convinced of their current belief than when they are uncertain. The third model was chosen to fill this void by basing its belief update function on Social Impact Theory (SIT) and adjusting this theory to the context of social media. This model is referred to as the 'SIT model'. The main findings of the project are the following:

1. There is a clear distinction between the models' optimal policies as well as their outcomes.
2. Differences in parameters do make a difference.
3. The models' optimal policies exhibit an order in how optimistic their outcomes are. This order (in descending direction) is DEFFUANT, SAMPLE, and SIT.
4. The outcomes of the DEFFUANT and the SAMPLE model are more similar to each other than to the SIT model.

The main methodological take-away is that the DMDU approach can bring substantial value to the field of ABM-based studies on the grand challenge of misinformation on social media platforms. While this is shown by a simple exploration of the structural uncertainty with respect to the belief update, many more insights could be gathered by utilizing the DMDU approach. For instance, the DMDU approach offers state-of-the-art methods to identify vulnerable scenarios, i.e., scenarios which would be particularly bleak. Another example could be to explore different problem formulations with different sets of objectives or other structural uncertainties such as the posting behavior. Furthermore, by utilizing the tools of DMDU, also society as a whole can benefit. By including multiple objectives and a wide range of considered uncertainties, the many different world-views and values of the diverse stakeholders can be taken into account in order to avoid potential policy gridlock situations. This could contribute to tackling the misinformation grand challenge more successfully and thus for instance lead to more people embracing evidence-based medical interventions.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1. Context

The influence of social media on people's beliefs and behaviors can have far-reaching consequences. This has once more become painfully clear during the current COVID-19 pandemic. Social media misinformation is fueling vaccine hesitancy despite overwhelming evidence that the vaccines are safer than waiting to get infected with COVID-19 without having been fully vaccinated (Wilson and Wiysonge, 2020). On social media, false information has been shown to spread faster and further than true information (Vosoughi et al., 2018). But before aiming to find a good approach to the problem of misinformation on social media, let us first settle what exactly we mean with the term *misinformation* and whether it really is such a relevant issue that is worth our time.

### 1.1.1. What is meant with misinformation?

Several definitions for 'misinformation' exist and academic literature often relies on definitions from a dictionary or refrains from specifying the exact intended meaning of the term (Karlova and Fisher, 2013). However, some dictionary entries are contradicting one another (Treen et al., 2020). Additionally, adding the related term of 'disinformation' to the mix can lead to confusions. What the two terms generally have in common is that they refer to information which is false or misleading. Furthermore, most define 'disinformation' as *intentionally* incorrect or deceiving. The aspect that seems to cause the bulk of the confusion is whether or not the term 'misinformation' implies that the information is *un*intentionally false or misleading. In other words, a key disagreement is whether 'disinformation' is a subset of 'misinformation'.

For the purposes of this project, 'misinformation' includes both, unintentional and intentional falsehoods (i.e., the subset-view is chosen). The main reason for this choice is that it is notoriously difficult to know with certainty what the intention behind sharing a specific piece was. Furthermore, it seems rather safe to assume, that whether the information and the belief spreads through the population, is not substantially influenced by whether the piece of content was deliberately misleading or whether the author is convinced of the message's truthfulness.

### 1.1.2. Is misinformation a grand challenge?

Grand challenges are high-level hurdles to important societal problems. The stakes of these issues are international or global, and the challenges are notoriously difficult to solve. Typically, finding solutions is further aggravated by incomplete or changing requirements.

Past research indicates that social media can effect users' beliefs (Allcott et al., 2020), increase polarization (Stray, 2021), and has enabled misinformation to spread wide and far (Kumar and Shah, 2018; Pierri and Ceri, 2019; Y. Wang et al., 2019a). Social media misinformation has been fueling vaccine hesitancy despite overwhelming evidence that the vaccines are safer than waiting to get infected with COVID-19 without having been fully vaccinated (Wilson and Wiysonge, 2020). Such misinformation has misled people during the COVID-19 pandemic and swayed them to abstain from prevention behaviors that are evidence-based (Kim & Tandoc Jr, 2022) (including vaccinations (Pierri et al., 2022)) and from effective care. This behavior can be life-threatening (Y. Wang et al., 2019b).

And next to influencing the course of pandemics, misinformation can impact many other of our current and future challenges. This is because the performance of democracies depends on the ability of their citizens and leaders to make informed decisions. But if we rely on low-quality information our effort of arriving at good decisions can derail . How successful we are at handling this grand challenge could make the difference of whether we are sliding towards being in a gridlocked and internally corroding society, or whether we arrive at an empowered society that can actively move forward to tackle any challenge that is thrown at it.

### 1.1.3. What is the current context of policy-making for this grand challenge?

Policy makers are aware of this grand challenge and aim to intervene by crafting legislative initiatives like the Digital Service Act package (DSA). Goals behind the DSA include making large platforms responsible for content that has been reported by users, as well as increasing transparency towards regulators in terms of the inner working of their algorithms, decisions around taking down content, and targeted advertising. Currently, the process behind new regulation like the DSA relies heavily on stakeholder participation (European Commission, 2021a). Stakeholder participation has identified various promising policy-directions (e.g., media literacy, 'safe design', accountability of recommender systems, flagging, etc.) (European Commission, 2021b). However, estimating the effect of a policy on a complex system is no easy feat.

### 1.1.4. Simulation modeling can support policy-making by putting the skills of computers to use.

If we want to interfere with a system to achieve some specific change, it is very valuable to understand the system well enough to be able to roughly predict how it might react to our perturbation. Without such understanding, we would bear the costs of our efforts while also risking that the result could be net-neutral or even net-negative. Such understanding depends on our ability to simulate the system. However, systems that are of interest for policy-making can easily become too big and too complicated for humans to accurately and reliably simulate. Simulation modeling can be even more helpful if the policy-making is aimed at a grand challenge. This is because grand challenges are usually happening in complex socio-technical systems. These systems have properties that make it difficult for humans to simulate them sufficiently well.[1]

One key characteristic of complex socio-technical systems is that they include feedback loops, delays, and thus exhibit non-linear dynamics (Vespignani, 2012). Humans are hard-wired to think linearly (De Langhe et al., 2017). This is a heuristic that proved incredibly useful in the environment to which we are evolutionary adapted ('Environment of Evolutionary Adaptedness') (Bennett, 2018; Bowlby, 1982). On the flip side, this heuristic makes it notoriously difficult for us to reliably think through the big and

[1] While simulation modeling can be incredibly helpful to support human thinking, I would like to add a cautious note. Model results run the risk of being seen as impartial, unbiased, and even as delivering accurate real-world prediction. However, models are no crystal balls, but just help us think through what all our included assumptions would imply if all these assumptions were correct and if they would be cover whole detail of what is going on in the world. But because we cannot encode any detail perfectly and because there are so many aspects that we are uncertain about, by default we cannot assume that a model's outcomes are accurate predictions of the their real-world counterpart. Therefore, if not clearly stated otherwise, I am referring to what is happening in the model.

complex systems of our modern world. Luckily, computers have complementary strengths here. They simply compute and do not face difficulties in including feedback loops or delays. Another characteristic that makes it harder for us humans to think through grand challenges is the uncertainty that these systems come with. When mentioning uncertainty, one might think of a coin toss. We cannot predict accurately what the result of any specific coin toss will be. Thus, the outcome of the toss is called *uncertain*. However, there is uncertainty that runs deeper than not being able to predict the outcome of an individual coin toss. We can distinguish five levels of uncertainty , ranging from *Marginal Uncertainty* to *Recognized Ignorance*. The structural uncertainty of the belief update function falls into the level 4 – 'deep uncertainty'. For this category, the *Decision-Making under Deep Uncertainty* (DMDU) approach is a suitable choice[2].

## 1.2. Related Work

In line with the societal interest in this grand challenge, quite some research on misinformation and how opinions are spreading within populations have previously been undertaken. It is to note that much of the work includes the implicit underlying assumption that beliefs are key drivers of human decision-making. This is in line with the Belief-Desire-Intention model, which based on research in psychology and 'artificial agency' (Bratman et al., 1988). I refer to this assumption as the 'BDI-assumption'.

When aiming for better understanding of this grand challenge, various paths can be chosen. There are various different approaches in modeling and simulation (for instance agent-based models, discrete time models, and system dynamics). And there are more ways to model this grand challenge, which commonly do not rely on computer simulation (e.g., comparative cognitive maps, resource dependence analyses, qualitative consequence tables, single or multi-actor system diagrams, or cooperative game theory). And naturally, there is also work that is purely empirical or philosophical in nature. All these different approaches are essentially different lenses to look at a given issue and understand it better. While various other approaches could have been employed in this project, I focus on agent-based models. In the following, we will see why an ABM is a suitable method for this project.

Agent-based models are a good candidate for modeling the spread and effects of misinformation on social media. This is because ABMs allow for heterogeneous and autonomous agents, which can have memory, adapt over time and display path-dependence (Bonabeau, 2002). Furthermore, ABMs are made to explore the effect of micro-level interactions of autonomous agents on the emerging macro-level patterns of the whole system (Axelrod, 1997). Because real social media users are heterogeneous (Lewandowsky et al., 2020) and autonomous[3], ABMs tend to be a suitable tool for approaching research questions in this field.

Before we start into the methods and findings of the performed literature review, I want to note that the review has been performed before the start of the thesis project. In the following, I summarize the essence of that review.

### 1.2.1. Methods of the Literature Review

The approach to the literature review had two phases – 1) a systematic search within Scopus, and 2) a forward and backward reference search. The process focuses on the niche of work that aims to understand misinformation on social media better by using opinion dynamic models, especially agent-based models.

The final selection of papers is thoroughly analyzed for four themes which they have in common. The two themes which are most important for this project are discussed below[4].

[2] For more information on the uncertainty levels, see appendix Section B.1.

[3] Autonomous agents are agents that act out of their own choice. Thus, each specific action is not predictable.

[4] More detailed information about finding and analyzing the related work can be found in appendix Section A.1.

### 1.2.2. Findings

The two themes that are most important for this project are theme 1 (Belief Representation and Belief Update), and theme 4 (Handling of Uncertainty). These two themes are discussed below[5]. After reading and summarizing the collected papers, common themes are identified and used as column headers within Table A.2. The themes judged as most relevant are described in the next subsections.

**Theme 1: Belief Representation & Belief Update**

Figure 1.1 provides a visual overview of the found papers with respect to belief representation and belief update. The belief representations are either discrete or continuous. For discrete representations the number of different values are encoded on the x-axis of the graph. The 'belief update' axis is nominal and roughly ordered by how realistic the belief update function seems intuitively.

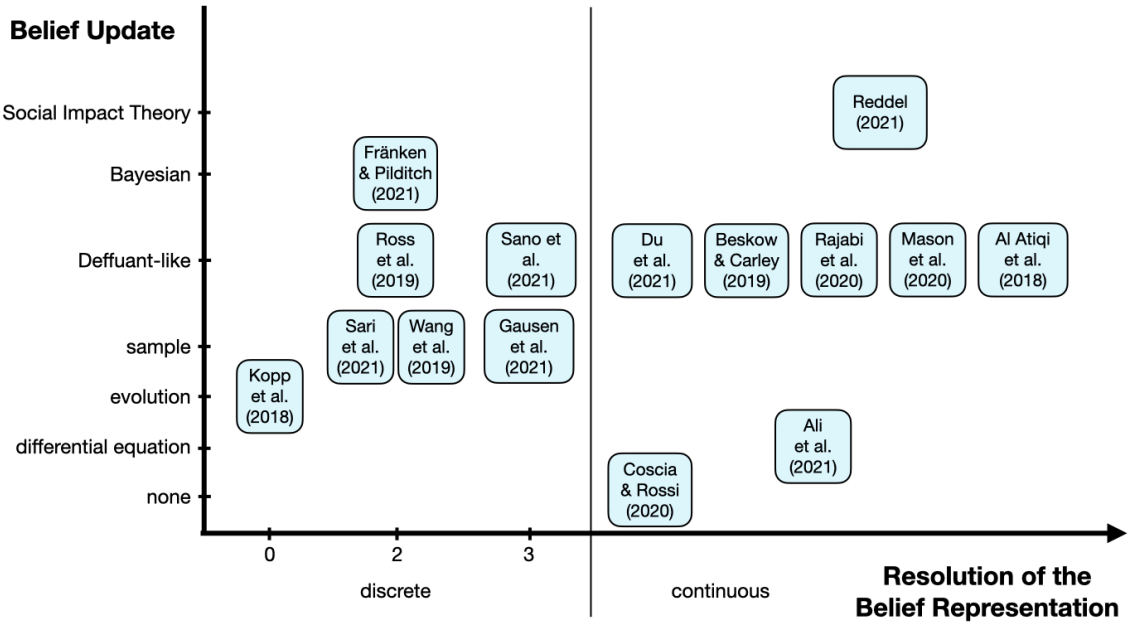



**Figure 1.1:** Belief Representation and Belief Update

For discrete *Belief Representations*, the number of values that any one belief can take are indicated in parentheses. For Kopp et al., 2018, it is 0 because beliefs are not explicitly, but by agent characteristics which are passed on via evolution.

The resolution of the belief representation can either be discrete or continuous. A continuous belief representation provides a more detailed sense of the belief distribution and the changes of beliefs during the model run. Furthermore, the model results can still be aggregated in various ways to make sense of them. However, a continuous representation might lead to higher computational expenses during the model runs. Most discrete representations of these papers leave the agent two different positions regarding each belief – i.e., the belief that a statement is true or that it is false. While discrete representations tend to be more computationally feasible, aggregation happens early and implicitly. Such early aggregation renders different aggregations or disaggregation virtually impossible.

A majority of papers lies within the upper part of the graph, indicating a more realistic belief update function. However, even the realism of most of these methods is likely rather limited. The Deffuant-like belief update category is a relatively broad class. The ways the included papers update the agents' beliefs can be boiled down to the Deffuant belief update (Deffuant et al., 2000), in which beliefs are always updated by a fix percentage towards the newly incoming information. The Bayesian update

[5]The details about the remaining themes can be found in appendix Section A.2.

is more detailed. However, it assumes that all agents are updating their beliefs in a perfectly rational manner. Neither the classic Deffuant update nor the Bayesian update takes into account social factors like how close the agent is to the source of the information or how much the agent trusts the source of information. In my previous internship (referred to as (Reddel, 2021)), I aimed to make progress at resolving these problems. How humans' real belief updating behavior is best represented introduces not only uncertainty around which parameters should be used, but also substantial structural uncertainty, i.e., what structure or what kind of function is most suitable.

**Theme 4: Handling of Uncertainty**

The handling of uncertainty is often underspecified. For many papers, it could not be found how often the model was run and with what parameters. Some, however, did mention that they averaged 10 runs (e.g., Mason et al., 2020) or that they averaged 50 runs (e.g., Coscia and Rossi, 2020). Structural uncertainty was not explored in any of these papers. Furthermore, none of the papers used their models to acknowledge and embrace uncertainty extensively with a DMDU approach.

Evaluations of policies is rare in the field of misinformation modeling (Gausen et al., 2021). Optimizing for policies requires the evaluation of policies and is even less common. The same is the case for work which thoroughly incorporates uncertainties (especially deep uncertainties). The main approach seems to be 1) selecting one set of assumptions, 2) model and simulate with that set of assumptions (potentially with a few runs to account for stochastics, i.e., for shallow uncertainties), 3) use the results of the model as conclusions for the real world. This approach is what may also be called *predictive* modeling. This approach works very well for systems that can be easily observed and that do not include higher levels of uncertainty (Kwakkel and Pruyt, 2013). Complex socio-technical systems, however, include a large number of components which cannot be easily observed. Furthermore, socio-technical systems include deep uncertainties. Under such conditions, the use of predictive modeling can be misleading. Some might argue that models cannot be useful under such circumstances (Cockerill, 2007). However, the set of techniques jointly referred to as Exploratory Modeling and Analysis (EMA) has been developed by the RAND Corporation for exactly such contexts. The central idea behind EMA is to use computer modeling to systematically explore the consequences of various kinds of uncertainties.

## 1.3. Deduced Research Objective

We have seen that despite the stark societal relevance of how we will choose to tackle the challenge of misinformation online, there is currently a lack of both, 1) research which explores apparent structural (and parametric) uncertainties, and 2) research which evaluates policies around misinformation on social media (especially research that would take into account multiple objectives of various stakeholders). I hypothesize that applying DMDU methods in our field of interest could contribute to both of these gaps. Consequently, I explore a method for handling this structural uncertainty while evaluating policies based on multiple objectives. The intended goal of this project is to be a use-case for showcasing the value that methods from the field of DMDU could have for the field of misinformation focused ABMs. Aiming to contribute to this gap is important, currently neglected, and – thanks to DMDU methods and exploratory modeling – also likely tractable[6].

[6] In appendix Section A.3, I elaborate on the rationale behind the deduced research gap, as well as why I use the common BDI-assumption.

### 1.3.1. Main Research Question

The chosen concrete and feasible research question for the deduced research objective is be the following:

**Main Research Question**

How does the consideration of **structural uncertainty** with respect to the choice between different **belief update functions** influence the resulting **Pareto-optimal policies** and their performance?

After having gathered more insights on this research question, we know more about whether it is relevant which belief update function we pick for our models and whether that choice significantly influences which policies seem the most promising. Such insights can help us to understand the complex system and the potential policies better, empowering us to tackle this grand challenge in a more evidence-based, realistic, and successful way.

### 1.3.2. Sub-Questions

To approach this main research question step by step, I use multiple sub-questions. The remainder of this section provides the high-level context of each of these sub-questions, as well as the information of which sections of this report are aimed to answering.

**Sub-Question 1**

Which **belief update functions** are commonly used in previous work?

Within the collection of the surfaced papers, there are various different belief update functions. It might seem to be the best choice to include each one of these functions. However, on second thought, it becomes clear that this would expand the scope of the project too far to be feasible as a master thesis project. Consequently, a suitable set of belief update functions needs to be selected. The details of how I answer this sub-question can be found in Section C.3.2.

**Sub-Question 2**

Which **objectives** are likely valuable for decision-makers and thus would be valuable to be included in the problem formulation?

For challenges in complex socio-technical systems, there are usually a multitude of diverse stakeholders, each having their set of objectives. Furthermore, each type of stakeholder is a simplified representative for a diverse set of people, who naturally differ again in what they care about. Consequently, it is not feasible to include the exhaustive list of objectives for such a diverse collection of people. What however can be done is aiming to include some of the major objectives for each of the prototypical stakeholders. The reasoning behind my final selection of included objectives can be found in Section C.4.11.

**Sub-Question 3**

What is an appropriate combination of **DMDU methods and techniques** to explore the structural uncertainty about belief update functions?

In the field of exploratory modeling and analysis, there is a multitude of approaches to explore uncertainties. The choice depends on the problem at hand and the available resources such as time and compute. How I choose the approach for this project is reported in Section 2.1.

**Sub-Question 4**

How do the found **Pareto-optimal policies** of the individual belief update functions differ with respect to lever-values and outcome-values?

The final and likely most interesting sub-question revolves around exploring the results of the performed computational experiments. Investigating and aggregating the Pareto-optimal policies that each of the belief update functions lead to is a crucial step before being able to compare between belief update functions. This can be found in Chapter 4. With the insights from these sub-questions, we are prepared to evaluate to what extent the different belief update functions lead to distinctly different types of Pareto-optimal policies. Therefore, we can evaluate whether the chosen focus speaks for or against utilizing the DMDU approach in the field misinformation.

# 2

# Decision-Making under Deep Uncertainty

The goal of this project is to explore the potential of DMDU methods for addressing the problem of structural uncertainty about the belief update functions in misinformation studies. In line with this goal, this chapter provides an overview of what different methods there are as well as an example of how a particular method can be selected. For the method which is selected for this project, I describe the mechanisms behind that method and provide an example of how the method itself could be adjusted to a particular use case.

## 2.1. Selecting the Method

As the main research question revolves around structural uncertainty – and thus around deep uncertainty – we want to use a DMDU (Decision-Making under Deep Uncertainty) approach. Various DMDU approaches exist. These include for instance the broad categories of Robust Decision-Making (RDM), Dynamic Adaptive Planning (DAP), Dynamic Adaptive Policy Pathways (DAPP), and Info-Gap Decision Theory (IG) (Marchau et al., 2019). For this project, the RDM category seems to be most suitable for a few reasons.

Firstly, RDM includes the approaches where multiple objectives can be taken into account (e.g., MORDM (Multi-Objective Robust Decision-Making), MS-MORDM (Multi-Scenario MORDM), and MORO (Multi-Objective Robust Optimization)). This is helpful because the aim behind the research question is to evaluate policies based on a multitude of objectives. Secondly, the parametric uncertainties in the model are not typical examples of the category of uncertainties that make DAP or DAPP a suitable choice. That category is defined by uncertainties which develop over time or which will decrease just by time passing by. Thirdly, a goal behind the research question is to evaluate numerous different policy candidates. This is perfectly suitable for RDM. At the same time, IG seems to focus more on choosing between a small number of alternatives and thus does not seem to have the best fit here. Fourthly, RDM is commonly used for consensus building between different stakeholders. This is also an important issue in this grand challenge and a recurring topic in this project. Lastly, I personally have most experience with the RDM approach. While this is only a bonus, it still should be helpful to increase the quality of the project. These points together made the choice for the RDM framework a straight-forward one.

Within the RDM approach, the MORDM method was selected. Many RDM methods enable the inclusion of multiple objectives. Thus, the main driver behind the choice for MORDM was the compu-

tational demand of the method. The computational requirements of MORDM are substantially lower than those of for instance MS-MORDM and especially than those of MORO. Consequently, MORDM it was selected.

## 2.2. The Method

The general working of the MORDM method is summarized in Figure 2.1. It includes the four depicted steps. The full benefits of MORDM are unfolding if the method is applied iteratively. In the following, these four steps are described.

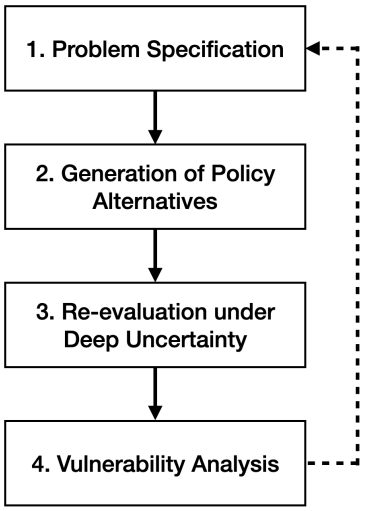



**Figure 2.1:** MORDM Overview

The graphic is based on (Kasprzyk et al., 2013)

### Problem Specification

The problem specification, also called problem formulation, is the step in which we aim to encode the problem. We want to specify what we include in the model and how we include it. This is typically done in the XLRM framework (Lempert, 2003; Lempert et al., 2006). The framework is visualized in Figure 2.2.

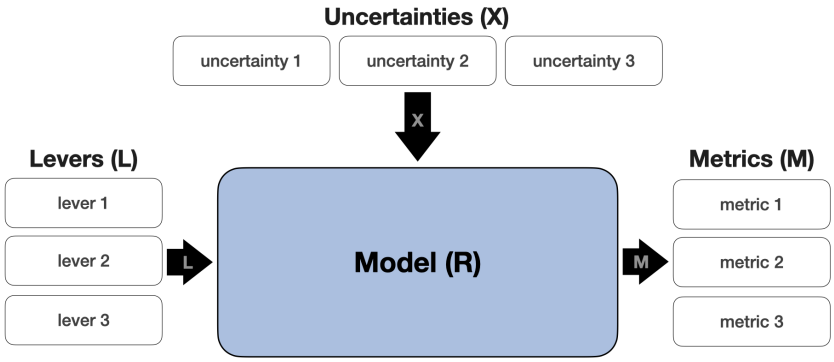



**Figure 2.2:** XLRM Overview

General overview of the XLRM framework. The number of included uncertainties, levers, and metrics depends on the specific model.

Each part of the XLRM acronym represents one component of the problem formulation. ”X” stands

for uncertainties. These are factors that are assumed to influence the model, while also being out of control of the decision-makers and not accurately known. ”L” represents the levers, i.e., the actions that decision-makers could decide to take to influence the system. These could be encoded as either binary, as categorical, or as continuous. ”R” stands for the relationships within the model. In other words, it refers to how the various inputs (i.e., X and L) are mapped to the outputs (i.e., M). Consequently, R is what we might intuitively think of as ’the model’. ”M” symbolizes the metrics that are included. These are the objectives that can be investigated to estimate success. Each metric is the encoded version of something some stakeholders care about. When all components of the XLRM framework are encoded, the problem which we aim to tackle is specified. Next, one can generate potential policies that can be evaluated and compared.

**Generation of Policy Alternatives**

To generate policy alternatives, usually a Multi-Objective Evolutionary Algorithm (MOEA) is used (e.g., epsilon-NSGA-II (Reed & Devireddy, 2004)). Provided a reference scenario (i.e., a valuation of the included uncertainties), the algorithm runs the model with randomly sampled policies (i.e., combinations of lever values). The success of the policy is then recorded and judged by the resulting outcome values (i.e., metric values). By strategically testing policies and saving the policies that are non-dominated by other policies, a set of Pareto-optimal policies is found and returned.

**Re-evaluation under deep uncertainty**

However, as all potential policies are only judged in a single scenario (i.e., in the reference scenario), it is unclear how well they would perform in other scenarios. To gather more evidence for judging their overall performance, the found policy candidates are *reevaluated under deep uncertainty*. What this means is that the Pareto-optimal policies are tested on a bigger set of other scenarios. As there likely are scenarios that are more ”difficult” – i.e., scenarios in which it is more difficult for policies to result in better metric values, it is important to test all policies on the same set of scenarios. This ensures that we can trust their comparative performances more because we know that there are no spurious results caused by situations in which some policy candidates are evaluated on a more ”difficult” set of scenarios than others (i.e., a set that includes more ”difficult” scenarios). After the re-evaluation under deep uncertainty, each policy is associated with many metric evaluations. To make sense of all this data and to facilitate comparisons between the performance of the policies, it is useful to aggregate the data. We often care about avoiding worst-case situations. Therefore, the *robustness* of policies is an important factor. A policy is said to be *robust* if it performs well in many different scenarios. There are numerous measures of robustness and the choice between them depends on what we care about with respect to the specific metrics. Furthermore, it is good practice to make use of multiple different robustness measures in order to get a clearer picture of the policies’ performance. With the addition of robustness information, it is possible to select policy candidates that seem the most promising based on the computational experiments. At the end of this step, we have a set of policies which seem promising with respect to their performance and robustness.

**Vulnerability Analysis**

Because also the number of scenarios in which we have evaluated the policies in the third step is likely quite small in comparison with all possible scenarios, it is likely that there are still some scenarios under which we are still vulnerable. The goal of vulnerability analysis is to discover exactly those vulnerable scenarios. We want to find scenarios in which the current Pareto-optimal policies fall short in performance. If we find such scenarios, we can do another round of MORDM in the hope of finding policies that would also cover those scenarios.

## 2.3. Applying the Method in This Project

### 2.3.1. Method Customization

As mentioned previously, the second step of a MORDM cycle usually uses a MOEA to generate policy candidates. However, for the levers included in this project, it can plausibly be assumed that the effect of each lever scales rather smoothly[1]. Our goal is to get a rough estimation of how the structural uncertainty in the belief update function makes a difference as well as to gain insights in the overall trade-offs and emergent patterns. Thus, for these purposes, we do not need to explore the infinite number of possible policies. A lower lever resolution is sufficient. Consequently, a choice is made for allowing levers either six or eleven possible values and to use a full-factorial design to arrive at the policies. Only one iteration is performed within this project (i.e., steps 1, 2, & 3). With the third step (reevaluation under deep uncertainty) not reaching the use of robustness metrics. In this third step, 50 scenarios are sampled and saved. Each of the scenarios os paired with each of the selected Pareto-optimal policies. And as previously, each computational experiment (i.e., each such combination) is run 30 times before aggregating over the outcomes of the 30 runs.

### 2.3.2. Method Input Preparation

As noted in the more general section above, the problem specification requires specifying each component of XLRM. This can be found in Chapter 3 which specifies the model (”R”), including the incorporated uncertainties (”X”), levers (”L”), and metrics (”M”). For the Generation of Policy Alternatives, a couple of inputs and some additional information is required. More specifically, four more things are needed:

1. the lever values, i.e., the values that each lever might take in the full-factorial design (to generate the policy alternatives)
2. the reference scenario (to evaluate the policies alternatives)
3. the number of replications for each experiment (to guarantee that the evaluation is not just a fluke caused by how the stochasticity played out during each individual run)
4. the epsilon values (to select a small, yet diverse set of Pareto-optimal policies)

In the following, a high-level overview is provided of how the choices behind these additional points were made.

#### Levers Values

The exact values that each lever might take for the full-factorial design are selected by balancing the need for computation with the additional information gained. It is decided that the values should be evenly spaced to be well-distributed over the lever dimension. Furthermore, the distance between a value and the next should be something that makes it easy to intuitively think about. As such, steps of 10% were chosen. The non-threshold levers can span the whole space. This includes the extremes of 0 and 100, and thus leads to eleven possible values. Threshold levers are related to the probability that a post is false. To reduce computational needs, I made the assumption that decision-makers would be less interested in having very strong values here. For instance, that they would not want to delete posts that are more likely true than false. As such, the threshold values only go up until the 50% mark. This results in six possible values.

#### Reference Scenario

The reference scenario for evaluating the policies is selected by aiming to find a ”median” scenario. What that exactly means will become clear in the remainder of this section. When starting to think about the choice of the reference scenario, it could make intuitive sense to pick a scenario that is

[1] For more about this assumption, see Section 5.2.

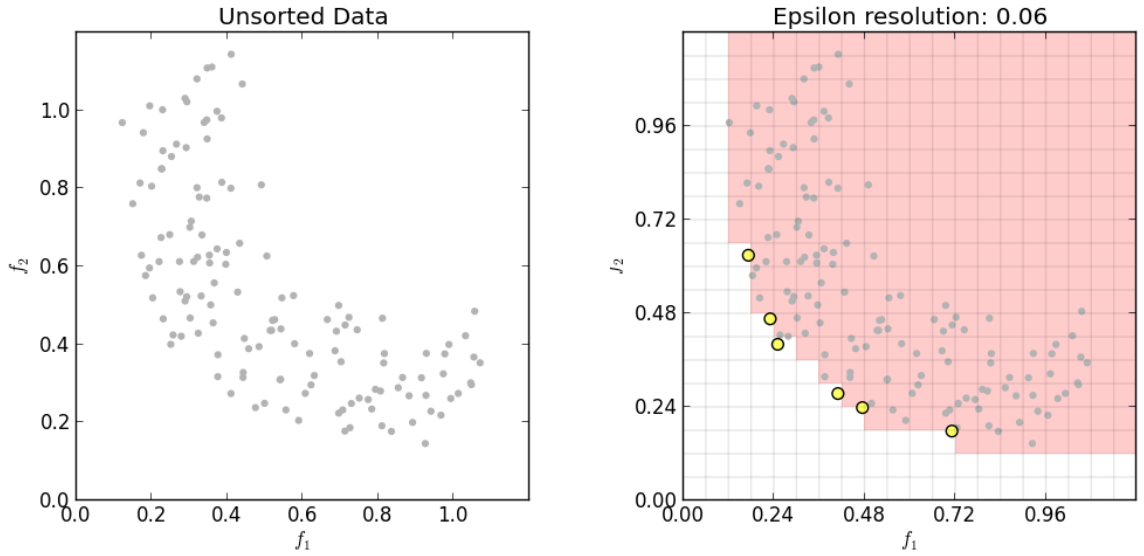
especially ”difficult”. We might hope that policies which can perform well in that scenario will have an easy time performing well on ”easier” scenarios. However, such ”difficult” scenarios might consist of extreme values whose real-world counterparts are not very likely to occur. And it might very well be the case that the optimal policy for such a worst-case scenario would be very different from what would be optimal in a less extreme scenario. As the project is limited in scope and will not get to a second round or MORDM, I aim to pick a scenario that would not be particularly ”difficult” or particularly ”easy” with respect to any of the metrics. With that goal in mind, an open exploration of the uncertainty space is performed. For this, the lever values are kept static at the ”do nothing” equivalent (also called the ”no policy” policy), and the uncertainty values are varied to explore the uncertainty space by testing various scenarios. The data gathered from this open exploration is then investigated to find the most ”median” scenario. For each metric, the median value is calculated. Then the policy is selected which is closest to the median values of all metrics.

**Number of Replications**

To arrive at the number of replications for each experiment, a model can be run with a number of different seeds. Then, it can be compared how the metric results converge when aggregating different numbers of runs. Finally, a suitable balance between computational costs and accuracy can be found.

**Epsilon Values**

To select a diverse set of Pareto-optimal policies from the large set of evaluated policies, epsilon values are used. The epsilon values define the resolution with which Pareto-optimal policies are chosen (out of the set of all evaluated policies). An example with only two metrics is presented in Figure 2.3.




**(a)** unsorted **(b)** sorted

**Figure 2.3:** Example Pareto Sorting

A two-dimensional example of how the Pareto-optimal policies are selected from all evaluated policies. In this example, the metrics of both axes are to be minimized. Each point represents one evaluated policy. The point’s location relative to each axis represents the policy’s outcome value. Graphic is from Woodruff and Herman, 2013.

Each metric dimension is sliced into segments. Each segment is of the size of the metric’s epsilon value. Together, the epsilon values of all metrics span a grid of subspaces. Each subspace has as many dimensions as there are metrics. Consequently, the simple example has two-dimensional subspaces,

and the project has 5-dimensional subspaces. For each subspace, the best performing policy is chosen. The performance is judged by the combination of all metrics. To arrive at a sets of Pareto-policies whose size is neither too restrictive nor too broad, we want to pick epsilon values that are neither too big nor too small for the final metric ranges. To select the five epsilon values, I aim to estimate a broad range of possible values for each dimension. With that goal, I openly explore what metric values would result from a broad range of policies and scenarios. 50 randomly sampled policies are evaluated in randomly sampled scenarios[2]. For each metric, the results from this open exploration are plotted in a violin plot to see their distribution Then the data is investigated to arrive at epsilon values that could be suitable. More details of this process are provided in the code base.

### 2.3.3. Additional Policy Selection (Single-Objective)

As is common when applying the method of MORDM, I select Pareto-optimal policies by considering multiple objectives. Using multiple objectives is useful for selecting policies that perform well in the eyes of different stakeholders. Yet, within the field of opinion dynamics and ABMs modeling misinformation on social media, usually just one objective is used for evaluation. I want to additionally provide an impression of the kind of policies that would be selected as a consequence of these current practices. To do this, I reuse the gathered data of the over 26.000 policies. Then, I simply rank the policies by the outcome of one main objective. For this objective, I use the belief-based metric of the number of agents above the belief threshold because it is the closest to what is commonly considered in the field of agent-based misinformation modeling. Then, I select those policies that share the highest performing value for this outcome as the promising or ”Pareto-optimal” set[3]. Which objectives are included for selecting the Pareto-optimal policies typically heavily influences which policies seem the most promising. Consequently, I expect that using a single objective, rather than all included ones, would change which policies are selected.

[2] Here, the uncertainty space and the lever space are combined to guarantee that the sampling via Latin-Hypercube Sampling has the best coverage.

[3] Pareto-optimal in quotation marks because when considering only one objective, it is not very useful to talk about Pareto-optimality (i.e., a non-dominated sort), rather than a conventional ranking based on one attribute.

# 3

# The MisinfoPy Model

The purpose of the model is to enable the testing to what extent the structural uncertainty in the agents' belief update function influences which Pareto-optimal policies are found. The focus of this chapter is to provide an overview of what is in the model as well as the essence of the reasoning behind model choices. The extended rationale behind the choices within the model, as well as details about the implementation, is provided in Appendix C. The noteworthy exception to this approach is when we get to the belief updating functions. Here, I provide also some more rationale and details because the structural uncertainty of these functions is central to this project.

This chapter is organized according to the previously mentioned XLRM framework. An overview of the model from the XLRM perspective is provided in Figure 3.1. Three different belief update functions are considered in this project. Therefore, three alternative models are built – each model solely differing from the others by their belief update function. 'R' represents the relations within the model and consequently is the main chunk of the model. Uncertainties (X), levers (L), and metrics (M) concern the inputs and outputs to the model.

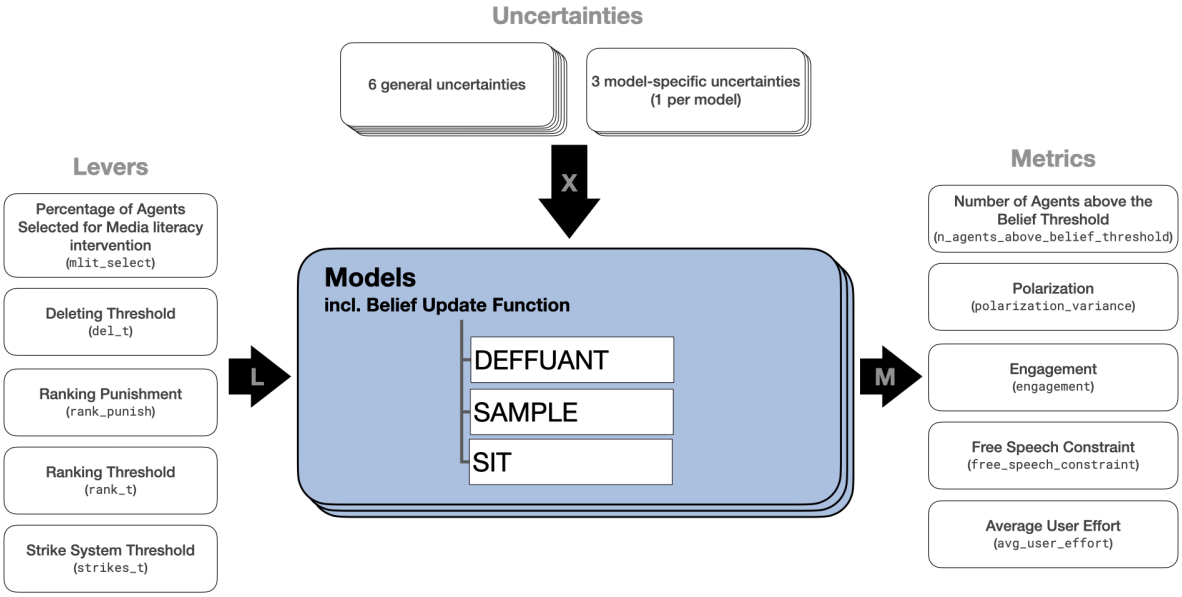



**Figure 3.1:** XLRM MisinfoPy Model

# 3.1. Model Description

## 3.1.1. Relations (R)

Before we dive into all the separate parts, a visual overview of what is happening over the course of a model run is provided in Figure 3.2.

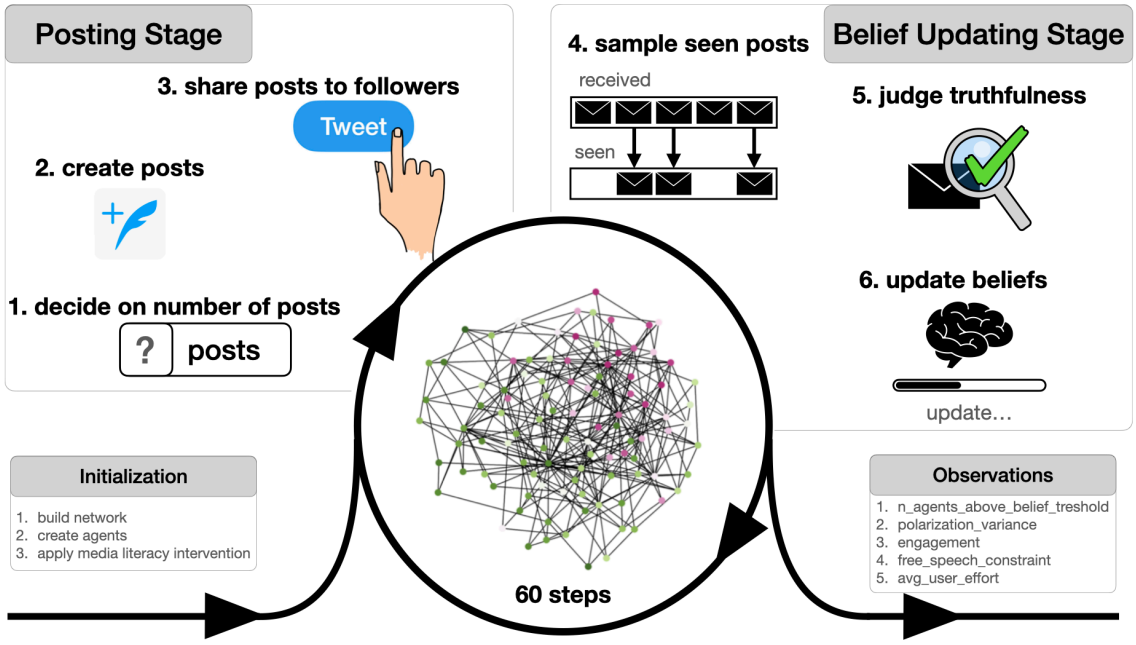



**Figure 3.2:** Visual Overview over a Model Run

### Space

The environment of agent includes everything that is both, outside of the specific agent and within the model. This means that other agents are part of the agent's environment. Agents gather information from their environment and provide information to their environment. A suitable environmental structure of a social media platform is a network in which nodes represent agents and edges represent the connections between agents (in this case, agents 'following' each other). Many different types of networks exist. The model builds *Barabási-Albert* networks because this type of network achieves high clustering (Barabási & Albert, 1999) and are therefore similar to real-world social media networks.

### Time

One model step represents one day passing by in the real world. The model can be run for a specified number of steps, but for the performed experiments the duration of one model run is consistently kept at 60 steps (i.e., representing roughly two months in the real world).

### Agents

As one would expect for the focus of this model, each agent can have beliefs. For the simple case study that I am aiming to do with this project, we focus on one single belief. This belief revolves around the safety of COVID-19 vaccines, also referred to as the *VAX statement* (see box below). The agents' belief refers to their agreement or disagreement with the VAX statement. An agents belief can range from 0 (total disagreement), up to 100 (total agreement). If an agent's belief is at 50, that agent is

completely uncertain about the statement. For the purposes of this project, this statement is assumed to be one that is objectively and certainly ”true”.

**VAX statement**

*”For most people, it is safer to get their COVID-19 vaccine jabs than it is to decline them (and consequently face COVID-19 encounters without the inoculation).”*

Next to their belief, agents also have a set of other agents they are following (`following`) and a set of agents that follow them (`followers`). Furthermore, each agent has attributes that specify how vocal they are, i.e., how frequently they want to post. Additionally, agents have an attribute of media literacy, which can either take the value LOW or HIGH. There are two types of agents – normal users and disinformers. The main difference between them is that disinformer agents are more vocal, are in strong disagreement with the VAX statement, and do not update their belief within the represented time frame of two months.

**Stages**

Each step of the model is split into two stages – the posting stage and the belief updating stage. While this is not a one-to-one representation of the real world, I chose this setup because it enables avoiding artifacts that are far from reality. If these two behaviors would not be split into separate stages, it would for instance cause that the agent that is first assigned to post and update their beliefs could not update its beliefs because it certainly would not yet have received any posts within this model step. Only for the very last agent it could be assumed that they certainly have received the posts from all the agents whom the agent is following. Both stages are explained here on a high level.

*Posting Stage*

In the posting stage, the entities of both types decide how many posts they would like to post, create the posts, and post them to their followers. In this stage, agents first decide on how many posts they want to create and post. Then, they create the posts and share them to their followers. How many posts an agent posts in a given step is modeled as a function of their natural vocality and how extreme its belief is. The more extreme their belief, the more they tend to post. Each post includes a belief that is represented in it. This value is based on the agent who creates the post. An agent can be assumed to post content that is roughly in line with their current beliefs (e.g., Ross et al., 2019).

*Belief Updating Stage*

In the belief updating stage, all ’Normal User’ agents are seeing some of the posts that they received, decide for each seen post on whether they judge the post as truthful, and if judged as truthful, the agents update their belief based on the post. For this updating, the model’s belief updating function is used.

To decide which belief update functions to include in the project, two aspects are taken into consideration. Firstly, some belief update functions should be those that are commonly used in previous work. Secondly, at least one belief update function should be one that is more deeply grounded in the science of other relevant fields such as social psychology and cognitive science. To strive for the first consideration, the two belief update functions that are the most commonly used in the papers that the literature review surfaced are selected. These two are the Deffuant-like belief update function (in the ’DEFFUANT model’) and the belief update function based on sampling (in the ’SAMPLE model’). And to accomplish also the goal of the second consideration, the belief update function that is conceptualized and implemented in the previous internship is utilized. This belief update function is based on Social

Impact Theory and adjusted for the context of social media. We refer to it as the 'SIT model'. The essence of these three ways of modeling how people change their minds on social media is provided in Figure 3.3.

**Belief Update Functions**

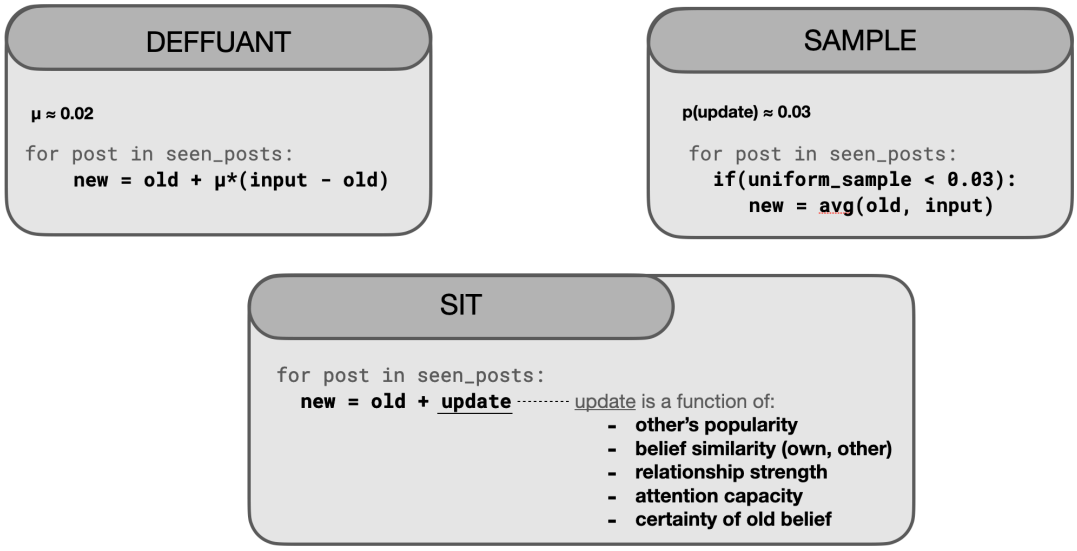



**Figure 3.3:** Selected Belief Update Functions

'new' is the new belief resulting from the update.
'old' is the previous belief of the agent.
'input' is the belief represented in the seen post.
'$\mu$' is the percentage by which the agent updates towards 'input'.
'p(update)' is the probability that the agent will update.
'uniform_sample' is a number between 0 and 1 drawn from a uniform distribution.
'avg(x, y)' represents a function that averages x and y.

In the DEFFUANT model, beliefs are always updated by a fix percentage towards the newly incoming information (i.e., the belief represented in the post). More specifically, to determine the new belief, the difference between the old belief and the new information is calculated. Then, in order to arrive at the size of the update, the update parameter $\mu$ is multiplied with that difference. Finally, the update is added to the old belief to arrive at the new belief. For instance, if $\mu$ is 0, the agent will not change its belief. If $\mu$ is 0.5, the new belief will be the average over the old belief and the new information. If $\mu$ is 1.0, the agent will adjust its belief to be the same as the new information it received.

The SAMPLE model samples whether an belief update happens or not. Agents update their beliefs only rarely. If an update happens, the new belief is the average between the previous belief and the newly incoming information.

In the SIT model, the agents update every time they see a post. How much they update towards the post belief depends on the context. Various aspects about themselves, the agent who posted the post, and their relationship to that agent influence how strongly they update. Social Impact Theory (SIT) aims to describe how individuals are influenced by others (Latané, 1981). The impact on an individual is described as the product of three components: strength, immediacy, and number of sources.

Strength represents the persuasiveness of the other (i.e., of the agent who is the source of the post). This persuasiveness is usually related to aspects such as age, gender, appearance, and perceived intelligence of the other perceived through the lens of the receiver. In the context of social media, this kind of status and persuasiveness seems to be strongly related to the number of followers an account has. Partially because more popular people will likely have more followers, and partly because people

that have a large base of followers are perceived as more popular. The persuasiveness and also the feeling of belonging to the same group. It is assumed to be based on whether the communicator is perceived to have similar beliefs to one's own. Consequently, within the SIT model, *strength* is a combination of the relative number of followers the communicator has and the belief-similarity between the agent's own belief and the estimated belief of the other agent. The other's beliefs are estimated by looking at the communicator's current and prior posts.

The immediacy component can be split into three parts – physical, temporal, and social immediacy (Chang et al., 2018). In the SIT model, physical and temporal immediacy are excluded because of the social media context and the chosen time resolution, respectively. Social immediacy is therefore the determining factor of the immediacy component and is itself represented by the relationship strength between the two agents.

The number of sources represents the number of people that are influencing an individual. If there are for instance only five people sending information to one person, each of them will have a bigger influence on the beliefs of that person than if one thousand people are aiming to influence that individual. In the SIT model, it is the case that the more other agents a user is following, the less they will update their beliefs based on each one of them.

The three components (strength, immediacy, number of sources) are combined and rescaled. The rescaling guarantees for instance that agents that are rather uncertain regarding the VAX statement update more than agents who are very certain.

### 3.1.2. Levers (L)

The included levers relate to interventions that are considered in decision-maker document (such as the European Commission's 'Strengthening the Code of Practice on Disinformation' (European Commission, 2021b)) and/or to policies that are currently in place at social media platforms like Twitter.

#### Media Literacy Intervention

Higher media literacy seems to help people in recognizing false information on COVID-19 as such and in avoiding updating their beliefs based on that false information (Austin et al., 2021). The lever of `mlit_select` represents the percentage of the population that is empowered by such an initiative. When an agent participates in the intervention, its media literacy is set to HIGH. This means that in this simple implementation, agents whose media literacy is already HIGH cannot further benefit from the intervention. With HIGH media literacy, agents have an 80% probability of correctly classifying a post as truthful (i.e., True) or not truthful (i.e., False). (Each post has a `groundtruth` label of either True or False.) Agents with LOW media literacy judge every post as truthful and are thus more vulnerable to misinformation.

#### Deleting Intervention

The deleting intervention works via the deleting threshold lever called `delete_t`. If a post's probability of being true is lower than the threshold, then it may be deleted. However, it is to note that not all misinformation on social media platforms is detected and acted upon. On Twitter, 41% of content that is fact-checked as false stays on Twitter (Courchesne et al., 2021). Consequently, also in the model, there is only a 41% chance that posts with a ground-truth label of 'False' are actually detected and acted upon. This holds for interventions of downranking content, deleting content, and for the strike system. A higher value for `delete_t` makes a policy delete many more posts. In the extreme case of the threshold value being at the upper extreme (i.e, at 50.0), it means that even if there is only a 50:50 chance that the post is misinformation, it can be deleted. If a post is deleted, it is deleted before any other agent can be influenced by it.

**Ranking Intervention**
With the ranking intervention, the visibility of content that has been detected as misinformation can be reduced by a specified percentage (`rank_punish`). However, the previously mentioned detect-and-act percentage of 41% applies here as well. Furthermore, downranking may only happen if there is enough certainty that the post is really misinformation. This threshold lever `rank_t` works the same way as the deleting threshold and the threshold for the strikes system. If a post's probability of being true is lower than the threshold, then it may be downranked if it is detected.

**Strike System Intervention**
The strike system is modeled after the one that Twitter implemented during the COVID-19 pandemic (TwitterHelpCenter, 2021). When a post of an agent is detected[1] and judged as misinformation (with a certainty that fulfills the threshold of `strikes_t`) causes the agent who is the source of that post to receive one strike. There is a simple mapping between number of strikes received and time for which the agent is blocked and cannot post anything. Agents may be blocked for one day, one week, or until the end of the run (i.e., they are suspended).

### 3.1.3. Metrics (M)
The main information we want to get out of the computational experiments is information that lets us judge how well a policy has performed in the context of the current scenario. In the context of the previously discussed complex socio-technical system with various stakeholders, some of the main stakeholders have been identified. Subsequently, for each selected stakeholder, their main objectives in the context of handling misinformation on social media have been deduced. An overview of the selected stakeholders and their metrics is visible in Figure 3.4.

| Stakeholder | Main Objectives |
|---|---|
| **government** | - number of agents above a belief threshold (`n_agents_above_belief_threshold`)<br>- polarization (`polarization_variance`) |
| **social media platforms** | - engagement (`engagement`) |
| **users** | - free speech (`free_speech_constraint`)<br>- effort for the user (`avg_user_effort`) |

**Figure 3.4:** Stakeholders and Objectives

The bullet points in the right column show the objectives. The terms below and in brackets represent the corresponding variable name as used in the code.

The first two metrics summarize over the whole distribution of agents' beliefs. The metric of the number of agents above the belief threshold just counts how many agents are above the set belief threshold. The exact threshold that would be useful here is uncertain. As such, it is one of the uncertainties. Polarization of beliefs roughly refers to how divided or separated a group of people is (Bauer, 2019). In this project, I represented it as overall fitting and commonly used measure of variance. The remaining three metrics are fairly straightforward in simply summing the aspect of interest over the

[1] It is to note that the previously mentioned detect-and-act percentage of 41% applies here as well.

whole run of the simulation and averaging them either over the number of agents (`engagement` and `avg_user_effort`) or the number of total posts (`free_speech_constraint`).

To calculate the engagement metric, the number of posts that each agent has seen (over the whole model run) are added up to arrive at the total number of posts seen by the entire network. In order to make the interpretation of this metric more intuitive and independent of the network size, the total number of posts seen is divided by the number of agents. Consequently, engagement is the average number of posts seen by an agent over the course of the whole model run.

`free_speech_constraint` represents the portion of posts that are ”deleted” compared to the number of posts that the agents wanted to post. If a post is blocked because the agent is currently still blocked due having received too many strikes, each such blocked post is counted as 1 ”deleted” post. If a post is downranked by 20%, it counts as 0.2 posts ”deleted”. To arrive at the metric value, the amount of ”deleted” posts is calculated and divided by the total number of posts that agents wanted to post.

The metric of average user effort is influenced by the media literacy intervention. Effort is counted as *time required*. The other interventions cause no effort for the users as they do not require any extra action or time from the users. However, there are indirect effects from the other levers as well. For instance, if many posts are deleted, downranked, or blocked from being posted by the strike system, it means that there are fewer posts to see. Consequently, also the overall effort for judging whether posts are truthful decreases.

### 3.1.4. Uncertainties (X)

Six general uncertainties and three model specific uncertainties are included. The general uncertainties relate to the ratio between the two agent types, to density of the built network, to the vocality of both agent types, and to the previously mentioned belief threshold for one of the metrics. For each model there is one model-specific uncertainty. For the DEFFUANT model, that relates to the update size (see '$\mu$' in Figure 3.3). Here it is to note that the the whole range of tested valuations is rather low. This is the case because in previous literature, this parameter is typically higher (e.g., 10%) and agents update their beliefs once based on every connected agent. In the MisinfoPy model, agents can update once per post and thus multiple times per agent. Therefore, this reference value seems suitable. For the SAMPLE model, it is about the probability that the agent will update its belief. And for the SIT model, this uncertainty relates to the number of posts that is used to estimate how similar the other agent's belief is to the agent's own belief.

## 3.2. Sensitivity to Stochastics

How sensitive the model is to stochastics is a relevant consideration for our interpretation of the data resulting from running the model. If the model is very sensitive to stochastics, it is useful to rerun the model under the same conditions for a number of times and aggregate the resulting data. An overview of where stochastics play a role is depicted in Figure 3.5.

To analyze the model's sensitivity to stochastics, the model is run one hundred times under the very same conditions (with respect to the levers and the values of the included uncertainties). Each run however used a different random seed. In this way, the only differences that occur in the observed metrics are the result of the random fluctuations due to the sources of randomness in the above table. The next step consists in finding out how much the aggregations over different number of replications differ. To get the most information out of the one hundred runs, the resulting data is bootstrapped. For instance, if we want to know what an average for five runs could be, we could take the runs of the first five seeds and aggregate those. However, we could also take the runs of the last five seeds, which would lead to a slightly different aggregated outcome. We can now take many samples of five runs each and arrive at a confidence interval. As the number of runs in each sample increases, the confidence interval becomes smaller and the median over these aggregates is changing less. In this

**Where stochasticity is used:**

**Initialization** (see appendix)
- building network
  - agent characteristics (type, init belief, media literacy, ...)
  - where to attach agent (preferential attachment)
  - how strong relation to other connected agents (tie_strength)
- selecting agents for media literacy intervention

**Posting Stage**
– creating new posts
- number of posts
- post's belief and ground truth

– sharing new posts
- whether a post is detected as misinfo

**Belief Updating Stage**
- whether agent sees a received post
- judging truthfulness of posts
- specific to Belief Update Function:
  - DEFFUANT: no stochasticity
  - SAMPLE: sample whether update or not
  - SIT: no stochasticity

**Figure 3.5:** Stochastics of the Model

way, we can explore the trade-off between the computational costs of running more replications of the same computational experiment and the increased accuracy of the results. The results of this analysis are provided in the Results section Section 4.1.2.

## 3.3. Validity

The purpose of the models is to enable the testing to what extent the structural uncertainty in the agents' belief update function influences the Pareto-optimal policies. To assess the models' fit for purpose, it is crucial to examine two factors. Firstly, the models should be built in a way that isolates the effect of that structural uncertainty. This is arguably the case because the models only differ in the used belief update function, and are otherwise identical. Therefore, any difference in the resulting Pareto-optimal policies can only stem from this structural uncertainty (or from unaccounted influence of stochastics, but this has been taken care of by choosing a sufficiently high number of replications). Secondly, the models should have a good fit for the kinds of questions one would want to tackle with the model (independently of which belief update function is used). One such example question could be revolving around comparing the robustness of different policy candidates, where each policy candidate is a valuation of the included policy levers. To evaluate this aspect of the models' fit for purpose, we need to inspect model behavior. For this purpose, the three behavioral patterns described in Appendix Section C.1 would be useful. In short, these patterns predict the following model behaviors. The first expected pattern is that the more extreme posts there are (summed over the whole population and the whole run) the higher levels of engagement. The second pattern states that the stronger the lever activation is, the lower the percentage of false posts that are seen by agents. The third pattern revolves around the distribution of number of posts posted over the agents. It postulates that a small percentage of agents produces the majority of the posts. These patterns can make a case for the models' fit for purpose because they either are grounded in empirical insights that we have from real-world social media platforms, or because they are based on central assumptions within the model. Together, these three patterns could make a strong point for the models' validity with respect to the displayed behaviors.

# 4

# Results

This chapter follows the general structure of the DMDU chapter. It presents the results as well as some interpretations and explanations. The rationale behind this choice is that having the figures close by facilitates making sense of the results. In the Discussion chapter (Chapter 5), the main findings are discussed.

## 4.1. Problem Specification

### 4.1.1. Scenario Selection

The selected reference scenario consists of the following uncertainty valuations:

1. `n_edges`: 3
2. `ratio_normal_user`: 0.993
3. `high_media_lit`: 0.285
4. `mean_disinformer`: 8.0
5. `mean_normal_user`: 0
6. `belief_metric_threshold`: 80.0
7. `deffuant_mu`: 0.019
8. `sampling_p_update`: 0.028
9. `n_posts_estimate_similarity`: 5

The values are rounded to three digits after the comma. In this reference scenario, the network is a bit denser and has a high ratio of normal users to disinformers. Uncertainties (4) and (5) are both rather on the lower end while the belief metric threshold is at it's upper end. As for the uncertainties related to the individual belief update functions, the one related to the DEFFUANT model is rather a medial value. In the SAMPLE model, the probability of an agent changing its mind is relatively high. Lastly, the number of last posts considered to update beliefs in the SIT model is relatively low.

### 4.1.2. Seed Analysis

The process of the seed analysis is performed for all three models. The results for all models and all included metrics are plotted in Figure 4.1. When increasing the number of included runs to higher than 30, the confidence range is not decreasing substantially anymore and the median is usually not changing much anymore. Consequently, a replication count of 30 is judged as sufficient.

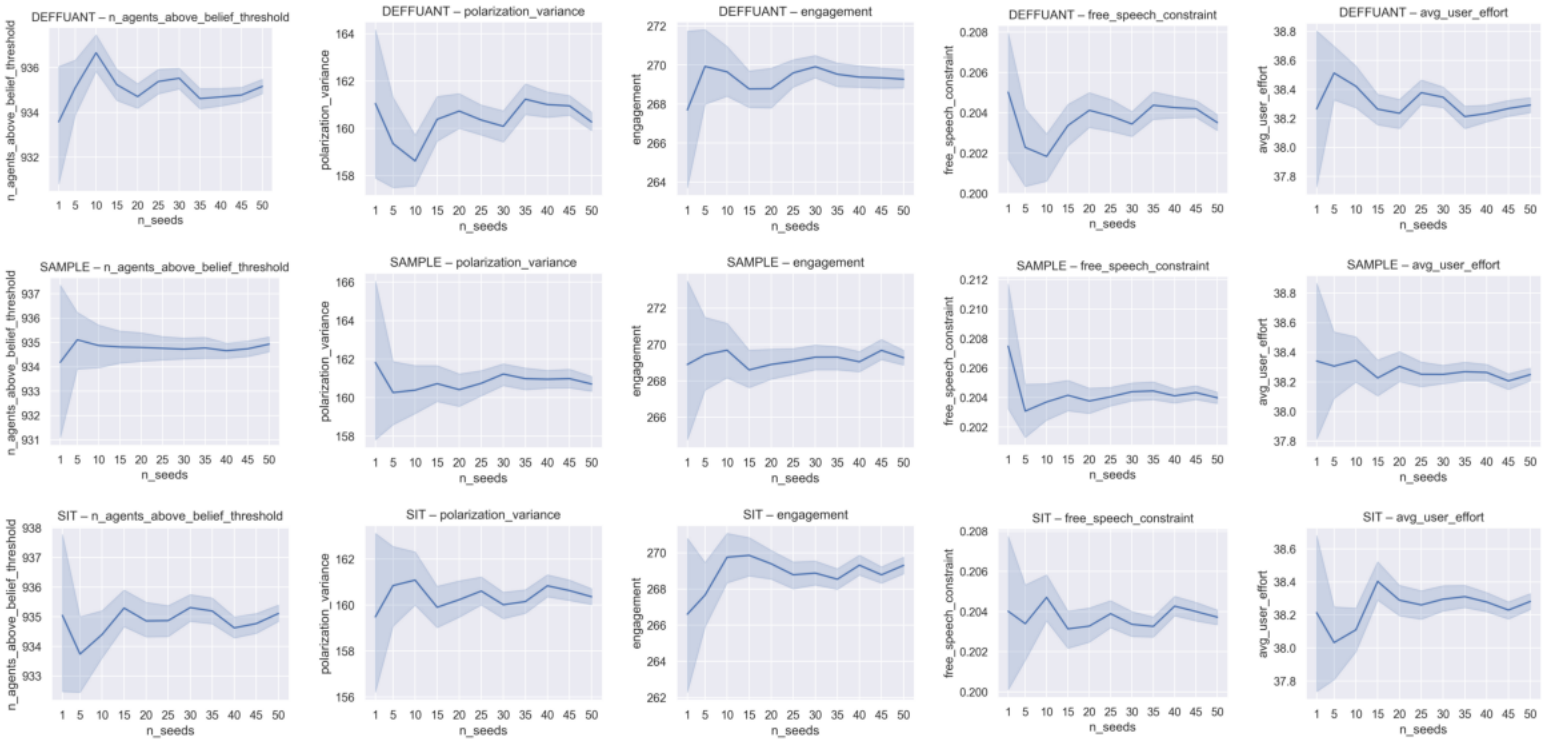



**Figure 4.1:** Seed Analysis

The results of the seed analysis are depicted for all models and all metrics. Each seed represents one way of how stochasticity can play out in the same computational experiment. The rows represent the models and the columns represent the metrics. The x-axis of each subplot specifies the number of seeds of which the results are averaged. The y-axis of each subplot refers to the metric of the subplot. The darker blue line is determined by the median of various sampled sets of runs (each set being of the respective `n_seeds` size). The lighter blue shade which wraps around the darker line specifies the 95% confidence interval of the averages.

## 4.2. Generation of Policy Alternatives

### 4.2.1. Epsilon Values

The resulting epsilon values are the following:

1. `n_agents_above_belief_threshold`: 2
2. `polarization_variance`: 2
3. `engagement`: 40
4. `free_speech_constraint`: 0.02
5. `avg_user_effort`: 1

In Figure 4.2 we see that in all models metric (1) `n_agents_above_belief_threshold` has a high density of high values. Thus, the open exploration runs, within which I sample policies randomly and do not filter them for performance, already many runs have relatively high values for the first metric. This indicates that the range of outcome values which is relevant for the later selected Pareto-optimal policies will likely be relatively small. To select enough policies that are high-performing on this metric, the chosen epsilon value of two is quite small (in comparison to the network size of 1000 agents). To a somewhat reduced extent, the analogous is the case for metrics (2) `polarization_variance`, (3) `engagement`, and (5) `avg_user_effort`. Therefore, also these values are relatively small when compared to the overall range of their resulting values. For metric (4) `free_speech_constraint`, the

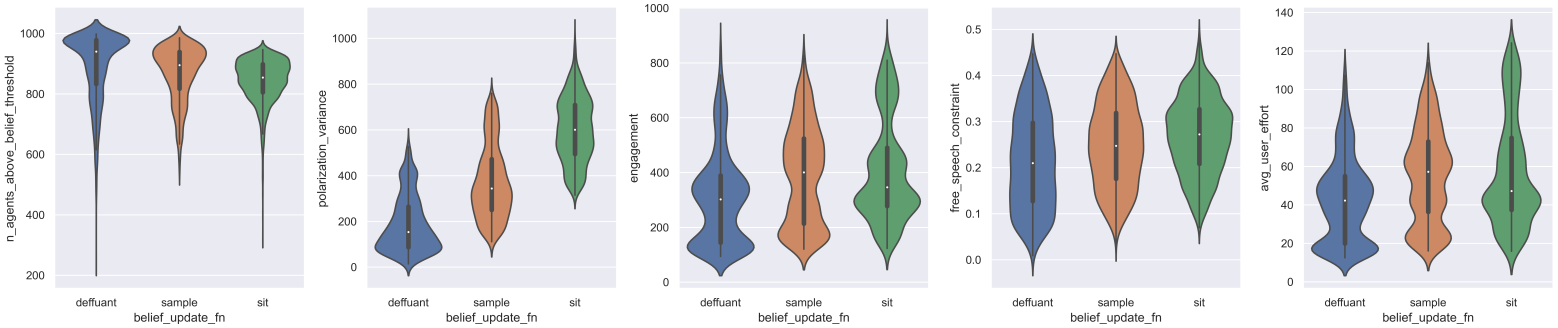



**Figure 4.2:** Visual Aid to Determine Epsilon Values

Each subplot depicts the values for one metric. For each model, the achieved values of a metric are depicted along the vertical dimension. The white dot in the middle depicts the median value. The darker area in the middle shows the interquartile range. The colorful area is a density plot, symmetrically depicted towards both sides. Engagement values refer to the average number of posts that an agent sees over the entirety of a run. The free speech constraint value refers to the percentage of posts deleted (e.g. 0.1 maps to 10%). The average user effort is the time that an agent spends over the entirety of a run (in minutes).

density of the values is a lot more evenly distributed. This implies a wider relevant range and a relatively large epsilon value (as compared to the common values). While the epsilon value of 0.02 is the smallest in absolute numbers, it is the relative size that counts. And in relative size, the 0.02 is a larger value. To exemplify let us compare metrics (1) `n_agents_above_belief_threshold` and (4) `free_speech_constraint`. In the DEFFUANT model for instance, the epsilon value of metric (1) leads to segments that each span 0.25% of the overall range, while the epsilon value of metric (4) implies segments that are as big as 4% of the whole range.

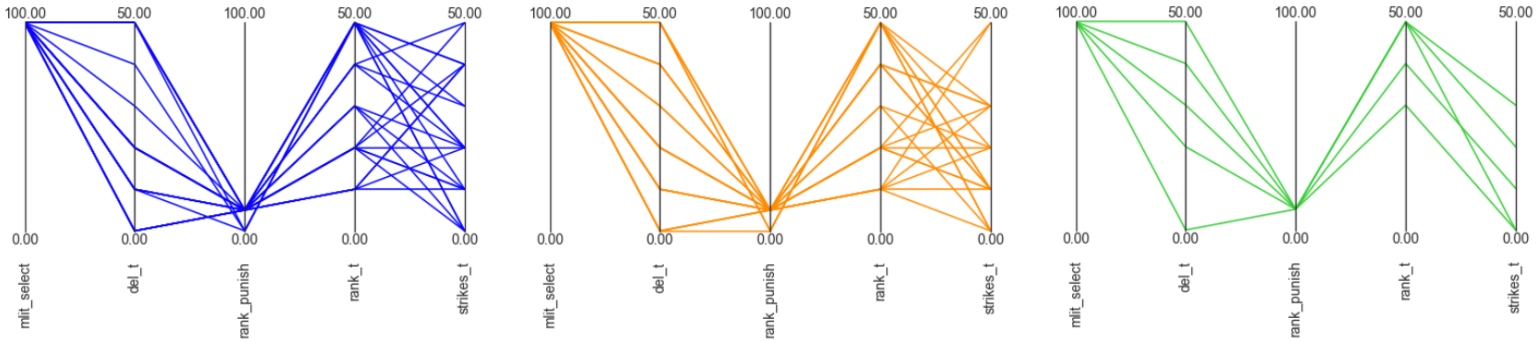
### 4.2.2. Pareto-optimal Policies

When using the above-specified epsilon values to find a diverse set of Pareto-optimal policies within the set of over 26.000 policies, the DEFFUANT model found 24 policies, the SAMPLE model 22 policies, and the SIT model 5 policies.

#### Levers

The policies resulting from the policy discovery process are depicted in Figure 4.3 and Figure 4.4. For an intuitive comparison, a *parallel axes plot* is used. In it, each vertical axis represents one lever and each policy is depicted as one line which connects its lever values. It is to note that for each lever, only a fix set of values is explored. That is why some lines are overlapping.

To answer sub-question 4 (Section 1.3.2) and finally the main research question (Section 1.3.1), we will first look at how these policies differ in lever values. A policy's position on the first lever represents how many percent of the population is selected to benefit from the media literacy intervention (`mlit_select`). All found Pareto-optimal policies have the highest possible evaluation of this lever – 100% of agents are selected to benefit from the media literacy intervention. This makes sense if we consider the model structure and the implementation of this lever. If an agent is selected to benefit from the intervention, its media literacy skills are leveled up (from *LOW* to *HIGH*) if they are not on *HIGH* already. For agents that could level-up, this means that without the intervention they could be influenced by 100% of the misinformation that they encounter. With the intervention, they only run a 20% risk of being influenced by seen misinformation. Consequently, this should result in a substantial benefit in terms of the metric of how many agents end up being above the belief threshold (`n_agents_above_belief_threshold`). And while the intervention leads to some increased user effort (metric `avg_user_effort`), it causes absolutely no costs in terms of restricting free speech (metric




**(a)** DEFFUANT
24 Policies

**(b)** SAMPLE
22 Policies

**(c)** SIT
5 Policies

**Figure 4.3:** Levers of Pareto-Optimal Policies, per Model.

These plots provide an overview of the found Pareto-optimal policies as parallel axes plots. Each policy is depicted as one line which connects its lever values.

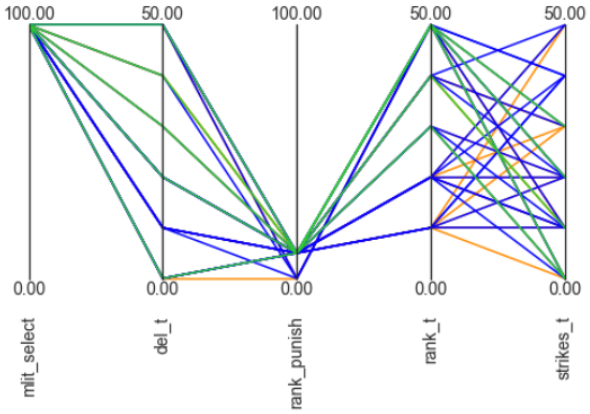



**Figure 4.4:** Pareto-Optimal Policies

The graphic provides an overview of the found Pareto-optimal policies as a parallel axes plot. Each policy is depicted as one line which connects its lever values. The policies are colored by the model that they stem from. Because only a fix set of values is explored for each lever, many lines are overlapping.

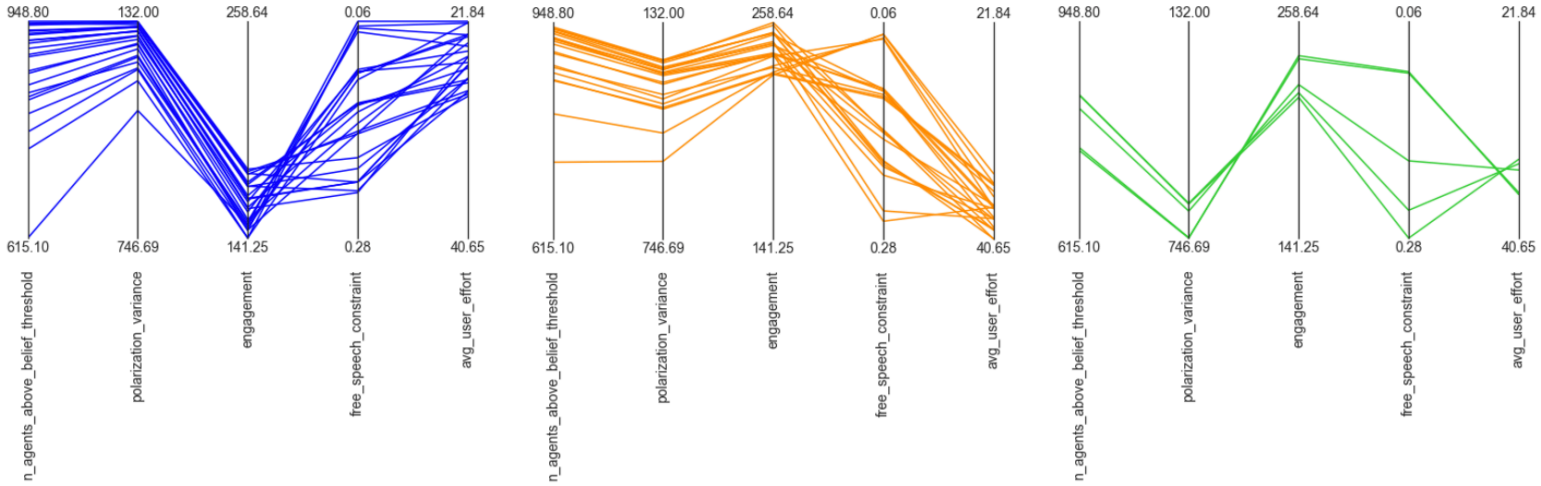
`free_speech_constraint`). As such, the strong benefits are likely to outweigh the costs. And this is exactly what we see with all the Pareto-optimal policies having the highest valuation of the media literacy intervention lever.

For the levers related to downranking posts that seem false (i.e., `rank_t` and `rank_punish`) and for the threshold levers more generally (i.e., `del_t`, `rank_t` and `strikes_t`), the story is a bit less clear. We can see that `rank_punish` is always very low (10% or 0%) for all Pareto-optimal policies of all three models. As such, even though the `rank_t` is rather distributed for all models, the overall impact of the downranking is rather weak. I am not sure what leads to this strong pattern in `rank_punish` and the distributed nature of the threshold parameters. However, the policy selection is directly determined by the outcomes, and as we will see in the following, the outcomes are well-explained by the models structures.

### Outcomes

Figure 4.5 displays the outcomes of each model's Pareto-optimal policies, and Figure 4.6 shows them all combined. As previously, each line represents one policy. But now, each vertical axis does not refer to one lever, but to one outcome. Each outcome is determined by how the combination of the model,

the policy, the scenario and the stochasticity play out. While they are averaged over 30 repetitions, the outcomes can still take a large amount of different values. The amount of exactly overlapping lines is therefore minimal. It is to note that the axes of outcomes which are supposed to be minimized (i.e., where lower values are considered better) are inverted. Like this, it becomes easier to intuitively interpret the quality of a policy because the best values for each metric are displayed on top of the figure. In the following, we will consider one outcome at a time. Each outcome part includes a description of what interesting differences can be found.




**(a)** DEFFUANT
24 Policies

**(b)** SAMPLE
22 Policies

**(c)** SIT
5 Policies

**Figure 4.5:** Outcomes of Pareto-Optimal Policies, per Model.

These plots provide an overview of the performance of the found Pareto-optimal policies as parallel axes plots. Each policy is depicted as one line which connects its outcome values in the reference scenario (averaged over 30 replications).

- `polarization_variance`

With this metric I aim to describe how polarized the population is. It is determined by the variance of the agents' belief distribution. In Figure 4.6, we can see that the DEFFUANT and SAMPLE models led to low polarization (i.e., low variance levels, which are depicted at the top). The SIT model led to policies with worse polarization values. The former two models are relatively simple in their belief update. More specifically, they are agent-independent. In the SIT model however, the size of the belief update depends on the belief confidence of the agent itself. At a belief of 50, the agent is completely uncertain and undecided about the statement in question. The further the agent's belief is from 50, the more certain it is in either agreeing (high values) or disagreeing (low values) with the statement. If an agent is close to either pole of the belief dimension (i.e., has a very high or a very low belief value), the agent is more certain. In the SIT model, such agents need more evidence, more posts to be moved far away from their extremely confident belief. This leads to a phenomenon that we will call here *sticky poles*. Once an agent is very close to a pole, it is less likely that the agent will again move further away from the pole, in other words, that pole becomes ”sticky”. Two sticky poles should lead to more polarization and that is exactly what we can observe in Figure 4.6.

- `n_agents_above_belief_threshold`

The metric of how many agents are above the belief threshold (`n_agents_above_belief_threshold`) is rather related to the previously considered one (`polarization_variance`). And also here we can see

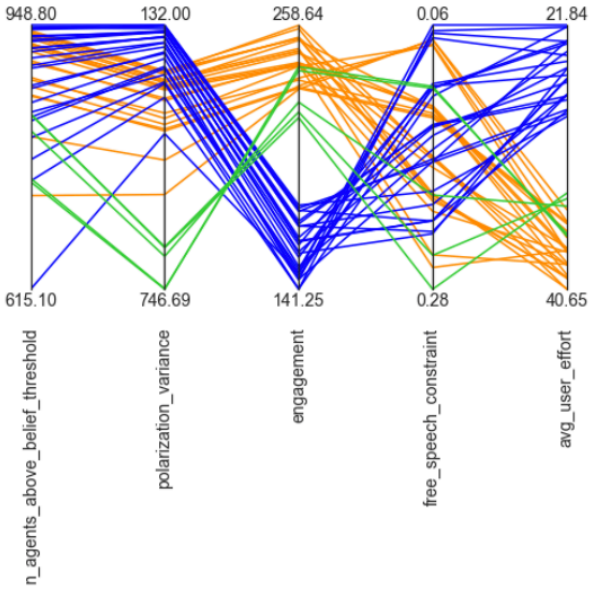



**Figure 4.6:** Outcomes of Pareto-Optimal Policies, Combined.

The graphic provides an overview of the performance of the found Pareto-optimal policies as a parallel axes plot. Each policy is depicted as one line which connects its outcome values in the reference scenario (averaged over 30 replications). The policies are colored by the model that they stem from.

that the selected policies from the SIT model are roughly lower than most policies stemming from the other two models. This pattern is linked to the initialization of the agents. The initial beliefs of normal users are placed in an uniformly random manner over the whole dimension, spanning from 0 to 100. The initial beliefs of the disinformers are placed in the same manner over a reduced span – between 0 and 10. Consequently, while both poles are sticky, there is a tendency for having more agents close to the lower pole than to the upper one. In the agent-independent belief updates of DEFFUANT and SAMPLE, the effect of this initial condition diminishes over the steps of the model run. In the SIT model, however, the poles being sticky means that initial imbalance has a stronger effect on how many agents are still below the belief threshold by the end of the run.

- `engagement`

In Figure 4.6, we can see a quite clear distinction between DEFFUANT on the one hand, and SAMPLE and SIT on the other hand. This is currently a bit puzzling. The engagement represents how many posts are seen by the agents. This is influenced by a few factors. First of all, as previously described in the model section, more extreme beliefs increase the amount of posts an agent aims to post. And all else being equal, more posts posted mean that more posts are seen. The other factors that influence the engagement metric are those lever values that can decrease how many posts are seen – `rank_punish`, `rank_t`, `del_t`, `strikes_t`. The lower the values for all of these levers, the higher the amounts of post that are seen. The lever of the media literacy does not affect this metric because the media literacy level of an agent does not influence how many posts they see. The main thing that SAMPLE and SIT have in common is that in both models, the agents always update whereas in the DEFFUANT model, agents update only very rarely. However, whether or not the agent actually updates on a seen post, does not influence how many posts are seen by the agent. Consequently, it might be the case that this stark difference could be caused by a difference in policies. However, judged by Figure 4.3, the relevant lever values seem quite similarly distributed in the different models. It may be the case that the previously mentioned overlap of the policies means that if for instance DEFFUANT's policies would

have vastly more high lever values than the other two, that would not be visible in this simple plot. To investigate this possibility, a different kind of plot could be added. A plot that gives an indication of the density of policy levers and/or boxplots could be a good choice.

- `free_speech_constraint`

In Figure 4.5, we can see that the outcomes for the free speech constraint are dispersed. There is no very clear distinction between the models possible. While there is no clear distinction between the models possible, if we look closely, we can spot that the policies resulting from the SIT model are overall rather a bit lower than those from the other two models.

The value of the free speech constraint depends strongly on the threshold lever values. The higher numbers for these levers, the more posts fall into the actionable category. Thus, there is a clear relation between the lever values and the expected outcome values. In Figure 4.3, we can observe that the threshold levers are themselves dispersed over their dimensions. Therefore, the observed dispersion is in line with the policies that lead to these outcome values.

To make sense of the somewhat worse performance of SIT-stemming policies, let us keep in mind that in this project, free speech can only be constrained in case of posts that are (likely) false. With the simple statement that is at the heart of the considered belief, posts that represent beliefs of lower values are more likely to be false. The more agents have low beliefs, the more posts with low belief values there are. And in turn, the more the free speech might be constrained by downranking or deleting posts and by blocking agents from posts (as a consequence of having received too many strikes for posting misinformation). As we have seen in the metric of `n_agents_above_belief_threshold`, SIT's policies have fewer agents above the belief threshold. Consequently, they have more agents with lower belief values, which increases the amount of posts that are actionable and thus also how strongly free speech is constrained. This could explain the somewhat worse performance of SIT-stemming policies with respect to the free speech constraint metric.

- `avg_user_effort`

In Figure 4.6, we can observe a clear split between the DEFFUANT model on the one side, and the SAMPLE and the SIT model on the other. The DEFFUANT model seems to lead to more optimistic valuations of average user effort, while the other two models indicate worse values. Furthermore, we can see that this dynamic is the same split between the models as we observed in the `engagement` metric. Low engagement levels are correlated with a desirable low user effort, while high engagement levels are correlated with an undesirable high user effort.

The final metric of how much effort the users have on average is mainly influenced by two factors – 1) the amount of agents with high media literacy, and 2) the engagement. High media literacy means that an agents takes more time to check the veracity of a post. All models led to policies that are at the highest level of the media literacy lever. This leads to rather high values in average user effort for the policies of all three models, but cannot explain the difference between the better performing policies (DEFFUANT model), and the worse performing policies (SAMPLE and SIT). To explain this difference, let us turn to the second factor (engagement). Higher engagement means higher number of seen posts. With each seen post, the overall user effort for judging the veracity increases. Thus, the model structure should lead to a correlation between engagement and user effort. High engagement implying high user effort implying, and low engagement implying low user effort. Because the engagement is visibly lower for the DEFFUANT model, we would predict lower average user effort. This is exactly what we observe.

## 4.3. Re-evaluation Under Deep Uncertainty

For the re-evaluation under deep uncertainty, the above-described policies are used. Consequently, we will refrain here from looking at them again and focus instead on their performance in terms of resulting metric values in the 50 scenarios. Figure 4.7 depicts these performances on the right. For easier comparison with shared axes, the original performances with the reference scenario is shown on the left. To compare between the models the plots are combined and annotated in Figure 4.8.
Before going into the differences between the models, let us first have a look at how similar the performance in the re-evaluation is to the original evaluation.

For metric (1) `n_agents_above_belief_threshold` and (2) `polarization_variance`, both evaluation phases indicate a rather clear split between the models. Though for the first metric, the DEFFUANT and the SAMPLE model indicate largely overlapping values. For both of these metrics, the SIT model anticipates somewhat lower values.

For metric (3) `engagement`, the original evaluation leads to unrepresentatively low values of engagement. We saw a clear split between the models with the DEFFUANT model indicating lowest engagement levels. This is still also the case in the re-evaluation under deep uncertainty. In all three models, but especially in the SAMPLE and the SIT model, the re-evaluation resulted in a much larger spread of resulting values.

The performances with respect to metric (4) `free_speech_constraint` have already been hardly distinguishable between the models. This is also the case for the re-evaluation under deep uncertainty.

Regarding metric (5) `avg_user_effort`, the performances after re-evaluation are a lot more dispersed than in the original evaluation. We can still make out that there is a higher density for the DEFFUANT (blue) and the SAMPLE (orange) model than for the SIT model, but there is a large overlap between the models' indicated performances on this metric. For the SIT model (green), there is also a higher density in the upper half of the resulting values, but the results are vastly more distributed after re-evaluation.

When comparing the scales of the various metrics, we can observe that the value ranges for the metrics that are more directly related to the beliefs (i.e., metrics (1) and (2)), the value range between the original evaluation with the single reference scenario is not much smaller than the value range after re-evaluation with 50 scenarios. For the remaining three metrics, this difference is more pronounced.

## 4.4. Selecting Policies Based on a Single Objective

### 4.4.1. Levers

For each of the models, the same number of policies are selected (110 policies). In Figure 4.9 we can spot a few interesting patterns within these policies.

Firstly, the lever that determines how many agents are benefitting from the media literacy intervention (`mlit_select`) and the lever which controls how strongly detected misinformation is down-ranked (`rank_punish`) are for all three models distributed over their whole possible ranges. The remaining levers are all threshold-levers. For each model, the selected policies are concentrated on a single value within the range of these threshold-levers. The threshold for deleting is either at 30% (for DEFFUANT and SAMPLE), or at 40% (for SIT). The thresholds for downranking and giving out strikes are for the DEFFUANT model both at 40%. For the SAMPLE and the SIT model, the downranking threshold is at the highest possible value of 50%, while the threshold for handing out strikes is at 30%. We therefore can observe that the policies stemming from the SAMPLE model combine the lever choices of the other two models. As a consequence, its policies are completely concealed in the figure that combines the models' policies (Figure 4.10).

When comparing the policies selected by the two methods of selection, we realize that there are some stark differences. In the multi-objective selection, the threshold-levers featured very distributed

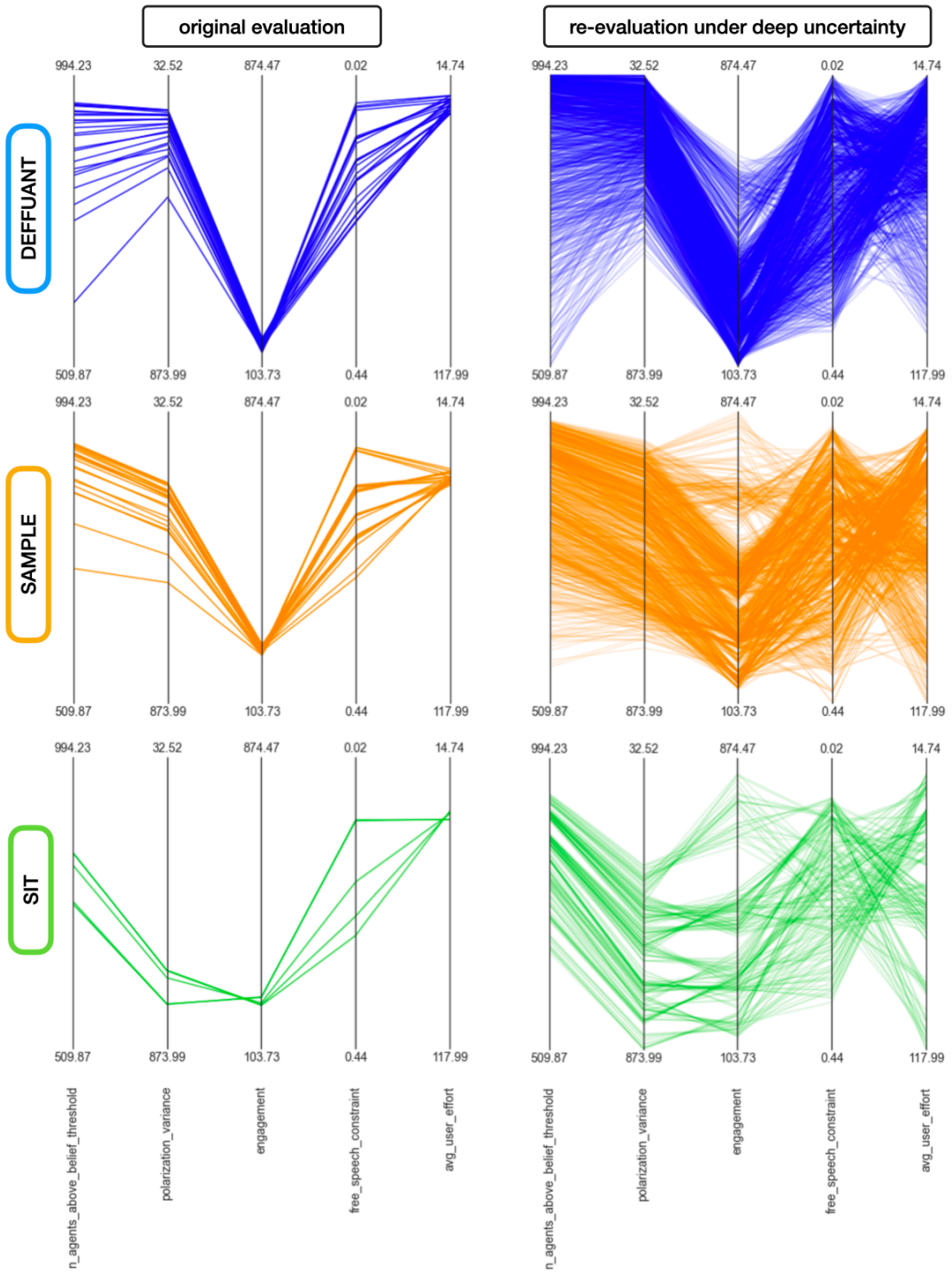



**Figure 4.7:** Outcomes of Pareto-Optimal Policies (per Model)

Original evaluation refers to the evaluation under the reference scenario. Re-evaluation under deep uncertainty refers to the evaluation in 50 sampled scenarios. As previously, each policy is depicted as one line which connects its outcome values in the reference scenario (averaged over 30 replications).

values. For the other two levers, the policies are concentrated on only a few values. The set of policies stemming from the single-objective selection exhibit the opposite trend.

For the policies stemming from the multi-objective selection, we specifically noticed that all selected

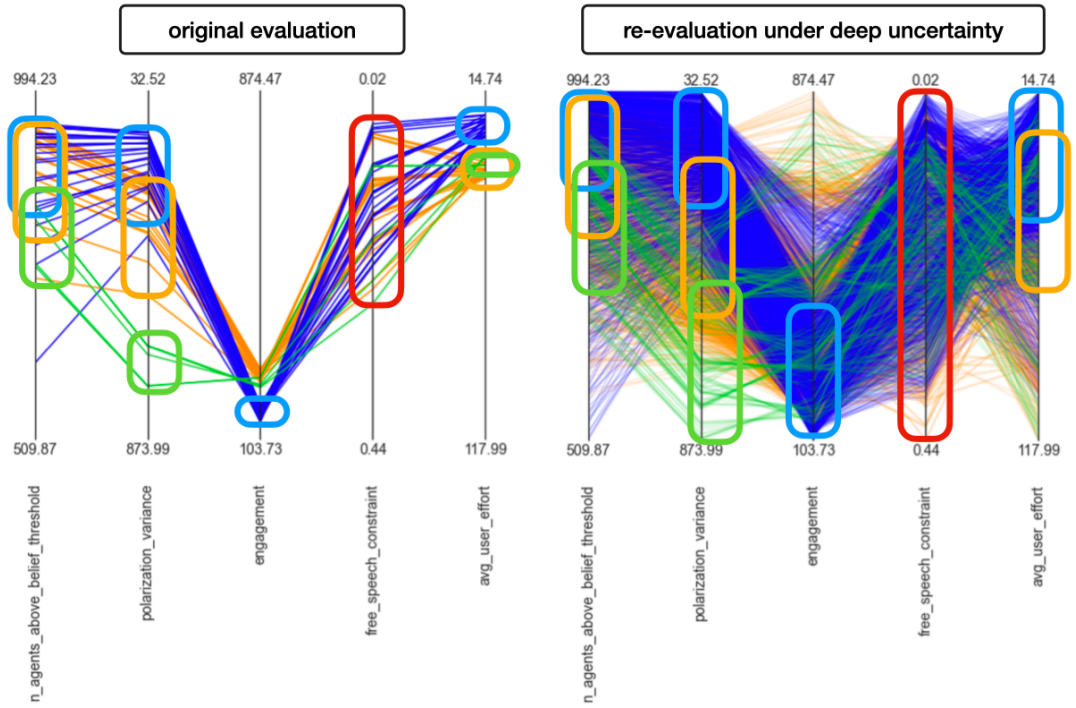
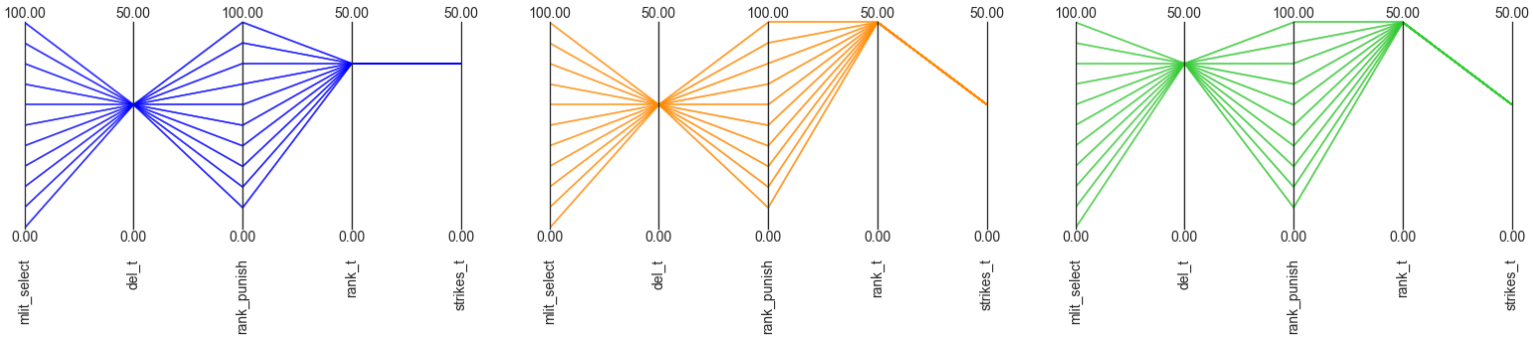



**Figure 4.8:** Outcomes of Pareto-Optimal Policies (Combined Models)

Analogous to Figure 4.7. With additional boxes to highlight the overall differences between the models. Blue maps to the DEFFUANT, orange to the SAMPLE, and green to the SIT model. The red box indicates that a lack of significant differences between the performance of the models' policies.




**(a)** DEFFUANT 110 Policies

**(b)** SAMPLE 110 Policies

**(c)** SIT 110 Policies

**Figure 4.9:** Levers of the Selected Policies, per Model.

These plots provide an overview of the found Pareto-optimal policies as parallel axes plots. Each policy is depicted as one line which connects its lever values.

policies empowered the entire agent population with the media literacy intervention. This is reasoned to be explained by the high efficacy of this intervention. The main disadvantage of the media literacy lever is that higher media literacy levels lead to more user effort. By applying the single-objective selection, we ignore this trade-off between high media literacy levels and user effort. Therefore, it is puzzling the values here are so dispersed when we are not considering the the metric of user effort for the policy selection.

What I would have expected to see with respect to the selected policies is that they are mainly

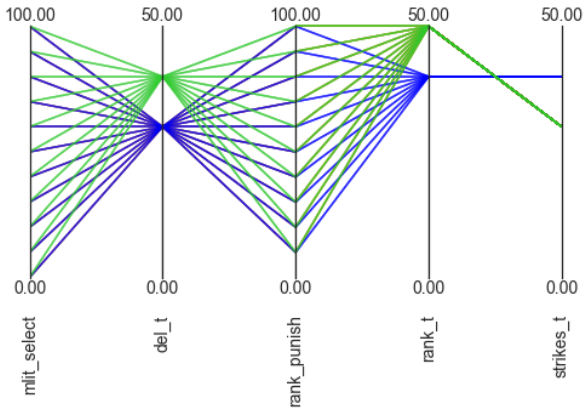


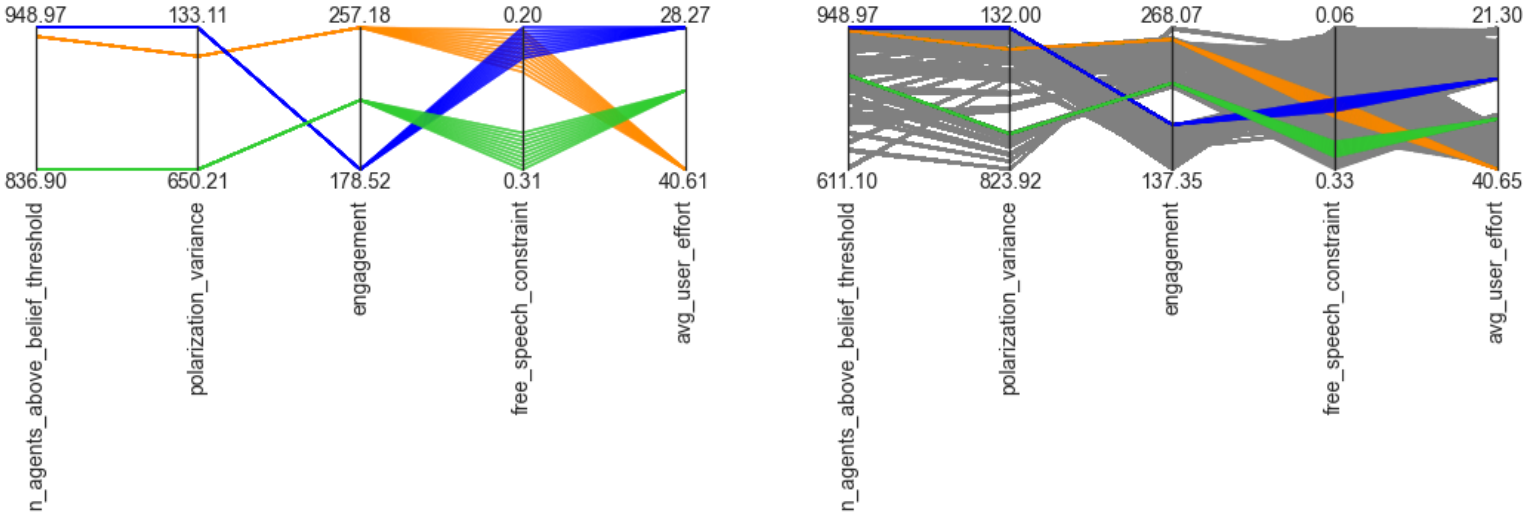

**Figure 4.10:** Pareto-Optimal Policies

The graphic provides an overview of the selected policies as a parallel axes plot. Each policy is depicted as one line which connects its lever values. The policies are colored by the model that they stem from. It is to note that the policies from the SAMPLE model are completely covered by the policies of the other to models. The SAMPLE model's policies use the same valuation of the `del_t` lever as the DEFFUANT model's policies, and the same values on the levers of `rank_t` and `strikes_t` as the SIT model.

focusing on the highest values of all levers. This is because the downsides which these levers cause for other metrics are irrelevant when only focusing metric (1) `n_agents_above_belief_threshold`. One reason why we see a different pattern could be that the outcome of this first metric is so easily brought to the absolute maximum value, that it does not matter so much how strongly the levers are activated. However, as we will see in the outcomes of the policies from the single-objective selection, this does not seem to be the case (see Figure 4.11). The number of agents above the threshold is for the SIT model substantially lower than the absolute maximum. The values for the policies of the DEFFUANT and the SAMPLE model are both quite high. However, they are not at the maximum value. In the reference scenario, the ratio of normal users to disinformers is around 0.993. With 1000 agents in total, the absolute number of normal users should therefore be around 993. If the policies would all already lead to the maximum number of agents above the threshold, the first metric should be around 993. However, even the model with the highest performing policies (DEFFUANT) "only" reaches around 950 agents above the belief threshold. In each model it should therefore be the case that stronger lever activation could achieve higher values on the first metric. That being the case, we are can rule out this hypothesis and therefore have another avenue for interesting future work.

### 4.4.2. Outcomes

Figure 4.11a shows how the models' selected policies performed in the reference scenario.




**(a)** Outcomes of Selected Policies

**(b)** Outcomes of Selected Policies (in color), and all other Policies (in grey).

**Figure 4.11:** Single-Objective Selection Outcomes, per Model.

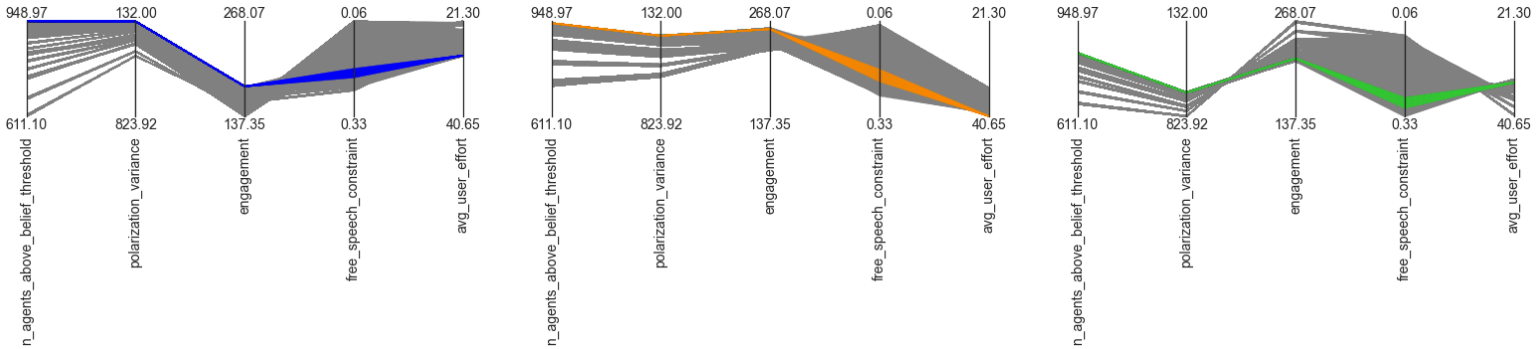
The policies which are selected when only considering the fist metric are drawn in color. Blue maps to DEFFUANT, orange to SAMPLE, green to SIT.

For each model, the outcomes of its policies are astonishingly consistent. The only metric in which there is some more pronounced spread is the amount of free speech constraint. For the first metric itself, its values are necessarily maximally concentrated because we all policies that we selected have the same value (i.e., the maximum of the achieved values). Regarding the concentration of values on metric (2), we can reason the following. With the extremely high number of agents above the belief threshold, there are not many options that would nevertheless lead to high variance. Especially for the DEFFUANT model with a very high number of agents above the threshold, it makes sense that the selected policies also lead to very concentrated values in the second metric. Also for the SAMPLE and the SIT model, there is hardly any spread on the polarization metric of the selected policies. This might be the case because the population polarized to such a degree, that the exact value of the variance mainly depends upon how many agents are in either of the two camps. The spread that we have seen with respect to the free speech constraint, is likely caused by the `rank_punish` lever. Of those levers that influence the free speech constraint, it is the only one with variation.

Furthermore, when focusing on the performance in the reference scenario, we can observe that the relative performance of the models selected policies lets a very similar ranking emerge as in the multi-objective selection. The policies from the single-objective selection almost seem to be the result of an attempt to distill the quint-essence of the relative performances from policies of the multi-objective selection. This is not very surprising, but is some evidence that the policies selected when only using the single objective are maybe also part of the sets of Pareto-optimal policies from the multi-objective selection.

When selecting policies by only considering the first metric, we still get high performance in terms of the second metric, but mostly worse performance in the remaining metrics (see Figure 4.12. For the DEFFUANT and the SAMPLE model, this trade-off hits especially metrics (4) and (5). For the SIT model, it is more pronounced for metrics (3) and (4). Even though for the DEFFUANT model, there seems to be no costs of the single-objective selection in terms of engagement, it is to note that the policies of the DEFFUANT model lead to lower engagement than those of the other two models (see

Figure 4.11b. In all models, metric (4) `free_speech_constraint` suffers when only considering the first metric to select policies. Given our above reasoning regarding the concentration of values for metrics (1) and (2), as well as the relation between these values, it makes sense that we do not observe a trade-off regarding the polarization. For the remaining metrics, we see some trade-off, which is what we would expect.




**(a)** DEFFUANT
110 Policies

**(b)** SAMPLE
110 Policies

**(c)** SIT
110 Policies

**Figure 4.12:** Outcomes of Pareto-Optimal Policies, per Model.

In grey, the outcomes of all considered policies are depicted. The policies which are selected when only considering the fist metric are drawn in color.

# 5

# Discussion

## 5.1. Main Findings

### 5.1.1. There is a clear distinction between the models' optimal policies as well as their outcomes.

In the Results section, we have observed differences between the policies resulting from the three alternative models. The same holds for the performances of the models' policies. On a higher level, we can therefore conclude that the structural uncertainty of the belief update function is a difference that makes a difference. This is especially the case for the two metrics that are more directly related to the agents' belief ((1) `n_agents_above_belief_threshold` and (2) `polarization_variance`). While after re-evaluation, the distinction is less clear for the remaining three metrics, we can still observe a difference (for instance in the engagement metric). Also the results from selecting policies based on the single objective provided additional evidence for this finding. Indeed, for these policies and their outcomes the differences are exceptionally pronounced, driving home the point of this distinction.

### 5.1.2. Differences in parameters do make a difference

We have seen that the outcomes of the original evaluation phase look a lot more clear-cut than those from the re-evaluation under deep uncertainty. The re-evaluation phase has shown that, dependent on the particular instantiation of uncertain parameters, the same policies can lead to rather different outcomes. This is the case despite we have been running each of the computational experiments 30 times before aggregating the resulting data and only depicting that. The results from the additional policy selection method of only using one metric as a criterion might seem at first like rather strong counter-evidence to this finding. However, we want to keep in mind that these results are all from evaluating the policies under the reference scenario. Therefore, the results from this selection process are only mild counter-evidence to this finding. Performing a re-evaluation of these policies under deep uncertainty could be a useful way forward to be more certain about the correctness of this finding.

### 5.1.3. Optimism of the models' policies in descending order: DEFFUANT, SAMPLE, SIT

With the "optimism of a model's policies", I mean how high the policies scored within the computational experiments (with respect to the included metrics). For metrics (1) `n_agents_above_belief_threshold` and (2) `polarization_variance`, and to a limited extent in metric (5) `avg_user_engagement`, we observe that as the 'optimism' of the model's policies follows the same order. The DEFFUANT model's

policies are the most optimistic when considering these two to three metrics. The SAMPLE model's policies are on the second position and the SIT models' policies are the least optimistic. Especially because metrics (1) and (2) are the most considered in related literature, this is a relevant finding. This finding is supported by the results of all three performed sets of (re-)evaluation, but is especially clear in the results from evaluating the policies from the single-objective selection.

### 5.1.4. The outcomes of the DEFFUANT and the SAMPLE model are are more similar to each other than to the SIT model

In at least two out of the five considered metrics ((1) `n_agents_above_belief_threshold` and (2) `polarization_variance`), the policies from the DEFFUANT model and from the SAMPLE model are substantially closer to each other than to those of the SIT model. This is the case when selecting Pareto-optimal policies by considering multiple metrics as is common within the field of DMDU. And this clustering is even more pronounced when selecting promising policies as is common within the field of studying misinformation with ABMs (i.e., by only considering a single objective belief-based metric). Again, this finding is particularly pronounced in the two most commonly used metrics (metrics (1) and (2)). This is a particularly interesting finding as I started out expecting that the structural uncertainty would make it difficult for policy-makers to use the resulting contradicting evidence. However, it turned out that the most commonly used belief update functions result in similar outcomes, but have the same blind-spots and shortcomings and therefore run the risk of being misleading. Thus, the academic field aiming to tackle misleading information online is currently at an unnecessarily high risk of producing misleading research.

## 5.2. Threats to Validity

In this section, we will consider some ways in which the validity of the results and the findings could be threatened. As such, we focus solely on aspects that could neutralize or negate the findings.

**Method Uncertainty**

In the beginning of the project, I decided to not explore method uncertainty within this project. This decision was made with good reasons and without regrets. However, one thing that plays nevertheless into method uncertainty and that seems good to mention is the following. In this project, it was assumed that the performance of policies would scale rather smoothly between the six or eleven considered values which each lever was allowed to have. While this seems reasonable, it is not a certainty. A rough estimation how valid this assumption really is could be done by investigating the gathered results. For instance, all else being equal, one lever value could be changed and the results of such changes could be compared. The results of having the lever set to 0% should be more similar to the results when the lever is set to 10% than when the lever is set to 90% (or any other value that is further from 0% than 100%). This could be done for all levers. In such a procedure, it is crucial to make sure to avoid situations where the remaining levers are set to their absolute minimal values. This is because that would cause the analysis to be blind to interaction effects between the levers. If one is willing to invest more computation, the assumption of smoothly scaling levers could be avoided by applying a more traditional MORDM approach. Thus, rather than evaluating the full-factorial combination of all lever values, one could perform an optimization with an evolutionary algorithm. Within this project, the risk of this assumption was somewhat mitigated by ensuring that the exact same policies were evaluated for each of the models. In case the assumption of smooth linear scaling would not hold, evaluating the same policies guarantees that we can still compare the performances of policies from different models.

**Evidence for Selection of Pareto-optimal Policies (Multiple Scenarios & Vulnerability Analysis)**

As typical for the MORDM approach, all potential policies are evaluated only on the chosen reference scenario. However, with which reference scenario policies are evaluated on can influence the results of

the evaluations. To understand why that is the case, let us consider the following toy example. Policy 1 performs well in scenario A, and badly in scenario B. The performance quality of policy 2 is the other way around. In this case, which of the two policies seems more promising mainly depends on the choice of the reference scenario. While there likely are scenarios that are just ”harder” to master for any policies, we cannot assume that this is always the case. We cannot assume a consistent ranking of policies across scenarios. Consequently, we can only have a limited amount of confidence in the overall quality of the evaluated policies. In order to improve this situation, we can evaluate the policies directly in multiple scenarios (similarly to Multi-Scenario Multi-Objective Robust Decision-Making (MS-MORDM)). We can then have a better estimate of the overall performance of a policy. Due to the full-factorial design, the identical policies would be evaluated on each of the reference scenarios. This makes it possible to combine the gathered information from the different scenarios easily. Then, robustness metrics could be used to select the Pareto-optimal policies. Subsequently, these policies could be reevaluated under deep uncertainty. With this approach, robustness could be prioritized more. To go even further, this approach could be combined with vulnerability analysis to do multiple rounds of (MS-)MORDM. This approach could change which policies are selected as Pareto-optimal policies and therefore is a threat to the validity of the results.

**Pattern-Oriented Modeling**
During the phase of the project in which I was focusing on grounding the model and model validation, I had not yet been aware of *Pattern-Oriented Modeling (POM)*. This seems to be a very useful way to figure out what to focus on in the pursuit of making sure that the model is fit for purpose (Grimm & Railsback, 2012; Grimm et al., 2005; Railsback & Grimm, 2019). Using this framework could have improved that project-phase and increased the odds that all crucial aspects are fit for purpose.

**Media Literacy Intervention**
The media literacy intervention of MisinfoPy is modeled in a rather simple way. This was done for a lack of scientific data to base more detail on. In the current version, the agents only can have two different levels of media literacy – high and low. A more detailed representation might be useful to capture the real-world situation better. Furthermore, the current implementation assumes that people with a high level of media literacy will really spend as much more time on evaluating the veracity of posts (averaging 30 seconds per post, while those with low media literacy only spend an average of 3 seconds). It is good that this does allow for the behavior in which people look at most posts only briefly to judge their truthfulness, and just rarely spend a bigger amount of time on posts that they really care about. However, extending the validation seems useful here. It might be the case that even though people still have the knowledge of the high media literacy level, they do not apply this knowledge as often as presumed by the current model. Similarly, the success rates of judging the truthfulness of posts could benefit from being grounded in real-world data. Additionally, these behavioral patterns might change over time. This is especially relevant because this intervention is very ”prominent” in the Pareto-optimal policies (They all have the highest possible value for this lever.). Fortunately, it is the case that the exact values do not matter so much because only the relative performance between policies is decisive which policies are selected. However, if a more realistic version of this intervention would be less potent and therefore less decisive in selecting Pareto-optimal policies, this could be a threat to the validity of the results.

**More Dynamic Model**
In the model, the connection between agents is kept static. In the real world, however, connections between users are dynamic. On social media platforms, users can follow and unfollow others on an ongoing basis. This is definitely something that happens in the real world over the course of the represented time frame of two months. And especially because agents only get to see the posts of agents that they follow, this could be a beneficial characteristic to include. For instance, let us assume that

some agent A receives every step a lot of posts from agent B (for instance because agent B is a disinformer). Furthermore, agent A's belief is very different from agent B. Then, agent A might perceive these posts as spam and unfollow agent B. Naturally, this would include a whole other set of uncertainties. And due to the limitations of real-world data and project scope, this endeavor was intentionally left unexplored. However, such dynamic connections should change how the overall belief distribution develops. Because multiple metrics are directly related to the belief distribution, the choice between static and dynamic agent connections could be decisive for the selection of Pareto-optimal policies. Along the analogous argument from previous subsections, this could indicate a threat to the results' validity. Furthermore, this could be an interesting aspect to explore because the two belief-distribution-related metrics have turned out to differ between the models' outcomes.

# 6

# Conclusions

In this chapter, we will tie everything together by concluding the research questions, considering the societal and academic implications of the answer to the research question (Section 6.2). We will also look ahead by considering promising avenues for future research (Section 6.3).

## 6.1. Concluding the Main Research Question

For easy reference, here is the main research question again:

**Main Research Question**

How does the consideration of **structural uncertainty** with respect to the choice between different **belief update functions** influence the resulting **Pareto-optimal policies** and their performance?

We are interested in this question because we want to find out whether it would be relevant to consider this structural uncertainty and how much DMDU methods could benefit the field of agent-based misinformation modeling. First of all, the findings of this research project indicate that the consideration of structural uncertainty with respect to the choice of belief update function does indeed influence which Pareto-optimal policies are resulting, as well as what the performance is caused by this set of Pareto-optimal policies. With respect to how the consideration of this structural uncertainty influences the resulting Pareto-optimal policies, we have a few main insights. We have seen that the structural uncertainty of the choice of the belief update function influences how desirable the outcomes of the Pareto-optimal policies are. The different belief update functions lead to consistently better (or worse) performance across the commonly considered metrics. This ranking has been especially pronounced when selecting policies by only considering a single metric, as is currently common in the field of modeling misinformation on social media. Interestingly, the belief update functions whose selected policies imply better performance turned out to be those that are the most commonly used in literature. This exacerbates the situation and stresses the relevance of discussing this structural uncertainty in order to avoid unrealistic expectations and a false sense of certainty. Furthermore, we have found that how simplistic (versus how well-embedded in social psychology research) the used belief update function is, influences the similarity between outcomes. Models with simpler belief update functions (DEFFUANT and SAMPLE) have resulted in more similar outcomes. This is the case despite the fact that the Pareto-optimal policies of all three models are fairly similar. This has additionally been supported by the evidence from the extra step of selecting policies by considering only one objective, as

is common on the field of modeling misinformation on social media. Overall, this project has provided that utilizing DMDU methods has plenty of value for the field of agent-based misinformation modeling. With the example of exploring the structural uncertainty about the belief update function, we have seen that if we would not consider this structural uncertainty and instead look only at one model individually, only a lower degree of certainty in the emerging dynamics would be warranted. If, however, we look at an ensemble of models and the observed dynamics overlap, we can gain more certainty on these dynamics. Furthermore, we can attain additional insights which would otherwise remain hidden.

## 6.2. Implications

The main motivation behind this project is to showcase how beneficial it could be to utilize DMDU methods within the field of agent-based misinformation modeling. In this section, we will aim to zoom out to see exactly these kinds of scientific implications. Furthermore, because science exists not in a closed bubble but because it influences society as a whole, we will also consider the societal implications our findings.

### 6.2.1. Scientific Implications

**Low-hanging Fruits for Researchers**

In this project, we have seen that the particular structural uncertainty which we explored, is a difference that makes a difference. This means that it could be promising to even just have a scientific discussion about for instance which belief update function is most fit-for-purpose under which conditions. However, the value that DMDU methods could bring to the field is not constrained to this structural uncertainty. There are other structural uncertainties that could be worth exploring (for instance the posting behavior). Furthermore, also parametric uncertainties or different sets of objectives for problem formulations are usually not discussed or explored within the field. We have also witnessed that differences with respect to chosen values for uncertain parameters can lead to relevant differences in the performance of a given policy. At the same time, we have observed that, even though the scenarios for the re-evaluation are so different from one another, the outcomes often still lead to values that are mainly focusing on a sub-range of the metrics ranges. This can lead to valuable insights and it makes a strong point for the importance of exploring uncertain model parameters. Additionally, the DMDU approach offers state-of-the-art methods to identify vulnerable scenarios, i.e., scenarios which would be particularly bleak. Overall, this means that we have identified a research niche that is not only relevant, but also neglected and tractable. This offers great career opportunities for researchers that are interested in the field. Conducting such research could make valuable contributions to the field of studying misinformation.

**Improving the Future of the Field**

For the structural uncertainty of the belief update function, it became clear that the most commonly used belief update functions lead to the most desirable outcomes in the commonly considered metrics. At the same time, the field of modeling misinformation on social media with ABMs is still turning a blind eye to relevant uncertainties. This implies a risk that the results from literature of the field causes unjustified optimism. This would likely lead to frustrations down the road when policies that were based on the modeling work do not achieve the hoped-for benefits. Next to frustration, this could also lead to modeling being less trusted and thus to decreased research funding. However, there is also the positive flip-side. By making use of DMDU methods, the field of modeling misinformation on social media could lead to fewer spurious take-aways and more robust insights. Consequently, the field would be more valued and thrive.

### 6.2.2. Implications at the Intersection of Science and Society

**Increasing the Research's Positive Real-World Impact**

Many researchers are likely not only in the field to advance the frontier. Many researchers likely also care about the real-world consequences of their work. We want to positively contribute to the world's pressing challenges. We want that new insights can benefit the world and that we do not exacerbate problems by accidentally producing research that has a misleading effect. For this reason, it is in the interest of many researchers to consider multiple objectives and thus reduce the risk that gridlock will stop their insights from benefiting the world. Along an analogous line of reasoning, it is of interest to researchers to have tools available that can help to detect current blind-spots or a false sense of certainty in the field. For a concrete example of how a false sense of certainty could arise, let us draw from the results of this project. The majority of the papers found in the literature review use either the DEFFUANT or the SAMPLE belief update function. In this project, the Pareto-optimal policies that resulted from those two models both indicated good performance with respect to metric (1) `n_agents_above_belief_threshold`. If the majority of the papers in the field is indicating that we have policies that perform with respect to this metric. Thus, the modelers and the policy-makers who rely on these papers, would be rather confident that with such a policy, we can expect good performance with respect to that metric. In other words, there would be the commonly drawn conclusion that if we implement such a policy, we can be relatively certain that after roughly two months, a large portion of the population will fall above the belief threshold. It goes without saying that even if this threshold is not the most accurate representation of the required confidence that people need to have in the safety of COVID-19 vaccines to decide to get vaccinated, overestimating how many agents will be willing to vaccinate can have devastating consequences. Making use of DMDU methods promises to be helpful for increasing the positive real-world impact of conducted by mitigating issues such as political gridlock, undetected harmful blind-spots or a false sense of certainty.

**DMDU, Validation, Funding**

In terms of academic implications, this project is a proof of concept for the value that DMDU methods can have for this grand challenge. As such, this project has potential to inspire other researchers in the field to make use of these potent methods used by other fields. Furthermore, while there is definitively room for even more validation, the project shows that more extensive validation is possible for these kinds of models. As such, the project could inspire the field to do more extensive validation. For instance, embedding especially central components (such as the way agents change their opinions) in literature while not shying away from benefiting from relevant insights gained in other disciplines. In summary, the academic implications of this study could be interdisciplinary cross-pollination. Such a shift towards acknowledging uncertainties and towards using methods that lead to more reliable, robust results could be a step towards additional blossoming of the field. If this would be joined by a development towards using models that have a better fit for purpose, it could not only speed up the progress of the field, but also have the societal benefit of increasing the value society can get out of the allocated funding.

### 6.2.3. Societal Implications

**Blanket Statement Against Modeling**

This brings us to the societal implications. Firstly, I would like to mention a risk which becomes clear when considering the political context of the project. Honest misinterpretation or strategic argumentation could lead to conclusions such as "All those models give contradictory results. We cannot trust any of them.". This overgeneralizes and thus misses the point. It makes sense that different choices behind models can lead to different outcomes. Also in non-modeling based reasoning different world-views and different assumptions lead to different conclusions. But 1) making the world-view and the assumptions explicit and by 2) utilizing the computational skills of computers to help us "think through"

complex systems, enables better understanding of the system and the potential for finding policies that are more robust and beneficial for many stakeholders. However, via statements such as the previously mentioned one, the results of this study risk being used as a blanket statement against taking modeling results into consideration during policy-making. While also simply gathering multiple previously existing studies could lead to such statements, this risk seems to be increased for this project because it highlights the differences in results of models which differ only at one single location.

**Avoiding Gridlock**
The second societal implication is more positive. An issue for bringing about policies that aim for societal benefit is that there are often other powerful stakeholders who feel that these policies would overly curtail their objectives. This can lead to gridlock situations, or even situations where these powerful actors can influence policies stronger than most would perceive as fitting (see for instance with the recent new insights about Uber's influence, or the extreme case of Boeing's influence on aviation policies in the US.). Also within the handling of the grand challenge of misinformation on social media itself, we can imagine a gridlock for policies. Policies might be seen as overshooting and ignoring the objectives of social media platforms. Consequently, their powerful lobbies might aim to avoid those potential policies. By explicitly including interests of a diverse set of stakeholders, and by aiming to find robust solutions that consider all these interests, such gridlocks could be mitigated.

**Improved Information Environment**
Last but not least, let us consider a broad and impactful portion of potential benefits that could be accrued over time. As mentioned above, I believe that this kind of project has the potential to spark a revolution in the academic field by infusing awareness of DMDU methods. The use of these methods could lead to better information support for decision-makers. This in turn would enable tackling the challenge of misinformation more successfully, leading to a better information environment. As mentioned in the introduction, better handling of misinformation on social media would strengthen democracy. It could enable us to successfully tackle many other challenges that require large numbers of people to support policies around challenges that we are currently facing – from climate change and pandemics, up to the policies around break-through technologies. It is hard to estimate the potential societal impact that such a revolution could have, but during the last pandemic we have seen how having a more successful handling of such challenges can save many lives, livelihoods and unnecessary frustrations. And this pandemic was just one rather short-term challenge. We can only try to imagine how much value could be generated across all grand challenges, just by enabling a better information sphere and joined sense-making.

I would like to end this section on the note that I hope my careful phrasing has made it clear that I do not expect that such major benefits are caused by this project. At the same time, I do believe that this project could be a stepping stone towards manifesting the described benefits.

## 6.3. Future Research

### 6.3.1. Directly Related Research

**Further Validation**
Naturally, future research could aim getting to grips with the outlined threats to the validity of the findings (Section 5.2). How this might be done follows rather straight-forwardly from that section. One promising direction for future research, which is very close to the research question of this project but goes a few steps further is to increase validation of the model. The clearer the model's fit for purpose is, the clearer it is how the findings of the computational experiments may be translated to the real world. Such translation and generalization is also referred to as 'escaping from model-land' (Thompson &

Smith, 2019). It is a non-trivial process and can easily lead to misunderstandings. As common in science and life, nothing (not even validation) can bring us to complete certainty. It is the case that even with a incredibly thoroughly tested and validated model, we could still encounter surprises if the model turns out to lack some necessary components or aspects. Striving for more certainty is nevertheless worthwhile. Validation is the foundation of successfully escaping model-land and therefore definitely important for future research.

**Evaluation Across Models**

If we could find policies that perform well independently of the belief update function, it provides hope that these policies (or the patterns that we see in them) could be worth-while to test in small-scale real-world experiments. If we would find such strategies, we would have policies that – at least in model-land – perform well, no matter which world-view we have in terms of how humans update their beliefs. To search for such policies, we could test the found Pareto-optimal policies of each model on the two other models. With that data, we can see whether there are policies that perform sufficiently well across the models. Furthermore, we might be able to find patterns such as that mainly the policies from one specific model perform well on other models as well. If that would be the case, it could be a good call to default to the belief update function of that model in case modelers still want to go ahead and only use one such function in their modeling efforts. Naturally, the results of this approach would be more reliable if the Pareto-optimal policies would stem from an evaluation on multiple scenarios, with a potential robustness filtering.

**More Restricted Range of Lever Values**

The whole set of possible lever values is explored and the Pareto-optimal policies are selected from all the considered policies. However, a more restricted range of lever values could have yielded results that could be more applicable to the real world. In order to make the results more applicable to the real world, it could be a valuable approach to not consider the whole lever ranges for choosing the Pareto-optimal policies, but mainly those ranges that seem feasible given the political and real-world context. For instance, depending on the exact real-world implementation of the media literacy intervention, it would not be feasible to offer the intervention to 100% of users. Reasons for such restrictions could be financial in nature or because the intervention could have the form of in-person workshops which not all users can or want to attend. A first rough estimation into this direction could be achieved basically without any additional computational costs. One could simply subset the existing results by the ranges that seem within a realistic range. While this would imply fewer considered values for each lever (i.e., a lower resolution), it could nevertheless be valuable to gain insights which policies would be Pareto-optimal under these more realistic conditions.

**Extended Policy Analysis**

More insights into the found policies could have been gained by extending the analysis of the found Pareto-optimal policies. For instance, selecting subsets of these policies to spot interesting patterns. Maybe it would become clear that the policies which already downranks posts even if not very certain that the post is misinformation (i.e., `rank_t` close to 50%), might go along with more prudent action in deleting posts and assigning strikes. A rather straight-forward way of doing this could be to use a library for interactive plots. These kind of patterns again might differ between the models. Furthermore, as these different actions are quite differently perceived by the general public, this could have real-world implications for what policies seem promising.

**Local Interactions**

Some of the differences between the SIT model and the other two models likely lead to a difference in local interactions. However, the data which was gathered when evaluating the policy candidates does not allow for analysis for such local interactions. This was not at the heart of this project, but could

be very interesting for two reasons. Firstly, analyzing the emerging patterns in these local interactions could be compared with what we can observe on real-world social media platforms. This could be used to validate, calibrate, and ground the model. Secondly, gathering suitable data could also enable the investigation of how the policy candidates could affect the commonly discussed local interaction types, such as echo chambers.

### 6.3.2. Ideas for More Advanced Research

After looking into the above ideas, that are closely related to this study, I would like to also outline some potential future research that would address broader or more advanced research questions. The model could be extended to pursue other research questions. These questions may be more advanced and could include many that are scientifically interesting and have societal relevance.

**Beliefs About Scientifically Contested Statements**

For instance, one might not only be interested in straight-forward topics where there is a scientifically well-established belief. Many topics, if not most, are not like that. In the real world, very few things can be cleanly proven. The scientific method cannot be expected to always provide us with clear black-and-white answers. Thus, phrases like ”following the science” can be detrimental because they suggest otherwise and are conversation stoppers. Limiting this project to such a relatively clear topic seemed like a necessary choice to stay within the scope of this project. However, relaxing this constraint could be a very useful direction. In line with this, one may also try to include an intervention of ’critical engagement’ where science-communicators are distributing posts on controversial topics by discussing the arguments of the current prominent conclusion of experts. This would enable critical engagement and joined sense-making rather than censoring.

**Multiple Beliefs, Bots, AI-generated Disinformation**

Other examples could focus on investigating the interaction between multiple beliefs (as people do not only post about a single topic) or including bots in the network (including for instance their particular characteristics such as the strength and number of their connections, as well as their location within the network). Another idea could be to explore what influence upcoming uses of artificial intelligence (AI) could have. This could include improvements in AI-based censorship or the increased capacity and use of AI to create and post misinformation or hate speech (similarly to the recent GPT-4Chan model (Kurenkov, 2022)).

# References


Al Atiqi, M., Chang, S., & Hiroshi, D. (2018). Agent-based approach to echo chamber reduction strategy in social media. *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*, 1301–1306. https://doi.org/10.1109/SCIS-ISIS.2018.00204

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, *50*(2), 510–530. https://doi.org/10.2307/2274239

Ali, R. N., Rubin, H., & Sarkar, S. (2021). Countering the potential re-emergence of a deadly infectious disease—information warfare, identifying strategic threats, launching countermeasures. *Plos One*, *16*(8). https://doi.org/10.1371/journal.pone.0256014

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629–76. https://doi.org/10.1257/aer.20190658

Austin, E. W., Borah, P., & Domgaard, S. (2021). COVID-19 disinformation and political engagement among communities of color: The role of media literacy. *The Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-58

Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. *Journal of Conflict Resolution*, *41*(2), 203–226. https://doi.org/10.1177/0022002797041002001

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Bauer, P. C. (2019). Conceptualizing and measuring polarization: A review. https://doi.org/10.31235/osf.io/e5vp8

Bennett, K. (2018). Environment of evolutionary adaptedness (EEA). *Encyclopedia of Personality and Individual Differences*, *1*(1627), 1–3. https://doi.org/10.1007/978-3-319-28099-8_1627-1

Beskow, D. M., & Carley, K. M. (2019). Agent based simulation of bot disinformation maneuvers in Twitter. *2019 Winter Simulation Conference (WSC)*, 750–761. https://doi.org/10.1109/WSC40007.2019.9004942

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(suppl_3), 7280–7287. https://doi.org/10.1073/pnas.082080899

Booth, R., Fermé, E., Konieczny, S., & Pérez, R. P. (2012). Credibility-limited revision operators in propositional logic. *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. https://www.researchgate.net/profile/Ramon-Pino-Perez/publication/268206433_Credibility-Limited_Revision_Operators_in_Propositional_Logic/links/546c752d0cf21e510f61a9f5/Credibility-Limited-Revision-Operators-in-Propositional-Logic.pdf

Bowlby, J. (1982). Attachment and loss: Retrospect and prospect. *American journal of Orthopsychiatry*, *52*(4), 664. https://doi.org/doi.org/10.1111/j.1939-0025.1982.tb01456.x

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., & Holman, B. (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science*, *84*(1), 115–159. https://doi.org/10.1086/688938

Bratman, M. E., Israel, D. J., & Pollack, M. E. (1988). Plans and resource-bounded practical reasoning. *Computational intelligence*, *4*(3), 349–355. https://doi.org/10.1111/j.1467-8640.1988.tb00284.x

Bromell, D. (2022). Challenges in regulating online content. *Regulating free speech in a digital age* (pp. 29–53). Springer. https://doi.org/https://doi.org/10.1007/978-3-030-95550-2_2

Carroll, M. D., Dragan, A., Russell, S., & Hadfield-Menell, D. (2022). Estimating and penalizing induced preference shifts in recommender systems. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning* (pp. 2686–2708). PMLR. https://proceedings.mlr.press/v162/carroll22a.html

Chang, J.-H., Zhu, Y.-Q., Wang, S.-H., & Li, Y.-J. (2018). Would you change your mind? An empirical study of social impact theory on Facebook. *Telematics and Informatics*, *35*(1), 282–292. https://doi.org/10.1016/j.tele.2017.11.009

Cockerill, K. (2007). Useless arithmetic: Why environmental scientists can't predict the future. *Environmental Practice*, *9*(3), 218–218. https://doi.org/10.1017/S1466046607070305

Coscia, M., & Rossi, L. (2020). Distortions of political bias in crowdsourced misinformation flagging. *Journal of the Royal Society Interface*, *17*(167). https://doi.org/10.1098/rsif.2020.0020

Courchesne, L., Ilhardt, J., & Shapiro, J. N. (2021). Review of social science research on the impact of countermeasures against influence operations. *Harvard Kennedy School Misinformation Review*. https://misinforeview.hks.harvard.edu/article/review-of-social-science-research-on-the-impact-of-countermeasures-against-influence-operations/

Daniel, K. (2017). Thinking, fast and slow. http://dspace.vnbrims.org:13000/jspui/bitstream/123456789/2224/1/Daniel-Kahneman-Thinking-Fast-and-Slow-.pdf

De Langhe, B., Puntoni, S., & Larrick, R. P. (2017). Linear thinking in a nonlinear world. *Harvard Business Review*, *2017*(May-June), 11. http://hdl.handle.net/1765/100448

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, *3*(01n04), 87–98. https://doi.org/10.1142/S0219525900000078

DiMaggio, P., Evans, J., & Bryson, B. (1996). Have American's social attitudes become more polarized? *American Journal of Sociology*, *102*(3), 690–755. https://doi.org/10.1086/230995

DROG. (2021a). Bad news. https://www.getbadnews.com/#intro

DROG. (2021b). Data-driven solutions to subversion. https://drog.group/

Du, E., Chen, E., Liu, J., & Zheng, C. (2021). How do social media and individual behaviors affect epidemic transmission and control? *Science of the Total Environment*, *761*, 144114. https://doi.org/10.1016/j.scitotenv.2020.144114

Duclos, J.-Y., Esteban, J., & Ray, D. (2004). Polarization: Concepts, measurement, estimation. *Econometrica*, *72*(6), 1737–1772. https://doi.org/10.1111/j.1468-0262.2004.00552.x

Erdös, P., & Rényi, A. (2011). On the evolution of random graphs. *The structure and dynamics of networks* (pp. 38–82). Princeton University Press. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.348.530&rep=rep1&type=pdf

Esteban, J., & Ray, D. (2012). Comparing polarization measures. *Oxford Handbook of Economics of Peace and Conflict*, 127–151. 10.1093/oxfordhb/9780195392777.013.0007

Esteban, J.-M., & Ray, D. (1994). On the measurement of polarization. *Econometrica: Journal of the Econometric Society*, 819–851. https://doi.org/10.2307/2951734

European Commission. (2021a). The Digital Services Act package. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

European Commission. (2021b). European Commission guidance on strengthening the Code of Practice on Disinformation. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021DC0262&qid

Fake news, propaganda, and disinformation: Learning to critically evaluate media sources: Infographic: Spot fake news. (2022). https://guides.library.cornell.edu/evaluate_news/infographic

Fiorina, M. P., Abrams, S. J. et al. (2008). Political polarization in the American public. *Annual Review of Political Science*, *11*, 563. https://doi.org/10.1146/annurev.polisci.11.053106.153836

Fränken, J.-P., & Pilditch, T. (2021). Cascades across networks are sufficient for the formation of echo chambers: An agent-based model. *Journal of Artificial Societies and Social Simulation*, *24*(3), 1. https://doi.org/10.18564/jasss.4566

Gadde, V., & Beykpour, K. (2018). Setting the record straight on shadow banning. https://blog.twitter.com/en_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning

Gausen, A., Luk, W., & Guo, C. (2021). Can we stop fake news? Using agent-based modelling to evaluate countermeasures for misinformation on social media. https://www.doc.ic.ac.uk/~cg1710/pub/mt21ag.pdf

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G., et al. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, *198*(1-2), 115–126. https://doi.org/10.1016/j.ecolmodel.2006.04.023

Grimm, V., & Railsback, S. F. (2012). Pattern-oriented modelling: A 'multi-scope'for predictive systems ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1586), 298–310. https://doi.org/10.1098/rstb.2011.0180

Grimm, V., Railsback, S. F., Vincenot, C. E., Berger, U., Gallagher, C., DeAngelis, D. L., Edmonds, B., Ge, J., Giske, J., Groeneveld, J., et al. (2020). The odd protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. *Journal of Artificial Societies and Social Simulation*, *23*(2). https://doi.org/10.18564/jasss.4259

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., Weiner, J., Wiegand, T., & DeAngelis, D. L. (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *science*, *310*(5750), 987–991. https://doi.org/10.1126/science.1116681

Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX* (tech. rep.). Los Alamos National Lab. (LANL), Los Alamos, NM (United States).

Hansson, S. O. (1999). A survey of non-prioritized belief revision. *Erkenntnis*, *50*(2), 413–427. https://doi.org/10.1023/A:1005534223776

Hansson, S. O., Fermé, E. L., Cantwell, J., & Falappa, M. A. (2001). Credibility limited revision. *The Journal of Symbolic Logic*, *66*(4), 1581–1596. https://doi.org/10.2307/2694963

Jaidka, K., Mukerjee, S., & Lelkes, Y. (2021). *An audit of Twitter's shadowban sanctions in the United States* (tech. rep.). EasyChair. https://scholar.google.fr/citations?view_op=view_citation&hl=fr&user=Y7_E1EIAAAAJ&citation_for_view=Y7_E1EIAAAAJ:qUcmZB5y_30C

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press. https://doi.org/10.1126/science.185.4157.112

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237. https://doi.org/10.1037/h0034747

Karlova, N. A., & Fisher, K. E. (2013). A social diffusion model of misinformation and disinformation for understanding human information behaviour. http://informationr.net/ir/18-1/paper573.html#.YvuWkpN78-Q

Kasprzyk, J. R., Nataraj, S., Reed, P. M., & Lempert, R. J. (2013). Many objective robust decision making for complex environmental systems undergoing change. *Environmental Modelling & Software*, *42*, 55–71. https://doi.org/10.1016/j.envsoft.2012.12.007

Khan, H. U., Nasir, S., Nasim, K., Shabbir, D., & Mahmood, A. (2021). Twitter trends: A ranking algorithm analysis on real time data. *Expert Systems with Applications*, *164*, 113990. https://doi.org/10.1016/j.eswa.2020.113990

Kim, H. K., & Tandoc Jr, E. C. (2022). Consequences of online misinformation on COVID-19: Two potential pathways and disparity by ehealth literacy. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.783909

Kopp, C., Korb, K. B., & Mills, B. I. (2018). Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to "fake news". *PloS One*, *13*(11). https://doi.org/10.1371/journal.pone.0207383

Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. *arXiv preprint*. https://doi.org/https://doi.org/10.48550/arXiv.1804.08559

Kurenkov, A. (2022). Lessons from the GPT-4Chan controversy. *The Gradient*. https://thegradient.pub/gpt-4chan-lessons

Kwakkel, J. H., & Pruyt, E. (2013). Exploratory modeling and analysis, an approach for model-based foresight under deep uncertainty. *Technological Forecasting and Social Change*, *80*(3), 419–431. https://doi.org/10.1016/j.techfore.2012.10.005

Kwakkel, J. H., Walker, W. E., & Marchau, V. A. (2010). Classifying and communicating uncertainties in model-based policy analysis. *International Journal of Technology, Policy and Management*, *10*(4), 299–315. https://doi.org/10.1504/IJTPM.2010.036918

Latané, B. (1981). The psychology of social impact. *American Psychologist*, *36*(4), 343. https://doi.org/10.1037/0003-066X.36.4.343

Lelkes, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*, *80*(S1), 392–410. https://doi.org/10.1093/poq/nfw005

Lempert, R. J. (2003). Shaping the next one hundred years: New methods for quantitative, long-term policy analysis. https://doi.org/10.7249/MR1626

Lempert, R. J., Groves, D. G., Popper, S. W., & Bankes, S. C. (2006). A general, analytic method for generating robust strategies and narrative scenarios. *Management Science*, *52*(4), 514–528. https://doi.org/10.1287/mnsc.1050.0472

Lewandowsky, S., Smillie, L., Garcia, D., Hertwig, R., Weatherall, J., Egidy, S., Robertson, R. E., O'Connor, C., Kozyreva, A., Lorenz-Spreen, P., et al. (2020). *Technology and democracy: Understanding the influence of online technologies on political behaviour and decision-making* (tech. rep.). Publications Office of the European Union. https://doi.org/10.2760/709177

Marchau, V. A., Walker, W. E., Bloemen, P. J., & Popper, S. W. (2019). *Decision making under deep uncertainty: From theory to practice*. Springer Nature. https://doi.org/10.1007/978-3-030-05252-2

Mason, C., van der Putten, P., & van Duijn, M. (2020). How identity and uncertainty affect online social influence. *Multidisciplinary International Symposium on Disinformation in Open Online Media*, 174–190. https://doi.org/10.1007/978-3-030-61841-4_12

Matthes, J., Rios Morrison, K., & Schemer, C. (2010). A spiral of silence for some: Attitude certainty and the expression of political minority opinions. *Communication Research*, *37*(6), 774–800. https://doi.org/10.1177%2F0093650210362685

Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology: General*, *64*(5). https://doi.org/10.1037/h0045106

Neumann-Böhme, S., Varghese, N. E., Sabat, I., Barros, P. P., Brouwer, W., van Exel, J., Schreyögg, J., & Stargardt, T. (2020). Once we have it, will we use it? a European survey on willingness to be vaccinated against COVID-19. https://doi.org/10.1007/s10198-020-01208-6

Nicholas, N. (2008). The black swan: The impact of the highly improbable. *Journal of the Management Training Institut*, *36*(3), 56. https://doi.org/10.5465/amp.25.2.87

Page, S. E. (2018). *The model thinker: What you need to know to make data work for you*. Basic Books, Inc.

Perra, N., & Rocha, L. E. (2019). Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific reports*, *9*(1), 1–11. https://doi.org/10.1038/s41598-019-43830-2

Pierri, F., & Ceri, S. (2019). False news on social media: A data-driven survey. *ACM Sigmod Record*, *48*(2), 18–27. https://doi.org/10.1145/3377330.3377334

Pierri, F., Perry, B. L., DeVerna, M. R., Yang, K.-C., Flammini, A., Menczer, F., & Bryden, J. (2022). Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific reports*, *12*(1), 1–7. https://doi.org/10.1038/s41598-022-10070-w

Pruyt, E., & Kwakkel, J. H. (2012). A bright future for system dynamics: From art to computational science and beyond. *Proceedings of the 30th International Conference of the System Dynamics Society, St. Gallen, Switzerland, 22-26 July 2012*. https://proceedings.systemdynamics.org/2012/proceed/papers/P1394.pdf

Pruyt, E., & Kwakkel, J. H. (2014). Radicalization under deep uncertainty: A multi-model exploration of activism, extremism, and terrorism. *System Dynamics Review*, *30*(1-2), 1–28. https://doi.org/10.1002/sdr.1510

Public Library, T. (n.d.). How to spot fake news. https://www.torontopubliclibrary.ca/spotfakenews/

Railsback, S. F., & Grimm, V. (2019). *Agent-based and individual-based modeling: A practical introduction*. Princeton university press.

Rajabi, A., Gunaratne, C., Mantzaris, A. V., & Garibay, I. (2020). Modeling disinformation and the effort to counter it: A cautionary tale of when the treatment can be worse than the disease. *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 1975–1977. https://tinyurl.com/43ewrnyp

Reddel, F. (2021). An extendable agent-based model of misinformation spread on social media. https://tinyurl.com/mrx9abbn

Reed, P., & Devireddy, V. (2004). Groundwater monitoring design: A case study combining epsilon dominance archiving and automatic parameterization for the NSGA-II. *Applications of multi-objective evolutionary algorithms* (pp. 79–100). World Scientific. https://doi.org/10.1142/9789812567796_0004

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, *28*(4), 394–412. https://doi.org/10.1080/0960085X.2018.1560920

Sanderson, Z., Brown, M. A., Bonneau, R., Nagler, J., & Tucker, J. A. (2021). Twitter flagged donald trump's tweets with election misinformation: They continued to spread both on and off the platform: Hks misinformation review. https://doi.org/10.37016/mr-2020-77

Sano, Y., Torii, H. A., Onoue, Y., & Uno, K. (2021). Simulation of information spreading on Twitter concerning radiation after the Fukushima nuclear power plant accident. *Frontiers in Physics*, *9*, 357. https://doi.org/10.3389/fphy.2021.640733

Sari, R. F., Ilmananda, A. S., & Romano, D. M. (2021). Social trust-based blockchain-enabled social media news verification system. *Journal of Universal Computer Science*, *27*(9), 979–998. https://doi.org/10.3897/jucs.68692

Spilsbury, L. (2018). Studies show lack of media literacy in students has negative impact. https://universe.byu.edu/2018/02/09/studies-show-lack-media-literacy-students-negative-impact/

Stray, J. (2021). Designing recommender systems to depolarize. *arXiv preprint arXiv:2107.04953*. https://doi.org/10.48550/arXiv.2107.04953

Thompson, E. L., & Smith, L. A. (2019). Escape from model-land. *Economics*, *13*(1). https://doi.org/10.5018/economics-ejournal.ja.2019-40

Treen, K. M. d., Williams, H. T., & O'Neill, S. J. (2020). Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, *11*(5), e665. https://doi.org/10.1002/wcc.665

TwitterHelpCenter. (2021). COVID-19 misleading information policy. https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy

TwitterSupport. (2020). *Sharing an article can spark conversation, so you may want to read it before you tweet it.* https://twitter.com/twittersupport/status/1270783537667551233?lang=en

Vespignani, A. (2012). Modelling dynamical processes in complex socio-technical systems. *Nature Physics*, *8*(1), 32–39. https://doi.org/10.1038/nphys2160

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Walker, W. E., Marchau, V. A., & Kwakkel, J. H. (2013). Uncertainty in the framework of policy analysis. *Public policy analysis* (pp. 215–261). Springer. https://doi.org/10.1007/978-1-4614-4602-6_9

Wang, C., Koh, J. M., Cheong, K. H., & Xie, N.-G. (2019). Progressive information polarization in a complex-network entropic social dynamics model. *IEEE*, *7*, 35394–35404. https://doi.org/10.1109/ACCESS.2019.2902400

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019a). Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, *240*. https://doi.org/10.1016/j.socscimed.2019.112552

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019b). Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, *240*, 112552. https://doi.org/10.1016/j.socscimed.2019.112552

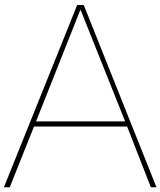
Watson, A. (2021). Topic: Fake news worldwide. https://www.statista.com/topics/6341/fake-news-worldwide/#dossierKeyfigures

Wilson, S. L., & Wiysonge, C. (2020). Social media and vaccine hesitancy. *BMJ Global Health*, *5*(10). https://doi.org/10.1136/bmjgh-2020-004206

Wojcik, S., & Hughes, A. (2021). Sizing up twitter users. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Woodruff, M., & Herman, J. (2013). *pareto.py: An $\epsilon$-nondomination sorting routine*. https://github.com/matthewjwoodruff/pareto.py

Yousefinaghani, S., Dara, R., Mubareka, S., Papadopoulos, A., & Sharif, S. (2021). An analysis of COVID-19 vaccine sentiments and opinions on Twitter. *International Journal of Infectious Diseases*, *108*, 256–262. https://doi.org/10.1016/j.ijid.2021.05.059



# Appendix A: Literature Review

## A.1. Methods

The approach to the literature review had two phases – 1) a systematic search within Scopus, and 2) a forward reference search and backward reference search. The final search query of the first stage is depicted in Table A.1. It combines sets of synonyms. Each of the three sets represents one relevant semantic cluster. Employing a model in order to study the proliferation of opinions is generally called *opinion dynamics*. Thus, this term is included in the search query.

**Table A.1:** Search Query for the Literature Search within Scopus.

Each of the terms within parentheses were added in a field to search within title, abstract, and keywords.

| | |
|---|---|
| **Search Query** | (”opinion dynamic*” OR ”**agent-based**” OR ”agent based” OR ”ABM”)<br>AND<br>( ”**social media**” )<br>AND<br>( ”**misinfo***” OR ”disinfo*” OR ”fake news” )) |
| **Number of Papers resulting from Search Query** | 26 |
| **Number of Papers after Title-& Abstract-Filtering** | 14 |
| **Number of Papers after In-Depth Examination** | 11 |

Because numerous papers use ABMs to do what is in essence opinion dynamics without mentioning ’opinion dynamic’ explicitly. To enable finding those papers as well, ABM-related terms were included in the first semantic cluster. In this fashion, a larger set of the relevant papers was collected. The second semantic cluster is self-explanatory. The third and final cluster relates to misinformation. Due to the various definitions of misinformation, the third set was slightly extended by very related terms. The resulting papers were first filtered by their titles, then by their abstracts.

## A.2. Findings

Table A.2 provides an overview over the selected papers along the dimensions of the identified themes. Below the table, the themes that are less relevant for this project are discussed.

**Table A.2:** Literature

For discrete *Belief Representations*, the number of values any one belief can take are indicated in parentheses. For Kopp et al., 2018, it is 0 because beliefs are not explicitly, but by agent characteristics which are passed on via evolution. The *Belief Update* value *Deffuant* means Deffuant-like and refers to the updating rule by Deffuant et al. (2000). To keep the Uncertainty theme at a high-level, it is aggregated to whether the papers applied DMDU methods or not.

| | *Theme 1* | | *Theme 2* | *Theme 3* | *Theme 4* |
|---|---|---|---|---|---|
| **Paper** | **Belief Representation** | **Belief Update** | **Agent Types** | **Space** | **DMDU** |
| Wang et al. (2019) | discrete (n) | sample | 1 | Barabási-Albert | No |
| Sano et al. (2021) | discrete (3) | Deffuant | 3 | Barabási-Albert | No |
| Du et al. (2021) | continuous | Deffuant | 1 | Grid | No |
| Beskow et al. (2019) | continuous | Deffuant | 3 | ? | No |
| Ross et al. (2019) | continuous | Deffuant | 1 | Barabási-Albert | No |
| Rajabi et al. (2020) | continuous | Deffuant | 3 | Barabási-Albert | No |
| Mason et al. (2020) | continuous | Deffuant | 1 | fully connected | No |
| Al Atiqi et al. (2018) | continuous | Deffuant | 2 | Barabási-Albert | No |
| Coscia & Rossi (2020) | continuous | - | 2 | Erdős-Rényi | No |
| Kopp et al. (2018) | discrete (0) | evolution | 1 | ? | No |
| Ali et al. (2021) | continuous | differential equation | 3 | Grid | No |
| Fränken & Pilditch (2021) | discrete (2) | Bayesian | 1 | Grid | No |
| Sari et al. (2021) | discrete (2) | sample | 2 | Barabási-Albert | No |
| Gausen et al. (2021) | discrete (2) | sample | 1 | ? | No |

### A.2.1. Theme 2: Agent Types

The number of agent types in these papers were either one, two, or three. In papers with just one agent type, information is produced and consumed by everyone. In the papers with two distinct agent types (Al Atiqi et al., 2018; Coscia and Rossi, 2020; Sari et al., 2021), the types differ with respect to their roles. One agent type is providing news, the other agent type is a normal user, consuming the news. In the papers with three distinct agent types (Ali et al., 2021; Beskow and Carley, 2019; Rajabi et al., 2020; Sano et al., 2021), types differ with respect to their beliefs. Two types being on opposite ends regarding their beliefs (e.g., one defending the truth, the other one aiming to disinform), and one type initially being undecided. One might make a case for seeing the choice for three types as an extension of the choice for two types; One type being the normal user, and the other two being opposing news providers.

### A.2.2. Theme 3: Space

In some papers, the kind of space chosen and implemented was not mentioned. However, most papers (at least 9) chose to implement a network to represent the social media connections. The most frequently chosen type of network type within those papers was the Barabási-Albert graph (Barabási and Albert, 1999). This graph type uses preferential attachment during the graph construction. This leads to a power law distribution in the number of the nodes' connections (i.e., the agents' connections). As such, it is well-suited to represent social media networks. The other type of network that was implemented is the Erdős-Rényi graph (Erdös and Rényi, 2011). This alternative way of creating a random graph, leads to low clustering, which is not very close to the connections within social media networks. Another approach was to either build or approximate a random network by using a grid (Ali et al., 2021; Du et al., 2021; Fränken and Pilditch, 2021). Here, each agent was given a random position. The distance to other agents determined the connection to other agents. This was done either by the distance translating into the connection strength or by connecting each agent to the closest five

agents. However, also this approach leads only to very minimal differences between the amount of connections the agents have. And consequently it is less suitable to represent social media network.

## A.3. Details of Rationale Behind Deduced Research Gap

The deduced research objective stems from the two high-level topics of exploring uncertainty, and including multiple objectives. They are discussed in the following two subsections.

### A.3.1. Exploring Uncertainties

Our system of interest is riddled with various kinds of uncertainties. For such systems, exploring such uncertainties holds to promise of arriving at significantly more robust conclusions than applying the traditional predictive forecasting approach. At the same time, the previously performed literature review had not surfaced any research that substantially incorporated uncertainties. As the aim behind this project is to improve the information and understanding that is available to decision-makers, utilizing the exploratory approach has high expected value.

However, within the limited scope of a master thesis, only a small part of the uncertainty space may be explored. Here again, I aimed to make this choice based on what would likely be the most valuable to decision-makers. EMA enables exploring different kinds of uncertainties – parametric uncertainties, structural uncertainties, and method uncertainties (Kwakkel and Pruyt, 2013).

What exact value a parameter is set to can have a substantial impact on the emerging higher-level patterns. As such, it would be valuable to include exploring parametric uncertainty in the project. The structural uncertainty of how people are modeled to update their beliefs is a key difference in the papers from the literature review. Furthermore, this structural uncertainty is not just any structural uncertainty, but it is arguably the one at the heart of the models. Consequently, it seems reasonable to expect that a different way of modeling how agents update their beliefs could be crucial to the emerging behavioral patterns. In other words, the modelers choice of how to model the agents' belief update could have a large impact on what policies seem optimal. This in turn would impact any decision-maker who aims to use the research to arrive at evidence-based policies. Thus, the choice of the belief update function could easily have substantial real-world consequences. However, as this point has not been discussed in the literature that I could find, it seems like this is currently a bit of a blind spot in the research community. These points made the structural uncertainty of the belief update a prime candidate for the main focus of the project. In contrast, the method uncertainties were judged as a weaker candidate for this project. The main reason being that the current discussion is not revolving around *which* EMA method to use, but rather to investigate *whether* to utilize exploratory modeling at all. Along a similar line of reasoning, I decided to keep the previously mentioned BDI assumption. The aim here was to stay closer to the related work in order to enable sensible comparisons and conclusions. A summary of the above considerations can be found in Figure A.1.

Following these considerations, I chose the structural uncertainty of the belief update a as the main focus of the project. The exploration of some parametric uncertainty is included as it seems to be a good deal when considering the ratio between value added (i.e., making results more robust and uncovering more relevant information) and the corresponding costs (i.e., amount of additional work and computation)

### A.3.2. Multiple Objectives

As mentioned previously, grand challenges involve many different stakeholders. Naturally, they tend care about multiple objectives. We want to arrive at solutions that are beneficial for collection of stakeholders. Thus, we want to evaluate solutions not only based on a single objective, but on the main objectives of the various stakeholders. However, the literature review has not lead to any papers that explicitly consider multiple objectives. While it is easier to mainly optimize for only one objective, this

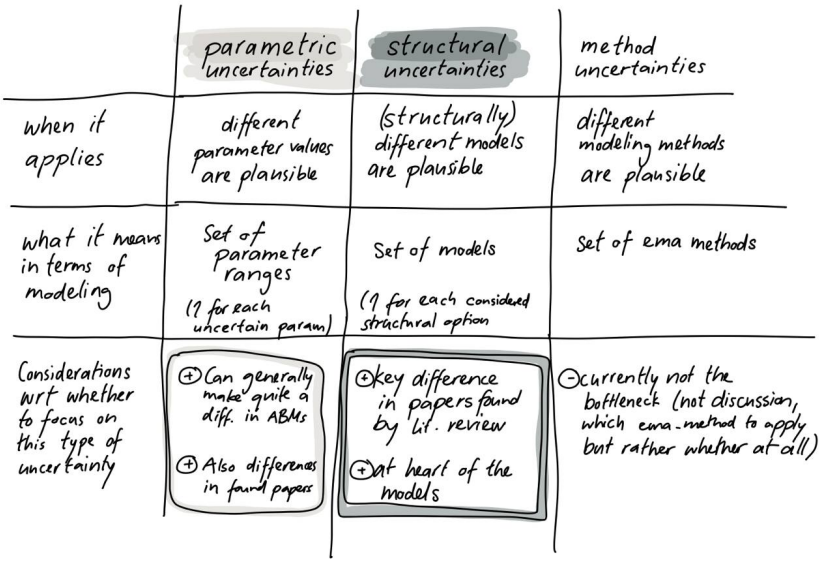



**Figure A.1:** Different Kinds of Uncertainties

can lead to a considerable amount of friction when aiming to implement the policy that has been decided on. For instance, many felt that the strong focus on decreasing the expected number of COVID-19 cases crowded out considerations for other objectives such as minimizing economic hindrances or minimizing the negative effects on mental health of people of various groups.

If only one objective is included, the performance of each policy can readily be compared to the performance of another policy. When multiple objectives should be included, trade-offs between objectives complicate this endeavor. If policy 1 performs better than policy 2 with respect to objective A, but worse with respect to objective B, it is unclear which policy would be the better solution. To find the most promising policies, it is possible to filter for policies that perform in each objective *at least as good* or *better* than the other policies. This we call a *non-dominated sort*, which results in the set of *Pareto-optimal policies* (Woodruff & Herman, 2013). Due to these thoughts, the Pareto-optimal policies are at the focus of this project. While this is in line with research in the field of exploratory modeling and analysis, it had not been done in the research which is on the content-level related (i.e., the research that was found in the literature review).

# B

# Appendix B: DMDU

## B.1. Levels of Uncertainty

In the following, the five levels are described.

| | Description of the level of uncertainty | Approaches for dealing with the level of uncertainty |
|---|---|---|
| Level 1: marginal uncertainty | Recognizing that one is not absolutely certain, but (dis)-qualifying uncertainty as a marginal issue | Performing sensitivity analysis on model parameters by marginally changing default values by e.g. ±10% |
| Level 2: shallow uncertainty | Being able to enumerate multiple alternatives and their corresponding probabilities | Enumerating multiple possible futures or generating alternative model outcomes and specify their probability of occurrence, e.g. Monte Carlo simulations with probabilistic interpretation |
| Level 3: medium uncertainty | Being able to enumerate multiple possibilities and rank order possibilities in terms of perceived likelihood. However, how much more likely or unlikely one alternative as compared to another cannot be specified | Enumerating multiple possible futures or alternative model structures and ranking them in terms of perceived likelihood, e.g. identifying different types of behavior patterns and ranking them as more, less, or equally likely |
| Level 4: deep uncertainty | Being able to enumerate/generate multiple possibilities without being able to rank them in terms of how likely or plausible they are | Enumerating multiple possible futures or specifying multiple alternative model structures and generating alternative outcomes without specifying their likelihood, e.g. exploratory SD with uncertainty and robustness analyses |
| Level 5: recognized ignorance | Being unable to enumerate multiple possibilities, because one does not or cannot know the generative mechanisms at play nor the possibilities that may be generated | Accepting the possibility of being wrong or surprised, because existing mental or formal models are known to be inadequate, e.g. planning for surprise |

**Table B.1:** Different Levels of Uncertainties

Table from Pruyt and Kwakkel (2014), which itself is adjusted from Kwakkel et al. (2010), Pruyt and Kwakkel (2012), and Walker et al. (2013)

**Level 1: Marginal Uncertainty**

Acknowledging that some aspects are not completely certain, for instance the measurement of some model parameters. Still a deterministic model, but something can be gained by performing sensitivity analysis, i.e., slightly adjusting those model parameters and explore the effects of these changes on the outcomes. One example situation would be to at what exact time your daily train arrives on the platform.

**Level 2: Shallow Uncertainty**

Some components are stochastic. This stochasticity can lead to a few different futures, each with their corresponding probability. Probability calculations and statistics can be used to guide decisions within this level of uncertainty. A simple example could be deciding whether you want chickpeas or lentils for lunch by tossing a coin.

**Level 3: Medium Uncertainty**
In this level, the few different futures canned be assigned specific probabilities. Luckily, in this uncertainty level, it we still have enough information to evaluate how good or bad a policy would be in each of the different futures would be. Like this, we can find and opt for policies that are performing sufficiently well over many different potential futures.

**Level 4: Deep uncertainty**
This level can be split into 4a and 4b. In 4a, there are not only a few different futures possible (as in level 3), but many. In 4b, the only thing we know is that we do not know. In 4a, we can still aim to apply the multi-model approach (Page, 2018). With this approach, we build multiple plausible models and try to learn from this ensemble of models. In 4b, we do not know how patterns were generated nor what kind of patterns might be generated. Here, we just can prepare to be surprised. *Black swan* events, which Taleb defines as being very surprising, having extreme impact, and being explainable only after their discovery (Nicholas, 2008) belong to this level. An example for 4b could be the historical discovery of black swans. An example for 4a might be that there are multiple theories of how people update their beliefs without knowing how likely each of the state-of-the-art theory is, or how many more and better theories there might be. For these levels, the *Decision-Making under Deep Uncertainty* (DMDU) approach is required to handle all uncertainties.

**Level 5: Recognized ignorance**
This is close to the opposite extreme of the uncertainty dimension. At this level, we can only be aware that we will likely be surprised because we know that our models are inadequate.

Even if we have concrete probabilities (as in the lower levels of uncertainty), as humans, we tend to have a hard time using and combining them in a consistent manner (e.g., Kahneman et al., 1982; Kahneman and Tversky, 1973). This makes it difficult for us to predict what the outcomes of a given policy might be when it is applied to system that includes many uncertainties and/or types of uncertainty. Here again, this is where computers come in. Their strength of accurate and fast computation means that they can complement and support human decision-making.

# C

# Appendix C: MisinfoPy Model

While the model was designed and implemented following the previously mentioned XLRM framework, the description of the model follows the ODD (**O**verview, **D**esign concepts, **D**etails) protocol (Grimm et al., 2006), as updated by Grimm et al., 2020. The ODD protocol was chosen because it is a common standard for describing individual- and agent-based models. Figure C.1 provides an overview how the following sections fit into the Overview, Design concepts, and Details structure. Furthermore, also an overview of the model from the XLRM perspective is provided in Figure C.2

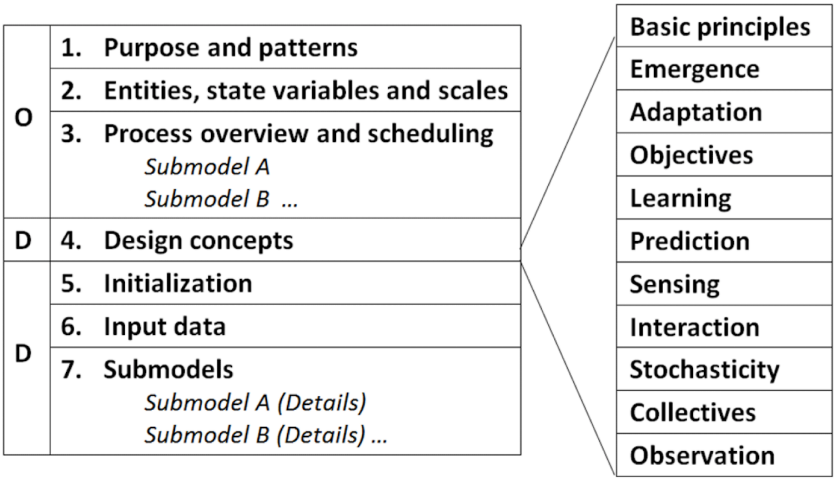



**Figure C.1:** ODD Overview

The graphic provides an overview how the following sections fit into the Overview, Design concepts, and Details structure. This is the original graphic from (Grimm et al., 2020).

Before we dive into all the separate parts, a visual overview is provided in Figure C.3

## C.1. Purpose and Patterns

The purpose of the model is to enable the testing to what extent the structural uncertainty in the agents' belief update function influences which Pareto-optimal policies are found.

In order to test and establish in how far the model is suitable to investigate this purpose, modelers can think of a number of patterns that should be emerging from the model. Depending on whether and how strongly these patterns appear, one might judge the model's fit-for-purpose as better or worse.

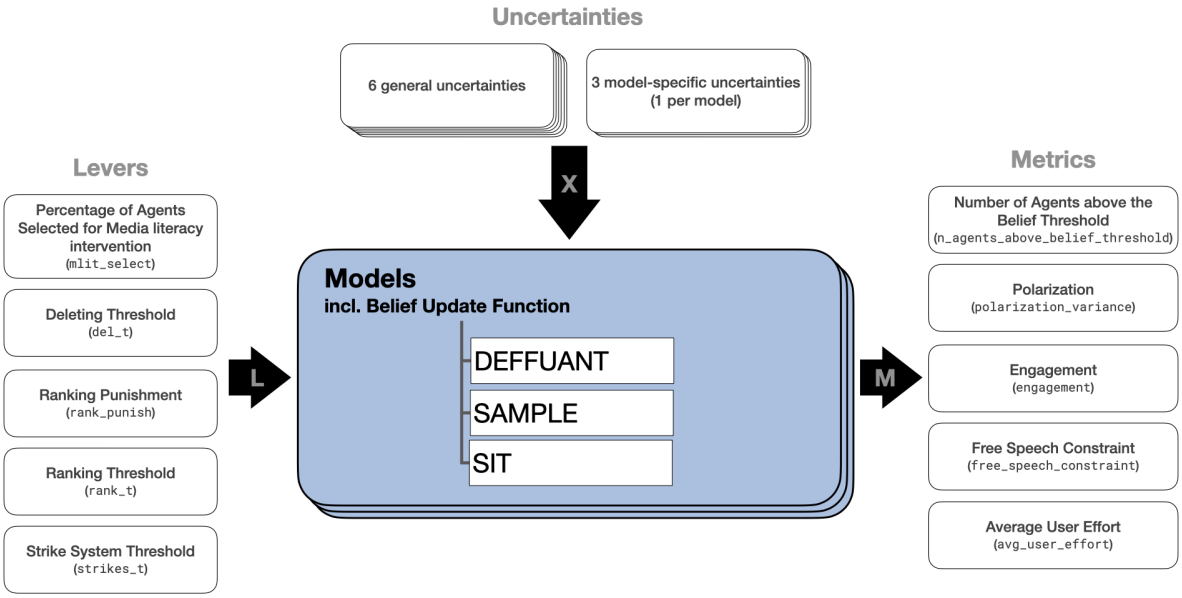



**Figure C.2:** XLRM MisinfoPy Model

XLRM perspective onto the MisinfoPy model.

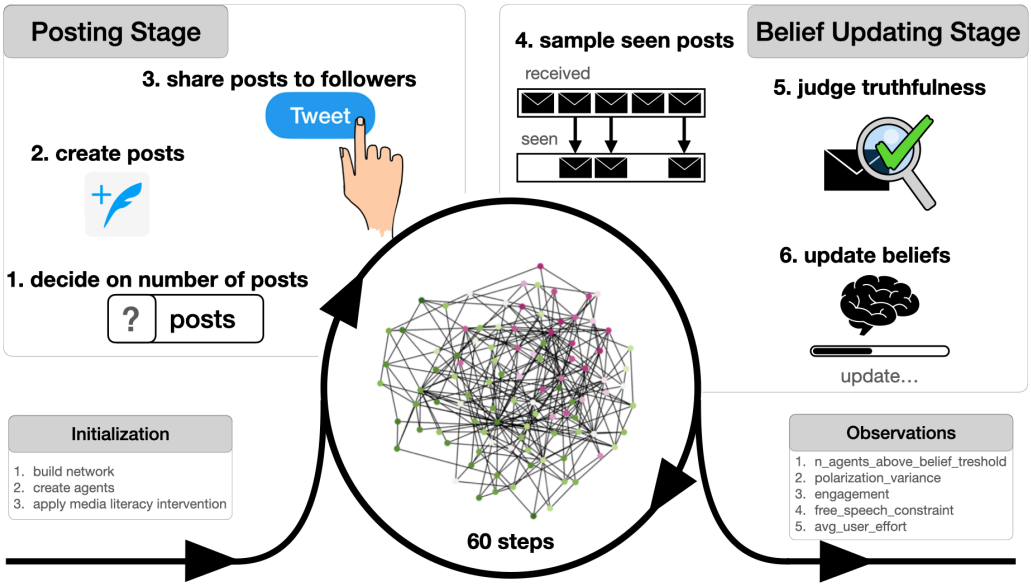



**Figure C.3:** Visual ODD

## C.1.1. Patterns

Three patterns have been identified. For each of these patterns, three aspects are documented: (1) why we have reason the belief that this pattern exists in the real world, (2) why we have reason to believe that the pattern is relevant for the purpose of the model, and (3) how it could be investigated in how far the pattern emerges from the simulation of the model. While these patterns should be relatively independent of the belief update functions, we could still be more confident about the conclusions if we evaluate each pattern for each of the three models.

**Pattern 1: More Extreme Posts → Higher Levels of Engagement**
According to (Watson, 2021), engagement levels are higher for content that represents more confident and extreme beliefs. This pattern is relevant for the purpose of the model because if we would not find this pattern, it would be evidence that the engagement metric might be misguided. As each metric is relevant to select Pareto-optimal policies, whether such a pattern emerges or not is important information to judge the models fit for purpose. The pattern that we would expect is depicted in Figure C.4. For a couple of model runs, the beliefs of all posted posts is depicted on one horizontal line. Each dot represents one posted post, where the location on the x-axis represents the posts belief. The y-axis position of the whole line represents the value of the engagement for the run. We would expect that in runs where most beliefs were around the uncertain belief of 50, are depicted as lines lower on the plot (indicating relatively low engagement levels). At the same time, runs that in which most posts lie close to the more extreme belief values (i.e., close to 0 or 100), the line of those runs is depicted higher in the plot, indicating higher levels of engagement. Consequently, we would expect a U-shaped pattern similarly to the one highlighted in the figure.

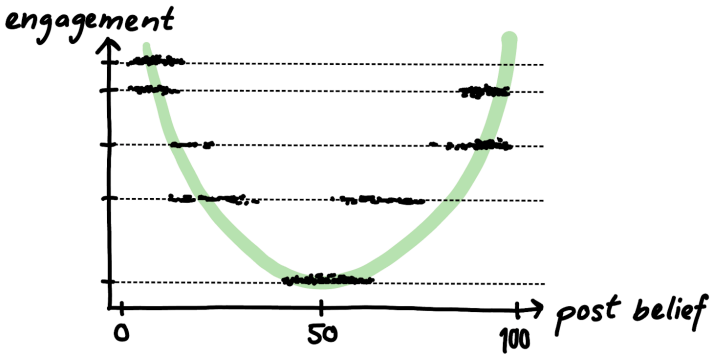



**Figure C.4:** Pattern 1: More Extreme Posts → Higher Levels of Engagement

In order to investigate whether this pattern emerges, we would want run the model a few times. The runs could differ in their instantiation of uncertainties and lever. This could even be helpful in order to arrive at a more divers set of resulting horizontal lines. In each run, the beliefs represented in each and every posted post would need to be saved. This data would need to be saved in combination with the run-id and the resulting engagement metric value at the end of the model run.

**Pattern 2: Stronger Lever Activation → Lower Percentage of False Posts Seen**
Activating policy levers that are based in literature and that are designed to mitigate the issue of misinformation on social media are assumed to decrease the percentage of false post seen. This pattern would be in line with intuitive reasoning about the system. Though often implicit, this assumption is very commonly made in related work. This pattern is relevant, because if we would find that the model does not produce this pattern, it would erode our trust that using the levers will even have the hoped for effect *direction*. To evaluate whether the model produces the pattern, three different lever activation levels could be defined. For instance, low, medium, and high. The model could be run a couple of times with each of the three lever activation values. Every run, the number of false posts seen, as well as the number of false posts created are gathered by aggregating over the whole agent population. From these two numbers, the proportion of false posts that were actually seen can be calculated and a dot can be placed at the corresponding position. From all these dots, also boxplots could be drawn to ease interpretation of the data. A visual impression of what we would expect is provided in Figure C.5. We would expect that the distribution along the y-axis has the biggest spread, the biggest boxplot, and the highest position for the low lever activation levels. With increasing lever activation levels, we would

expect the distributions to become more narrow, boxplots to become smaller, and the position of the boxplots to decrease with respect to the y-axis.

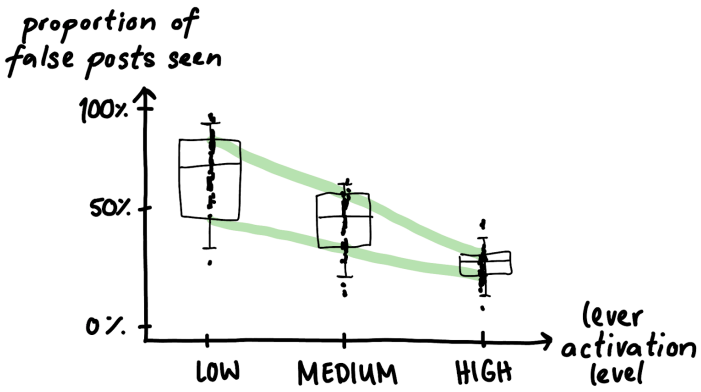



**Figure C.5:** Pattern 2: Stronger Lever Activation → Lower Percentage of False Posts

**Pattern 3: Behavior Distribution (`n_posts_posted`) over Agents: Small Percentage of Agents Produces the Majority of Posts**

Data from Twitter indicates that a majority of the posts stems from only a small proportion of agents(Wojcik & Hughes, 2021). This pattern is relevant because it would be a strong point to justify the relatively strong measure of the strike system, which involves locking users accounts – potentially indefinitely. If only a small proportion of the users are virtually vandalizing online discussion and joined sense-making it could be a reasonable measure to hinder this trouble-maker minority in order to increase the value the vast majority of users can get out of the platform. Furthermore, this is an important pattern observed on the real-world Twitter and thus would indicate that the model is reasonably well approximating this aspect. The pattern that we would expect would look somewhat like the green lines in Figure C.6, and would not cross the depicted red line. To gather the data needed for this investigation, one would need to run the model a few times, while each time gathering the information of how many posts the individual agents have posted. Such a list could then be sorted in descending fashion before calculating and plotting the cumulative sums over both agents (to arrive at the percentage of agents), and the number of posts (to arrive at the proportion of posts posted by that percentage of agents).

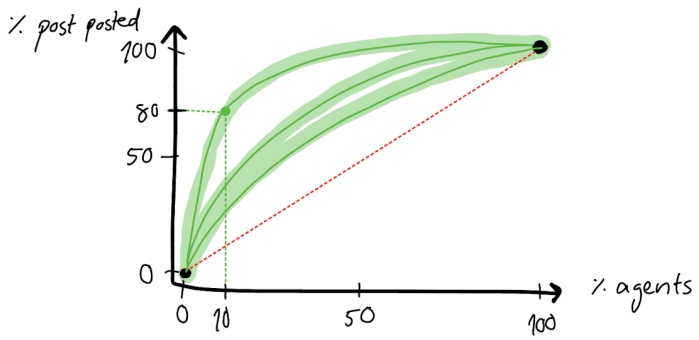



**Figure C.6:** Pattern 3: Small Percentage of Agents Produces the Majority of Posts

## C.2. Entities, State Variables and Scales

The model includes the following entities: agents representing the social media users (i.e., normal users and disinformers), network nodes (i.e., the virtual non-geographical location of the agents), and the global environment which represents the broader social media context. According to the ODD protocol, this section should include *state variables* (dynamic over a model run), but no *parameters* (static over a model run). The global environment of this model only includes the state variable representing time. Consequently, most details about the global environment and its parameters are provided in Section C.7.

### C.2.1. Spatial Units: Network Nodes

The environment of agent includes everything that is both, outside of the specific agent and within the model. This means that other agents are part of the agent's environment. Agents gather information from their environment and provide information to their environment. A suitable environmental structure of a social media platform is a network in which nodes represent agents and edges represent the connections between agents (in this case, agents 'following' each other). The number of agents an agent is connected to is generally referred to as its *degree* or the degree of its node. Network nodes are "locations" in the non-geographic space of the model. Figure C.1 shows an overview over the main state variable of a network node.

**Table C.1:** State Variables of the Network Nodes

| State Variable | Scale |
|---|---|
| `edges` | - represents the connections between users<br><br>- **list of edges**<br>(at least `n_edges` long (see `ref_scenario`))<br><br>- each edge includes the information of:<br>- **from** which **to** which agent the edge is<br>- the **`tie_strength`** between the two agents,<br>i.e., how strong their relationship is (this is encoded as the `edge_weight`)<br>- the range of edge weights is [0,100]<br><br>- **static** over a simulation run (edges, incl. edge weights) |

Many different types of networks exist. As social media networks in the real world have high clustering, it is beneficial if the modelled network also has a similar degree distribution. Consequently, a scale-free network with its approximately power-law degree distribution is a suitable representation. *Barabási-Albert* networks fulfill this characteristic via *preferential attachment*(Barabási & Albert, 1999). During the creation of the network, new nodes have a higher probability to be attached to nodes that already have a higher degree (i.e., they are preferably attached to those nodes). This graph type is also a common choice in related work (e.g., Al Atiqi et al., 2018; Rajabi et al., 2020; Ross et al., 2019; Sano et al., 2021; Sari et al., 2021; C. Wang et al., 2019). An example degree distribution of a network with 100 agents is depicted in Figure C.7.

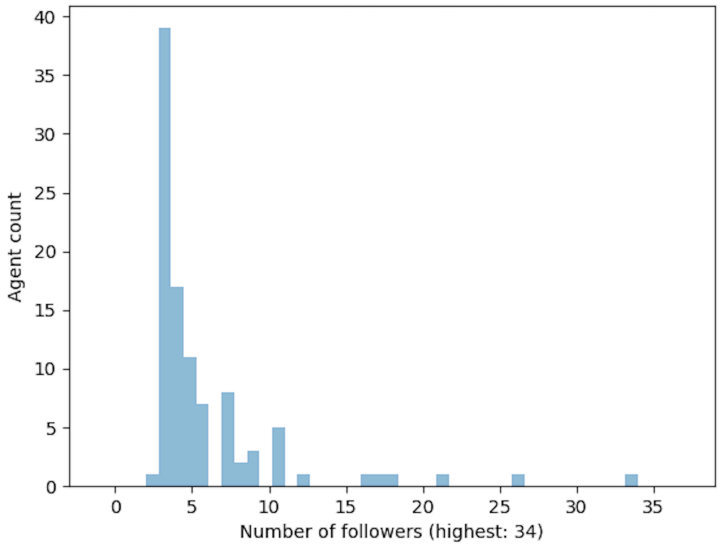
Other choices in related work include the Erdős-Rényi graph, fully connected networks, and grid-based networks (see Table A.2). However, these networks do not display a power-law degree distribution and thus tend to be less realistic representations of social media networks.

### C.2.2. Agents

The tables for the state variables of both agent types (Figure C.2 for Normal Users and Figure C.3 for Disinformers) should largely include enough detail to not require additional explanation. However, for ease of delving into the project, here are still a few high-level notes about the beliefs and the media literacy of agents. Each belief represents the subjective probability of an agent about a statement. This project restricts itself to just one belief. The belief is the following statement in Section C.2.2. A belief of 50 indicates complete uncertainty. The beliefs of 0 and 100 represent the maximal certainty that the

**Figure C.7:** Power-Law Degree Distribution in a Barabási-Albert Network

An example of what a power-law degree distribution looks like. The histogram depicts how many agents have how many connections within an example graph. The graph is of the type Barabási-Albert graph, has 100 agents, and an average number of connections (i.e., degree) of three. Most agents have a small number of connections, while very few agents have a very high number of connections.




statement does not or does hold, respectively.

**The VAX statement**

The agents' belief refers to their agreement or disagreement with the below statement:
*"For most people, it is safer to get their COVID-19 vaccine jabs than it is to decline them (and consequently face COVID-19 encounters without the inoculation)."*

The media literacy of an agent represents the skill of a user to judge whether a post is truthful or not. This attribute is relevant because it influences whether agents adjust their beliefs based on a post. Agents update their beliefs based on a belief only if they judge the post as truthful.

Agents have two key behaviors – **1) sharing posts** and **2) updating their beliefs**. The former is about outgoing information towards the agent's environment, the latter about incoming information from the agent's environment. These two processes are explained in Section C.3 (overview) and Section C.7 (details).

### C.2.3. The Global Environment

The choice of the resolution of time was made with consideration of compatibility with levers and metrics, ease of intuitive interpretation, and computational feasibility. The chosen resolution of 1 step representing 1 day passing by in the real world scores high in all three aspects. The next-higher intuitive time resolution would be 1 model step representing 1 hour. However, because the distribution of posts over the hours in a day is not required for the approaching the main research question. None of the levers would apply differently depending on the hour in which a post was posted. The analogous holds for the objectives. As such, such a high resolution is not required and would only unnecessarily increase computational costs. The next-lower intuitive resolution would be that 1 model step represents 1 week. This however, would be at odds with the strike system intervention. This intervention benefits from a

**Table C.2:** State Variables of Normal Users

| State Variable | Scale |
|---|---|
| `followers` | - represents the number of followers a person has<br><br>- **list of agents**<br><br>- numbers of followers in range `[n_edges, n_agents-3]`<br><br>- **static** over a simulation run |
| `following` | - represents the number of how many people a person is following<br><br>- **list of agents**<br><br>- numbers of followings in range `[n_edges, n_agents-3]`<br><br>- **static** over a simulation run |
| `vocality` | - represents how many posts an agent wants to post each step (represented by a normal distribution)<br><br>- a dictionary of two entries defining the normal distribution<br>- **mean**: `mean_normal_user`<br>- (`std_dev`: sigma (0.7))<br>(always the same for everyone → parameter, not state variable)<br><br>- **static** over a simulation run |
| `media_literacy` | - represents the media literacy skills of an agent, specifically how skilled they are at distinguishing misinformation from true information<br><br>- **enumeration**, can take two values: **`LOW`** and **`HIGH`**<br><br>- **static** over a simulation run |
| `beliefs` | - represents a person's belief with respect to the "VAX" statement<br><br>- dictionary of **beliefs** and their **values**<br><br>- for this project:<br>- only one belief option: **VAX**<br>- value range: **`[0,100]`**<br><br>- **dynamic** over a simulation run |
| `received_posts` | - represents the posts that a person has in their feed<br><br>- **list of posts** that the agent received within one time step<br><br>- **dynamic** over a simulation run |
| `visible_posts` | - represents the posts a persons has posted<br><br>- **list of all posts** that the agent successfully posted (i.e., were not deleted)<br><br>- **dynamic** |
| `n_strikes` | - represents the number of strikes a user has already accumulated<br><br>- **integer** number of strikes<br><br>- **dynamic** over a simulation run |
| `blocked_until` | - representing the simulation time from which on the agent is allowed to post again (if it had been blocked due to the strike system)<br><br>- **integer** value of time step<br><br>- range: **`[0,math.inf]`** (agents can be blocked infinitely)<br><br>- **dynamic** over a simulation run |

**Table C.3:** State Variables of Disinformers

| State Variable | Scale |
|---|---|
| `followers` | same as NormalUser |
| `following` | same as NormalUser |
| `vocality` | - only difference from NormalUser:<br>- **mean**: `mean_disinformer` |
| `beliefs` | - differences from NormalUser:<br>- value range: **`[0,10]`**<br><br>- **static** over a simulation run |
| `received_posts` | same as NormalUser |
| `visible_posts` | same as NormalUser |
| `n_strikes` | same as NormalUser |
| `blocked_until` | same as NormalUser |

resolution that is at least as high as 1 model step representing 1 day in order to block agents for a specific number of days. Additionally, evaluating the outcomes is intuitively harder to grasp if they are

**Table C.4:** State Variable of the Global Environment

| State Variable | Scale |
|---|---|
| `step` | - represents time (1 step = 1 day)<br>- **integer**<br>- in range `[0, steps]` → **`[0,60]`**<br>- **dynamic** |

per week than if they are per day. Furthermore, most data about when users are posting in the real world, do not explicitly include the week number, but the day. Consequently, the resolution of 1 model step representing 1 day seems to be the best solution.

## C.3. Process Overview and Scheduling

Each step of the model is split into two stages – the posting stage and the belief updating stage. While this is not a one-to-one representation of the real world, I chose this setup because it enables avoiding artifacts that are far from reality. If these two behaviors would not be split into separate stages, it would for instance cause that the agent that is first assigned to post and update their beliefs could not update its beliefs because it certainly would not yet have received any posts within this model step. Only for the very last agent it could be assumed that they certainly have received the posts from all the agents whom the agent is following. Both stages are explained here on a high level. In line with the ODD protocol, more details for these stages or submodels are provided in Section C.5 (initialization details), and in Section C.7 (all remaining details).

### C.3.1. Posting Stage

In the posting stage, the entities of type 'Normal User' and 'Disinformer' decide how many posts they would like to post, create the posts, and post them to their followers. No state variables are adjusted in this step. The order of agents performing these actions is irrelevant because it only affects the order of the posts to which agents will update their beliefs in the next stage. However, as the agents process all received posts the exact same way, it does not matter in which order they see them[1].

In this stage, agents first decide on how many posts they want to create and post. Then, they create the posts and share them to their followers. How many posts an agent posts in a given step is modeled as a function of their natural vocality and how extreme its belief is. The more extreme their belief, the more they tend to post. [2] This is in line with evidence from social media networks and seems to be (at least partially) caused by humans that are very certain in their beliefs are inclined to avoid self-censoring Matthes et al., 2010. Each post includes a belief that is represented in it. This value is based on the agent who creates the post. An agent can be assumed to post content that is roughly in line with their current beliefs (e.g., Ross et al., 2019). The post content is modeled to be not the perfect representation of the agents beliefs because agents in order to mimic that agents are resharing content from others that just roughly agrees with their own view. Furthermore, agents cannot be assumed to always find the words that perfectly represent their current beliefs. As such, the post's belief stems from the agent's current belief combined with some additional noise. Furthermore, many of the policy levers are applied within this stage (downranking, deleting, strike system).

[1] Psychological effects such as primacy and recency effects according to which people better remember the first and last information they have received (Murdock Jr, 1962) is not included in the model. This choice has been made because it is not clear whether this is crucial for whether (or by how much) a person changes their mind based on the seen information, and because it would substantially increase model complexity.

[2] Furthermore, Disinformer agents tend to post more than NormalUsers.

### C.3.2. Belief Updating Stage

In this scheduling stage, all 'Normal User' agents are seeing some of the posts that they received, decide for each seen post on whether they judge the post as truthful, and if judged as truthful, the agents update their belief based on the post. For this updating the belief updating function is used with which the model was instantiated. The order in which the agents update their beliefs does not matter because this step is very modular and isolated. During it, agents cannot have any direct influence on other agents, nor be influenced by other agents.

**Rationale Behind Selecting Belief Update Functions**

To decide which belief update functions to include in the project, two aspects were taken into consideration. Firstly, some belief update functions should be those that are commonly used in previous work. Secondly, at least one belief update function should be one that is more deeply grounded in the science of other relevant fields such as social psychology and cognitive science. To strive for the first consideration, the two belief update functions that were the most commonly used in the papers that the literature review surfaced were selected. And to accomplish also the goal of the second consideration, the belief update function that was conceptualized and implemented in the previous internship was utilized. This belief update function is based on Social Impact Theory and adjusted for the context of social media. The essence of the three ways of modeling how people change their minds on social media is provided in Figure C.8.

**Overview Belief Update Functions**

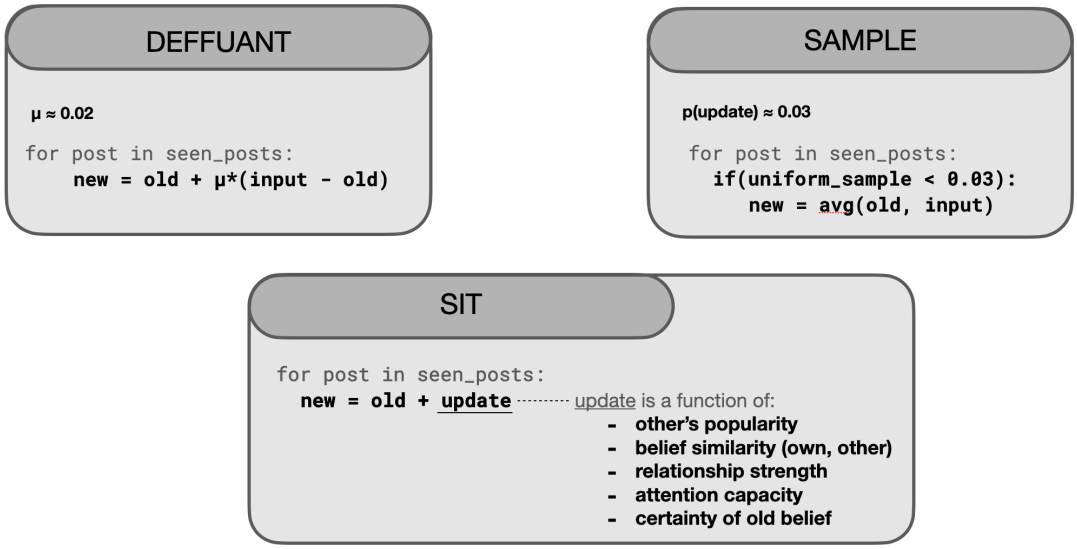



**Figure C.8:** Selected Belief Update Functions

In the DEFFUANT model, agents update every time they see a post. They always update by the same percentage of the difference between their old belief and the post belief. In the SAMPLE model, agents update only very rarely. But if they update, they update to the belief exaclty between their old belief and the post belief. In the SIT model (Social Impact Theory), the agents update every time they see a post. How much they update towards the post belief depends on the context. Various aspects about themselves, the agent who posted the post, and their relationship to that agent influence how strongly they update. The details about the implementation of the belief update functions can be found in the Details part of the ODD framework (in Section C.7).

### C.3.3. Policy Levers

The stages are affected by the policy levers. In order to get a full overview over the model, the specific policy levers and the rationale for including them are provided in this subsection. The details of how the levers are implemented can be found in the sections focusing on details (i.e., in Section C.5 and Section C.7).

**Levers Included in the Base Model and the Extended Model**

An overview over which levers are included in the two models is provided in Figure C.12. The base model was platform-agnostic. Consequently, the interventions that were selected to be included in the base model were not based on what interventions a specific social media platform is considering. Rather, this selection was based on which interventions governments are contemplating. Various possible interventions are included in documents like the European Commission's 'Strengthening the Code of Practice on Disinformation' (European Commission, 2021b). These include for media literacy interventions and ranking interventions. As such, simple versions of these two interventions were already implemented in the base model. In the extended model, however, the choice was made to focus on a use-case of a Twitter-like platform. This was done in order to facilitate grounding of the model. With the orientation on Twitter, interventions were not only selected if governments considered them, but also if Twitter had implemented them. Next to Ranking, these interventions additionally include flagging, removing content (i.e., deleting), and a strike system.

Flagging refers to the practice of attaching a visible label to a post in order to inform people who see the post that this post for instance has been fact-checked as being false. Flagging was not included because empirical evidence suggests that it is not very effective or even counter-productive at the goal of slowing the spread of these posts on the platform (Sanderson et al., 2021).

In Twitter's strike systems, accounts can receive strikes if they post Tweets that contain COVID-19 related misinformation (including about the safety of vaccines). For each number of accrued strikes, the account is locked for a specified amount of time. After five strikes, the account is permanently suspended(TwitterHelpCenter, 2021). The model orients itself in the mapping of number of strikes to duration of locked account.

In current literature, there is a discrepancy between interventions studied (mainly fact-checking) and the interventions that are implemented by real-world platforms (e.g., downranking, content moderation, and deplatforming accounts) (Courchesne et al., 2021). Including interventions that are implemented by real-world platforms is a step towards addressing this gap.

**Media Literacy Intervention**

**Rationale behind Media Literacy Intervention** The evidence from papers like Austin et al. (2021), indicates that higher media literacy seems to help people in recognizing false information on COVID-19 as such and in avoiding updating their beliefs based on that false information. However, previous work has commonly not included this effect of media literacy. In classical belief revision, agents update their beliefs based on every incoming information (Alchourrón et al., 1985). Many authors have criticized this and other approaches have been suggested (Hansson, 1999). For instance the 'credibility limited revision approach' presumes that an agent judges others as credible if the information coming from them does not contradict the agent's current belief (Booth et al., 2012; Hansson et al., 2001). This approach is somewhat related to Bounded Confidence models. In these models, agents only update if the other's belief is within a certain distance from the own belief Figure C.9. However, own set of problems. For instance, there is the unrealistic aspect of the arbitrary cutoff at the border of the confidence interval. Within the interval, agents update their beliefs strongly, and just outside of it, they do not update at all.

In contrast, in the MisinfoPy an agent's media literacy is included and affects whether they judge a given piece of information as truthful. This in turn, determines whether the agent updates their beliefs based on the piece of information.

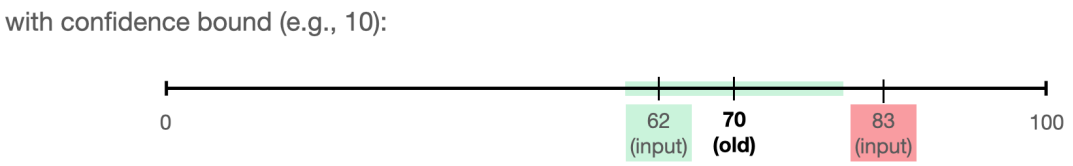



**Figure C.9:** Bounded Confidence Models

**Real-Life Counter-Part of Media Literacy Intervention** The lever of media literacy intervention represents an initiative that aims to increase netizen's media literacy. Such an initiative could for instance take the from of a workshop, an informative website, or a digital games. An ideal candidate to represent the media literacy intervention in this modeling effort would likely have the following attributes:

1. realistic that a government would opt for the intervention (to make the modeling work more realistic and more applicable to governments)
2. can be offered in a flexible manner to the users in the network (to be able to use the lever to different degrees)
3. available without a substantial initial investment and time delay (to make it more likely that the modeling work could have a timely impact)

The selected candidate is a serious game such as the one developed by DROG (DROG, 2021a, 2021b). The goal behind the game is to empower people in their media literacy. It seems to check all three boxes: (a) The Dutch Ministry of the Interior and Kingdom Relations supports DROG. Furthermore, other official partners by DROG include the UK Cabinet Office and the European Commission. (b) Their serious game is digital, playable in the browser, and easily shareable and accessible via a link. As such, their game is a very suitable use case for the media literacy intervention. (c) The game is currently already available. Consequently, the current implementations models the case of the Dutch government deciding to spread the digital game by DROG.

**Implementation of Media Literacy Intervention** Media literacy and its effect are modeled in a simplified way. Agents' media literacy can only be either 'LOW' or 'HIGH'. An agent with low media literacy is updating to any information that it receives. An agent with high media literacy has more success in judging the truthfulness of content but is not perfect in this skill. The rates at which these agents judge content as truthful are arbitrarily set at 80% for true content and at 20% for false content Figure C.10.

The game can be offered to a specified percentage of the overall population. If less than 100% of the agents will be offered the game, a way of selecting the agents that get the offer needs to be chosen. For this, agents are selected randomly. Selecting agents randomly is the simplest form. Implementing this in real life could for instance take the shape of purchasing ads for the game on the social media platform. [3]

When an agent participates in the intervention, its media literacy is set to HIGH. This means that in this simple implementation, agents whose media literacy is already HIGH cannot further benefit from

[3] For most efficiency of the intervention, one would want to offer the media literacy intervention to people of low media literacy. Because these people can benefit from the game most, being able to offer it to them first should achieve the greatest results. However, in the real world, one cannot just know the media literacy levels of a person[4]. However, if an agent has low media literacy, they are more likely to share content that has been fact-checked as false. As such, sharing such content is a promising indicator of low media literacy. Furthermore, also implementing such a distribution of the game in the real world is rather straightforwardly possible for the social media platform[5]. It could, however, be another idea to select agents in an ongoing basis during the run, based on the content that they are sharing. Agents that are sharing content which has been fact-checked as false could be selected to distribute the resources of the intervention mainly to those that can benefit most from it.

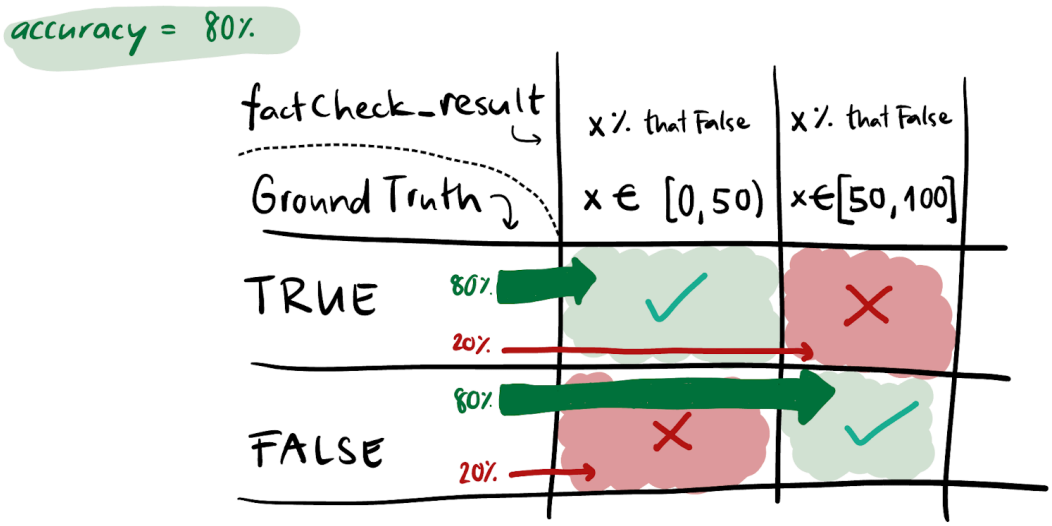



**Figure C.10:** Effect of HIGH and LOW Media Literacy

the intervention. This simple implementation was chosen for the sake of simplicity and computational feasibility. However, we want to keep in mind that in the real world, there are many degrees of media literacy and it could be the case that also people who already have a rather high media literacy can benefit from playing the serious game.

**Deleting**

Generally, if a Tweet is deleted, it cannot be seen by any user and thus cannot have an effect on other users. It should therefore be a rather effective intervention. Furthermore, Twitter is removing content from the platform. These two points together make a rather strong argument for including a deleting intervention.

The deleting intervention works via the deleting threshold lever called `delete_t`. If a post's probability of being true is lower than the threshold, then it may be deleted. However, it is to note that not all misinformation on social media platforms is detected and acted upon. On Twitter, 41% of content that was fact-checked as false stays on Twitter (Courchesne et al., 2021). Consequently, also in the model, there is only a 41% chance that posts with a ground-truth label of 'False' are actually detected and acted upon. This holds for interventions of downranking content, deleting content, and for the strike system. A higher value for `delete_t` makes a policy delete many more posts. In the extreme case of the threshold value being at the upper extreme (i.e, at 50.0), it means that even if there is only a 50:50 chance that the post is misinformation, it can be deleted.

A notable difference between how the deleting works in the model and in the real world is the following. In the real world, posts are usually not immediately detected and removed. Some users will likely see a post before it is removed. However, because these intricacies of time are not within the scope of this project, a simpler implementation was chosen; If there is enough certainty that a post is false, and if it would be detected as misinformation (based on the mentioned 41% detection rate), it is removed before it is even posted to any of the agents followers.

**Ranking**

Empirical evidence speaks for the power of ranking interventions (Perra & Rocha, 2019). This, combined with governments' interest in the intervention played a crucial role in the decision of whether to include a ranking intervention.

In the real world, every social media platform has their own trade-secret algorithms for ranking the content shown to the users. Additionally, these rankings are personalized for individual users. It is in

the interest of many to analyze these algorithms in order to adjust what they are posting to maximize the reach they can achieve and also academic research is done into this direction (e.g., Khan et al., 2021). However, these algorithms are adjusted frequently to fit the topics that users are currently interested in. Furthermore, platforms aim to be keep their proprietary ranking algorithms secret. In order to not get lost in the details, the MisinfoPy model aims to capture the high-level commonality that these ranking algorithms have with one another. The platforms generally have in common that they optimize for user engagement (Bromell, 2022). Particularly, the more posts are seen by the users, the more ads can be shown and the more ad-revenue the platforms receive. There is no single ingredient that makes for content that will cause the greatest user engagement. However, sensational content tends to include more extreme and confident post beliefs like sensationalist content and click-bait and attract more engagement (Watson, 2021). Consequently, to arrive at a simple representation of engagement, the extremeness or confidence of the post beliefs included in a post are taken to represent the expected user engagement. Thus, posts with post beliefs that are either closest to 0 or closest to 100 will have the highest visibility for agents, and those with post beliefs close to 50 will be seen by the least amount of agents.

With the ranking intervention, the visibility of content that has been detected as misinformation can be reduced by a specified percentage (`rank_punish`). However, the previously mentioned detect-and-act percentage of 41% applies here as well. Furthermore, downranking may only happen if there is enough certainty that the post is really misinformation. This threshold lever `rank_t` works the same way as the deleting threshold and the threshold for the strikes system. If a post's probability of being true is lower than the threshold, then it may be downranked if it is detected.

If that is the case, and the of percentage by which visibility is adjusted is 0% (i.e., `rank_punish` is 0.0), the ranking (i.e., visibility) is not adjusted. If the percentage is 50% (i.e., `rank_punish` is 50.0), the ranking (i.e., visibility) is decreased by 50%. At a ranking adjustment of 100% (i.e., `rank_punish` is 100.0), the post will not be seen by any agent who received it. This can also be called "shadow-banning" the post (Gadde and Beykpour, 2018; Jaidka et al., 2021) [6].

**Strike System**

The main reason for including the strike system is that Twitter implemented it already during the COVID-19 pandemic. Because the choice of the model's time resolution took the strike system into account, it was possible to implement a very similar mapping number of received strikes and account-lock duration (see Figure C.11). The system is straight-forward and does not require more explanation. It is to note that the previously mentioned detect-and-act percentage of 41% applies here as well.

## C.4. Design Concepts

In this section, we go one by one through the eleven common design concepts of ABMs.

### C.4.1. Basic Principles

The MisinfoPy model tackles the issue of how opinions of people influence one another. This is generally called opinion dynamics and there has been done a substantial amount of previous work on this. What sets the base model apart from previous literature is its novel interpretation of the classic Social Impact Theory by (Latané, 1981). This theory has been adjusted to the social media environment. (As such, the contribution of the base model focuses on the agent-level, while it falls in line with previous literature on the system level.) Within this project, the base model has been extended both on the system level as well as on the agent-level. For an overview, see Figure C.12. On the system level, the extended model (referred to simply as MisinfoPy) adds to existing literature by including multiple uncertainties and policy levers. This enabled the exploration of how potential policies (i.e.,

[6] Shadow-banning a *post* is vastly different from shadow-banning an *agent*. If an agent would be shadow-banned, all of their posts would be completely hidden from other agents (or almost completely hidden, depending on the implementation).

| Nr. Strikes | Steps blocked |
|---|---|
| 0 | 0 |
| 1 | 0 |
| 2 | abort current step |
| 3 | abort current step |
| 4 | 7 |
| >=5 | ∞ |

**Figure C.11:** Strike System

lever combinations) might perform in certain scenarios (i.e., combinations of uncertainty instantiations). Furthermore, the extended model includes various metrics for the evaluation of the policy candidates. While this approach is common within the field of Decision-Making under Deep Uncertainty (DMDU) and Robust Decision-Making (RDM), to the best of my knowledge it has not previously been used for the issue at hand. Previous work rarely evaluated policies (Gausen et al., 2021), and if it did, only one metric and very few policies were included. On the agent level, the model contributes by enabling the use of different ways of belief updating. Next to the belief update based on social impact theory, the extended model also includes the commonly used Deffuant-based (as in e.g., Al Atiqi et al., 2018; Beskow and Carley, 2019; Du et al., 2021; Mason et al., 2020; Rajabi et al., 2020) and sampling-based (as in e.g., Gausen et al., 2021; Sari et al., 2021; C. Wang et al., 2019) belief update functions.

### C.4.2. Emergence

The emergent output of the model is how the resulting belief distributions over the whole agent population, as well as how the differences in model inputs map to model outputs in terms of the defined metrics.

### C.4.3. Adaptation

Below, the adaptive agent behaviors are explained on a high level. The third adaptive behavior differs between the models. For each model, a separate table is provided below.

### C.4.4. Objectives

Agents implicitly have the goal of arriving at the most accurate beliefs. Thus, e.g., only update their beliefs if they judge a post as truthful, and have the heuristic that agents to which they have a stronger relationship are likely more trustworthy and thus posts from such agents warrant a stronger belief update. However, this model does not include explicit direct objective seeking. Agents do not calculate one variable which they aim to maximize or minimize. As such, there are no objectives to describe here (Supplementary file S1 to Grimm et al., 2020).

| | Base | Extended |
|---|---|---|
| **X** | - agent ratio<br>- network density | - agent ratio<br>- network density<br>**- vocality of both agent types**<br>- belief update function based:<br>**- DEFFUANT: probability of update**<br>**- SAMPLE: update size**<br>**- SIT: number of posts considered to estimate belief similarity to other agent** |
| **L** | - interventions<br>- media literacy (simple)<br>- ranking (simple) | - interventions<br>- media literacy (extended)<br>- ranking (extended)<br>**- deleting**<br>**- strike system** |
| **R** | - Time<br>- Space<br>- Agents<br>- Posts<br>- Belief updating<br>- SIT | - Time<br>- Space<br>- Agents<br>- Posts<br>- Belief updating:<br>**- DEFFUANT**<br>**- SAMPLE**<br>- SIT |
| **M** | - Belief-based<br>(visual inspection of resulting belief distribution over agents) | **- Belief-based**<br>**(number of agents above belief threshold)**<br>**- Polarization (variance)**<br>**- Engagement**<br>**- Free speech constraint**<br>**- Average user effort** |

**Figure C.12:** Overview Base Model and Extended Model

The graphic provides an overview how the base model has been extended. The extensions are highlighted in light-blue and bold.

**Table C.5:** Adaptive Behavior 1

| **Adaptive Behavior 1** | |
|---|---|
| What decision is made: | - decide how many posts to post in the current model step |
| Alternatives to choose from: | - any integer between 0 and the maximum reachable by sampling from the normal distribution defined by the agent's vocality and current belief |
| Decision drivers: | - vocality<br>- current belief |
| Modelled as direct or indirect objective-seeking: | - indirect objective-seeking<br>(Agents do not explicitly rank different options and decide which one seems better to reach their objective of informing others and spreading their beliefs. Agents simply the sample number of posts based on vocality and current belief.) |

**Table C.6:** Adaptive Behavior 2

| **Adaptive Behavior 2** | |
|---|---|
| What decision is made: | - decide whether or not to trust a post<br>(and thus whether to update beliefs based on it) |
| Alternatives to choose from: | - any integer between 0 and the maximum reachable by sampling from the normal distribution defined by the agent's vocality and current belief |
| Decision drivers: | - media literacy level (of the agent)<br>- ground truth (of the post) |
| Modelled as direct or indirect objective-seeking: | - indirect objective-seeking<br>(Agents do not explicitly rank the two options and decide which one seems better to reach their objective of arriving at the most accurate belief. Agents simply rely on how their media literacy skills.) |

### C.4.5. Learning

No learning mechanism has been included in the model. The reason for this choice is that the agents do not get feedback on the quality of their past decisions (wrt adaptive behaviors). Consequently, they do not have information that could enable a learning process. If the scope of the model would be

**Table C.7:** Adaptive Behavior 3

| Adaptive Behavior 3 (DEFFUANT) | |
|---|---|
| What decision is made: | - decide how to update beliefs |
| Alternatives to choose from: | - the agent adjusts its belief by a fix percentage towards the post's belief |
| Decision drivers: | - `deffuant_mu` (a fix percentage of around 2%) |
| Modelled as direct or indirect objective-seeking: | - indirect objective-seeking<br>(Agents do not explicitly rank the two options and decide which one seems better to reach their objective of arriving at the most accurate belief.) |

| Adaptive Behavior 3 (SAMPLE) | |
|---|---|
| What decision is made: | - decide how to update beliefs |
| Alternatives to choose from: | - 0 (i.e., keep current belief)<br>- 50% towards the post's belief |
| Decision drivers: | - `sampling_p_update` (a fix probability of around 3%) |
| Modelled as direct or indirect objective-seeking: | - indirect objective-seeking<br>(Agents do not explicitly rank the two options and decide which one seems better to reach their objective of arriving at the most accurate belief.) |

| Adaptive Behavior 3 (SIT) | |
|---|---|
| What decision is made: | - decide how to update beliefs |
| Alternatives to choose from: | - the agent can adjust its belief by any value between the minimal and the maximal update:<br>- minimal update: 0 (i.e., keep current belief)<br>- maximal update: the post's belief |
| Decision drivers: | - popularity of the other agent (i.e., the source of the post)<br>- belief similarity (between itself and the other agent)<br>- relationship strength (between itself and the other agent)<br>- attention capacity (of the agent)<br>- current belief |
| Modelled as direct or indirect objective-seeking: | - indirect objective-seeking<br>(Agents do not explicitly rank the two options and decide which one seems better to reach their objective of arriving at the most accurate belief.) |

larger and also include consequences of the agents' decisions (for instance whether they or people in their environment that hold similar beliefs had negative health outcomes), a learning process could be included. I decided against this kind of extension because it would increase the scope of the model beyond feasibility of a master project

### C.4.6. Prediction

There are a couple of implicit predictions that the agents make in the adaptive behaviors. When deciding how much to post (Adaptive Behavior 1), it includes the implicit prediction "If you are more confident in your belief (i.e., have a belief value that is further from 50), it is more valuable to communicate your belief to others.". In the context of judging the truthfulness of a post, it includes the implicit prediction that "Updating your belief to posts that you judge as truthful will improve the quality of your belief.". And when deciding how to update the belief, which predictions are implicitly included depends on the model. For each of the models, the statement "To arrive at the most accurate belief, it is best to ..." would be differently completed. In the DEFFUANT model, this assumption would likely be something along the following lines: "To arrive at the most accurate belief, it is best to **update a little bit based on everything that you judge as truthful.**" For the SAMPLE model, it could be something like: "To arrive at the most accurate belief, it is best to **only update very rarely (e.g., when you are very sure that they are truthful), but then update strongly.**" And for the SIT model: "To arrive at the most accurate belief, it is best to **take various factors about yourself, the post, and the agent that posted the post into account.**" This itself includes the following implicit predictions:

- If an agent is more popular (i.e., has more followers), that is likely because they have earned these followers by providing valuable, high-quality information. Thus, it is useful to strongly update beliefs based on their posts.
- If an agent has a similar belief to my own, we likely have similar values and a similar social group. Thus, they probably just have some more information. If I would also have that information, I

would also update towards their position. Consequently, it is useful to strongly update beliefs based on their posts.

- If we have a strong relationship, they care about me and are not trying to mislead me but to help me. Thus, the stronger our relationship (i.e., `tie_strength`) is, the more I can trust their information and update my beliefs based on it.
- There is no implicit prediction *by the agent* behind the 'attention capacity' factor. But an prediction stemming from the social impact theory: If an agent is following many agents, the influence of each those agents over the agent is smaller.
- If one has strong confidence in a belief, there are good reasons for that. Consequently, it should take more evidence and input in order to change the belief.

As humans usually do not explicitly reason through whether they want to update their beliefs after having seen a post, the implicit representation seems to represent the real world situation better than an explicit one. Consequently, the implicit representation was chosen.

### C.4.7. Sensing

For the agents' behaviors it is crucial what they "know" – what information they have access to, and with what accuracy or uncertainty they have access to it. This applies to state variables of the agent itself, but also about state variables of other entities (e.g., other agents or the global environment). In the following, the choices around such sensing, as well as the rationale behind them, are explained. Figure C.13 provides an overview about what state variables of the agent itself and of other agents can be accessed. With respect to the global environment, there is not much to say because the agents do not need to be able to explicitly sense the time indication of model steps.

**Own States**

The agents are assumed to be able to use their own state variables. There is no uncertainty in what their values are. Humans would likely struggle to pinpoint their belief to an exact value on a scale from 0 to 100. Also in the model, agents do not need to have such conscious accurate access. The processes of deciding how many posts to post, whether to judge a post as truthful, and how much to update to a post that was judged as truthful are assumed to not happen within the explicit thinking of a human (i.e., not within system 2 in Kahneman's distinction (Daniel, 2017)), but they are assumed to happen in a more automatic fashion into which the human doesn't have easy and explicit access (i.e., system 1 with respect to Kahneman's distinction (Daniel, 2017)). As such, the values are "accessible" to the agent in the sense that these system 1 based processes can play out accurately. The agent however is assumed to not e.g., be able to communicate their exact belief value to others. In line with that, within the creation of new posts, agents do not present their exact belief value but aim to capture this value, leading to a represented belief that is somewhat off from the agent's exact belief. Furthermore, the number of strikes the agent has accumulated as well as the time until when the agent is blocked from posting (if at all) is accessible to the agent itself, but it is not used by the agent. Rather, the model uses and does not give the agent a chance to post if it is still blocked. There is no uncertainty for the global model environment in sensing how many strikes an agent has and whether it is still blocked.

**Others' States**

The central state variable in this model as well as in most related models are the agents' beliefs. In previous literature, it is common that agents can directly perceive the beliefs of other agents (e.g., Du et al., 2021; Sano et al., 2021; C. Wang et al., 2019). This however is drastically different in the real world. In this model, agents can sense only a few state variables of other agents. Importantly, as in the real world, agents cannot directly perceive the beliefs of other agents. Instead, agents can see what other agents have posted previously and aim to use that to estimate their beliefs. This again is in line

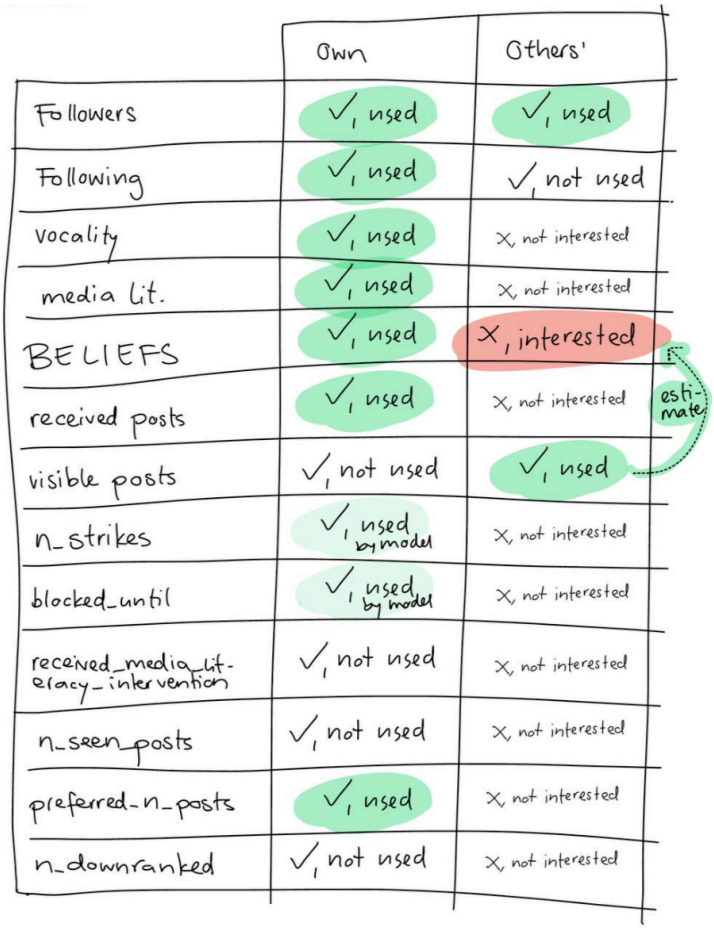



**Figure C.13:** What Agents Can Sense

The graphic shows what an agent can sense about itself and about other agents.

with the real-world situation on social media. There will always be uncertainty on what the real belief of the other agent is because the posts represent the agent's belief imperfectly and because only the last few posts are considered for the estimation. Other than this, agents have access to others' number of followers and following. The remaining state variables are kept private because there is no reason why agents would (need to) be able to access the remaining state variables of others.

### C.4.8. Interaction

The interactions between agents are modeled in a straight-forward and somewhat simplified way. Agents can only directly interact with other agents that they are directly connected to. Specifically, they can only share posts to agents that follow them, and they can only see posts of agents that they follow. This is somewhat simplified from the real-world social media networks where it is common that also posts from accounts that one does not follow are displayed. However, to keep the model rather simple and to steer clear of modeling proprietary recommendation algorithms. As such, there are only direct interactions possible. The only interaction that might be described as mediated (rather than direct) is the following: Imagine an agent A being followed by agent B, who in turn is followed by agent C. Now, agent A can directly interact with agent B, leading to a change in B's beliefs. Agent B's beliefs in turn can directly interact with agent C and lead to a change in C's beliefs. Like this, one could argue that agent A has indirectly interacted with agent C – in other words, agent A had mediated interaction with agent C. With whom an agent can interact is determined during the creation process of the network

(and remains static over the run of a simulation, as described in Section C.2).

### C.4.9. Stochasticity

An overview of where stochastics play a role is depicted in Figure C.14.

| **Where stochasticity is used:** |
| --- |
| **Initialization** (see appendix)<br>- building network<br>- agent characteristics (type, init belief, media literacy, ...)<br>- where to attach agent (preferential attachment)<br>- how strong relation to other connected agents (tie_strength)<br><br>- selecting agents for media literacy intervention |
| **Posting Stage**<br>– creating new posts<br>- number of posts<br>- post's belief and ground truth<br><br>– sharing new posts<br>- whether a post is detected as misinfo |
| **Belief Updating Stage**<br>- whether agent sees a received post<br>- judging truthfulness of posts<br>- specific to Belief Update Function:<br>- DEFFUANT: no stochasticity<br>- SAMPLE: sample whether update or not<br>- SIT: no stochasticity |

**Figure C.14:** Stochastics of the Model

#### Initialization

During the initialization of the model, there are a few processes that include stochastics. Most of them are part of the network creation. During network creation, the agent characteristics (like agent type, initial belief, and media literacy level) are sampled. Furthermore, stochasticity involved in determining to which other agents an agent is connected to (including how strong that connection is). Stochasticity is used here to achieve variability of the initial conditions. In this manner, characteristics of real social media networks can be replicated in many different concrete networks. Thus, avoiding that the structure of any one specific network has an overly strong influence and may lead to spurious results.

Apart from the network creation process, there is one more procedure during initialization that uses stochastics – the selection of agents for the media literacy intervention. Here, the reason for using stochastics is to keep this process simple. By selecting the agents this way, the model results are more independent of what the real-life choices behind such an intervention might be (e.g., how the organizers would aim to distribute the intervention and how that might correlate with the specific kind of intervention they choose). As previously discussed, in this simplified model, we assume a digital game as the intervention. A random distribution of the intervention here could take the form of digital ads to distribute the intervention independently of the network structure itself.

#### Posting Stage

Each time the model enters the posting stage, the number of posts that each agent wishes to post, as well as the characteristics of that post (e.g., post belief and its ground truth) are determined via making use of stochastics. This implementation was chosen in order to not model in detail how people decide on how many posts they want to post, or how successful they are in posting posts that are close to

their current beliefs. Modeling these processes explicitly would have included far too much detail for the purposes of this project.

An other attribute of the post is whether it is detected as misinformation. Setting this attribute involves stochastics for analogous reasons. It would go far beyond the scope of this project to model explicitly how this detection system works. In the real world, this would include a combination of automatic detection systems, reports from users, and human detection and decisions. Consequently, this aspect was modeled with lower resolution by just making sure that it would replicate the real world rate of misinformation that is not being acted upon.

**Belief Updating Stage**

In the stage where agents update their beliefs, there are some stochastics involved that apply to all three models. These include the modeling of whether an agent sees a received post and whether an agent judges a given post as truthful. Whether an agent sees a post is dependent on the visibility of the post. It could have been an option to model this more explicitly by for instance modeling explicitly how much time the agents spend online. However, for the research question and the scope of this project, this was deemed unnecessary and would have mainly meant to include another set of uncertainties. Modeling the judging whether a post is truthful a stochastic process was also done with the goal of simplifying this submodel while recreating agent behavior at a rate that might be seen in the real world.

Stochastics that are specific to the individual models and their belief update functions are few. The DEFFUANT model and the SIT model both do not include stochasticity. The SAMPLE model includes the sampling of whether an agent will update its beliefs or not. This was presumably also done to simplify the modeling of this process and lead to outcomes that are hopefully representative of the real world.

**Observation and Analysis**

The final stage of observation and analysis does not include relevant stochastic procedures.

### C.4.10. Collectives

No collectives have been modeled in the model as was not at the focus of this project and thus has not been explored. However, if one would be interested in that, one might explore whether something resembling filter-bubbles or echo-chambers might have emerged from the agent behaviors.

### C.4.11. Observation

**Goal of observation: being able to evaluate policies**

This section focuses on what we observe from the model, when we gather that data, and how it is processed. The main information we want to get out of the computational experiments is information that lets us judge how well a policy has performed in the context of the current scenario. In the context of the previously discussed complex socio-technical system with various stakeholders, some of the main stakeholders have been identified. Subsequently, for each selected stakeholder, their main objectives in the context of handling misinformation on social media have been deduced. An overview of the selected stakeholders and their metrics is visible in Figure C.15. An attempt has been made at capturing the essence of these objectives in a quantifiable way that is within the model's scope. Naturally, these operationalizations are somewhat crude and imperfect. Operationalizing values accurately is notoriously hard. However, I hope that the implementations of the values point roughly into the right direction. I currently believe the implementations are sufficiently close to justify using them to evaluate the policies and have a decent discussion around the findings.

**Overview**

All metrics are observed at the end of the run. To enable that the model may be used for optimization, they are all measures of a central tendency (Optimization needs one single value to evaluate a policy).

| Stakeholder | Main Objectives |
|---|---|
| **government** | - number of agents above a belief threshold (`n_agents_above_belief_threshold`)<br><br>- polarization (`polarization_variance`) |
| **social media platforms** | - engagement (`engagement`) |
| **users** | - free speech (`free_speech_constraint`)<br><br>- effort for the user (`avg_user_effort`) |

**Figure C.15:** Stakeholders and Objectives

The first two metrics are more complex, summarizing over the whole distribution of agents' beliefs (`n_agents_above_belief_threshold` and `polarization_variance`). The remaining three metrics are fairly straightforward in simply summing the aspect of interest over the whole run of the simulation and averaging them either over the number of agents (`engagement` and `avg_user_effort`) or the number of total posts (`free_speech_constraint`). In the following, each of the five metrics is described and elaborated on.

`n_agents_above_belief_threshold`

With the assumption of BDI that people's beliefs have an influence on what actions they take, it is straight-forward to reason that having "false" beliefs can have unwanted consequences. "False" beliefs refers here to beliefs that are at odds with how the world actually works. In this project, the focus lies on a statement for which there is assumed to be a "scientifically settled answer". Thus, conditional on these assumptions, a person's belief about whether this specific statement is true or false should influence their decision of whether they want to get vaccinated or not. Naturally, there are a lot more factors included of whether a person actually gets vaccinated. For instance, even if a person would strongly believe the statement is true, they might still not get vaccinated because of a lack of accessibility. But even if believing that it is the safer choice to get vaccinated is not a sufficient condition to get vaccinated, it is probably a necessary one. Empirical research has also confirmed that safety concerns are a relevant issue influencing COVID-19 vaccination willingness (Neumann-Böhme et al., 2020). Consequently, it makes sense to use the belief regarding the VAX statement as an indication for vaccination willingness.

However, where the cutoff point might lie between being certain enough about the benefit of the vaccine to actually want to get the jab is unclear. Furthermore, such a threshold is difficult the verify. While it could be an idea to at least roughly calibrate this threshold using the number of people getting vaccinated, this seems very difficult to do in a scientifically sound manner. One reason are the other reasons that can hinder people from getting vaccinated even if they think it would be the safer choice. Another reason is that we do not have access to the real-world "belief distribution" of the people that got vaccinated and those that did not. Consequently, the chosen approach was to acknowledge this uncertainty and make the threshold an uncertainty that can be explored.

`polarization_variance`

Polarization of beliefs roughly refers to how divided or separated a group of people is (Bauer, 2019). How to measure polarization is debated and could be said to be polarized itself (Bramson et al., 2017; Duclos et al., 2004; J. Esteban & Ray, 2012; J.-M. Esteban & Ray, 1994; Lelkes, 2016). Instead of aim-

ing to find the perfect measure for polarization, it can be a good choice to implement multiple measures and learn from the combination of these measures (e.g., (DiMaggio et al., 1996; Fiorina, Abrams, et al., 2008; Lelkes, 2016)). Which measures are suitable depends for instance on the dimensionality of the data and on its scale type (Bauer, 2019). In this project, we focus on data of only one dimension (only the VAX belief). The data is continuous, but could reasonably be discretized (e.g., into integers). For unidimensional data that is either continuous or discrete, measures such as Variance, Kurtosis, DER Index, Average Synthetic Opinion Score, d-Squared and ER Index could be considered (Bauer, 2019). The most promising candidate out of these was variance. With variance, unimodal distributions result in low variance values (i.e., low polarization). Furthermore, the variance value of mirrored distributions is identical. Last but not least, it is a very established measure in previous literature. The main concern about it is that if one is asked to picture a distribution with high polarization, it would likely be a strongly bipolar one and not a very flat distribution. The variance measure however judges such flat distributions (i.e., platykurtic distributions) as very high in polarization. As such, variance does not perfectly capture our intuitions about polarization. What could cover these cases well would be to implement a Kullback-Leibler Divergence (KL-divergence) measure. KL-divergence measures the difference between two (probability) distributions and thus could be implemented to calculate the difference between a belief distribution and our prototype of a polarized distribution. The classical KL-divergence is not symmetric depending on whether we measure the how far distribution A is from distribution B as when calculated from distribution B to distribution A. But by calculating both directions and taking the average, the KL-distance is a symmetric measure. Because of these benefits, the symmetric KL-divergence has also been implemented. However, when we are evaluating policies based on multiple objectives, it is valuable if these metrics are as independent from one another as possible. One option could be to combine the variance and symmetric KL-divergence into a new and potentially better measure of polarization. However, as the attempt at improving of polarization metrics would be a whole project in itself and in order to keep things simpler where possible, I decided to opt for the tried-and-tested option of variance.

`engagement`

The business model of social media platforms is to sell ads. As long as this business model does not drastically change, it is a key motivation of the platforms to increase engagement, to increase the number and duration of ”eye-balls” on their services because that increases their ad revenue and their profit. The more posts a user sees, the more ads can be shown to that user. There are definitely also other factors also playing into the ad revenue. For instance, users are more predictable in their beliefs, they are also more predictable in their actions. This makes it easier to show them ads that they will act upon, increasing the value for ad buyers and by extension for the social media platforms themselves. The actions of people with more confident or more extreme beliefs likely will have stronger preferences on what kind of content they prefer. As such, they are to predict, leading to higher engagement (Carroll et al., 2022). Thus, one could make the case to include the belief values into the calculation of the engagement metric. However, this would complicate things more and introduce more uncertainties. Consequently, the choice was made to take number of posts seen as a metric that is simple, straightforward to measure and yet likely capturing the essence of this objective rather well.

To calculate the engagement metric, the number of posts that each agent has seen (over the whole model run) are added up to arrive at the total number of posts seen by the entire network. In order to make the interpretation of this metric more intuitive and independent of the network size, the total number of posts seen is divided by the number of agents. Consequently, engagement is the average number of posts seen by an agent over the course of the whole model run.

`avg_user_effort`

A measure of how much effort a policy implies for the users is important. If such a measure is not included, it risks over-burdening the user. One result of that could be that the users are dissatisfied.

However, another consequence could a false sense of security. If the policy is causing too much effort for the users, they will likely try to avoid as much as possible of that effort. For instance, if people have high media literacy knowledge, but they act more as if they would have low media literacy levels because that is less effort, that could decrease or even negate the impact of the policy while decision-makers might wrongly assume that the grand challenge of misinformation on social media is mitigated because so many people have been empowered by a media literacy intervention.

The metric of average user effort is influenced by the media literacy intervention. Effort is counted as *time required*. The other interventions cause no effort for the users as they do not require any extra action or time from the users. However, there are indirect effects from the other levers as well. For instance, if many posts are deleted, downranked, or blocked from being posted by the strike system, it means that there are fewer posts to see. Consequently, also the overall effort for judging whether posts are truthful decreases.

For the media literacy intervention, there is an initial amount of time required for the actual media literacy intervention. This could for instance be a computer game like the one developed by DROG with the support of the Dutch government (DROG, 2021a) or a workshop. Furthermore, the effort for users is increased with each post that they see and want to judge the truthfulness of. The initial investment is currently set to one hour and the time per post is on average 3 seconds for agents with low media literacy and 30 seconds for agents with high media literacy. Grounding of the amounts of the initial investment as well as the investment per post is difficult. I have not succeeded at finding data on how long people take to judge truthfulness of a post or anything related. However, where the actions of people with higher or lower media literacy differ is what steps they take to judge the truthfulness of a post. The behaviors that are recommended for higher media literacy are rather costly in their required time investment (see e.g., ("Fake news, propaganda, and disinformation: Learning to critically evaluate media sources: Infographic: Spot fake news", 2022; Public Library, n.d.)). For media literacy levels to really be higher, some of those recommendations are followed, causing people more effort. In line with this, people with low media literacy levels spend less time on judging trustworthiness (Spilsbury, 2018). While there is support for the qualitative difference, it is difficult to quantify the invested amount of time differs by media literacy level of the person. The lack of scientific data makes the grounding of the size of these investments difficult. Fortunately, the estimates of these times do not need to be very precise. This is because we only care about the *relative* differences between the policies. The goal here is not to find out whether people would actually go spend these amounts of times. It is to rank policies while taking into consideration how they differ in terms of user effort. The values that were chosen are on average 30 seconds per post for an agent with high media literacy and 3 seconds for an agent with low media literacy. For a rough check of how sensible these values are, see the example provided in Figure C.16.

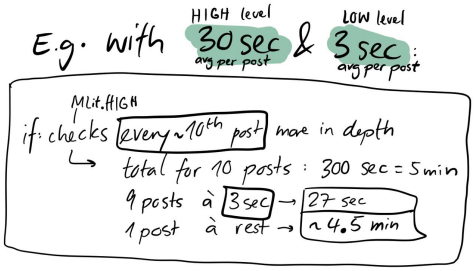



**Figure C.16:** Average User Effort: Example

Time that agents of differing media literacy levels require to judge truthfulness of posts. This is just to see that the numbers seem rough okay because there is no need for precision. We are only interested in the relative differences between policies, not in the absolute values.

With ease of interpretation in mind, the effort is provided in minutes rather than seconds. Furthermore, like for the engagement metric, also this metric was not only aggregated over all agents, but then also divided by the number of agents. The makes the interpretation of this metric more intuitive and independent of the network size, the total number of posts seen is divided by the number of agents.

`free_speech_constraint`
Restricting the freedom of speech can have many first order and higher order effects in a society. In the simplistic case of the VAX belief, this might seem less relevant because posts that are deleted are false and harmful to the individual and to the group. However, historically, censors have thought of themselves as doing a good, even heroic thing by protecting the vulnerable. Even if we would likely judge today that at least some cases of such censorship have caused more harm than good. As such, we want to be careful and indicate that, all else being equal, policies should be judged as better if they restrict free speech less.

This metric represents the portion of posts that were ”deleted” compared to the number of posts that the agents wanted to post. If a post is blocked because the agent is currently still blocked due having received too many strikes, each such blocked post is counted as 1 ”deleted” post. If a post was downranked by 20%, it counts as 0.2 posts ”deleted”. To arrive at the metric value, the amount of ”deleted” posts is calculated and divided by the total number of posts that agents wanted to post.

## C.5. Initialization

The initialization of the model includes a couple of processes:

1. **Build the network**: a Barabási-Albert graph, after (Barabási & Albert, 1999). For instance using the NetworkX library, which is a Python-based package focusing on networks (Hagberg et al., 2008). It was chosen to create the networks within this project because its powerful functionality also includes the creation of random graphs of the *Barabási-Albert* type. The number of nodes is determined by the parameter `n_agents`. It is to note that this network is at first still without agents. The network density is determined by the uncertainty of `n_edges`. This parameter determines with how many edges a node is attached to already existing nodes (during the process of building the network).
2. **Create agents and add each of them to a node.** Agents are created including all their (initial) attributes, such as their initial beliefs. Regarding the initial belief distribution: Accurately estimating the belief distribution of a large group of humans is currently unworkable. One way of grounding the initial distribution is to use sentiment analysis on the posts of users in real-world networks. For Twitter, for instance, these analyses use the Twitter API (Application Programming Interface) and a Python library like TweePy to access the API. This setup offers the functionality of gathering tweets to a specific topic which is specified using a number of search terms. A number of characteristics of the tweets are collected. The underlying sentiment of each tweet can then be analysed using different machine learning methods. This classification typically distinguishes only three classes – in this case either pro vaccines, neutral, or against vaccines. Consequently, the resolution is a lot lower than what is represented in the model. The difference in resolution between the sentiment analysis and the representation in the model represents a point of future improvement with respect to model validation. The current state of validation still relies on the assumption that a generalization between the two resolutions can be made. With reliable sentiment analysis data that has a higher resolution, the validation of the belief initialization could be improved. Nevertheless, the results from such real world data and sentiment analysis can provide some grounding to the belief initialization. According to the Twitter data gathered by Yousefinaghani et al. (2021), the sentiment regarding the COVID-19 vaccine was quite uniformly distributed at beginning. Based on this data, the beliefs of the population of Normal Users are

initialized by sampling values from a uniform distribution between 0 and 100.

3. **Apply media literacy intervention** (select agents randomly, upgrade their media literacy to HIGH)

## C.6. Input Data

In this model, no input data was used to represent time-varying processes.

## C.7. Submodels

### C.7.1. Posting Stage

In Posting Stage a couple of actions happen:
(All the following points happen in each model step once for each agent)

1. Sample how many posts the agent would prefer to post. (This is sampled from a normal distribution that is defined by the agent's vocality parameters and its current belief.)
2. Check whether the agent is blocked.
    - If it is blocked, do not post any posts, but just register for the free speech metric that the number of preferred posts were ”deleted”.
    - If it is not blocked, do for each post the following steps:
        - Create the post (incl. all its attributes)
        - Apply the deleting and the strike system intervention in case they apply (i.e., in case the post was detected as misinfo with a certainty that is beyond the corresponding threshold).
3. Share all the created posts to the agent's followers and save them into list of posts that are visible on the agent's own profile.

### C.7.2. Belief Updating Stage

In the Belief Updating Stage, each Normal User performs the following steps:

1. Sample which of the received posts are seen.
2. For each of the seen posts, do these actions:
    - Judge the truthfulness of the post. (This depends on the agent's media literacy level and on the ground truth of the post.)
    - If the post was judged as truthful, update the belief using the belief update function of the model.
        - DEFFUANT: as in Figure C.8
        - SAMPLE: as in Figure C.8
        - SIT: as explained in Reddel (2021)