# Research Proposal for the Go No-Go Meeting

Emma den Brok

June 22, 2020

## Contents

**1 Introduction** **2**
1.1 Online disinformation as a problem of growing societal concern . 2
1.2 What is disinformation? . . . . . 2
1.3 Scope and goals of this proposal . . . . . 2

**2 Literature Review** **2**
2.1 Disinformation Research . . . . . 2
2.1.1 Computer Science Perspective - What does disinformation look like? . . . . . 3
2.1.2 Psychological perspective - Why does disinformation work? 3
2.1.3 Societal perspective - What are the effects of disinformation? 4
2.2 Network Science . . . . . 5
2.2.1 Why is network science relevant to disinformation? . . . . 5
2.2.2 Information Diffusion, (Complex) Contagion and Network Strucutres . . . . . 6
2.2.3 Multiplex networks . . . . . 7
2.3 Psychology of advertising . . . . . 8
2.4 Propaganda . . . . . 10
2.5 Simulation modelling . . . . . 10

**3 Research Gap** **11**

**4 Research Approach** **11**
4.1 Disinformation is a post-normal problem . . . . . 11
4.2 A Metamodel to enable theoretical exploration . . . . . 12
4.3 Research Goal & Questions . . . . . 13

**5 Methodology** **13**
5.1 Defining the purpose and requirements of the metamodel . . . . 14
5.2 First Design Cycle . . . . . 14
5.3 Simulating Case Studies . . . . . 15
5.4 Evaluation of the metamodel . . . . . 16

# 1 Introduction

## 1.1 Online disinformation as a problem of growing societal concern

## 1.2 What is disinformation?

## 1.3 Scope and goals of this proposal

socio-technical systems perspective.
does it matter, for research, if it is disinfo or just (hidden) propaganda

# 2 Literature Review

## 2.1 Disinformation Research

Disinformation can be defined as false information created with the intent to mislead (Fallis, 2015). Commonly associated terms in literature are misinformation, conspiracy theories, fake news, and hoaxes. Actors involved in the spread of disinformation range from bots and trolls to partisan media platforms and (national) government bodies (Tucker et al., 2018). The goals of those involved in the spread of disinformation are also varied, ranging from political influence (Keller, Schoch, Stier, & Yang, 2020), distracting the public (King, Pan, & Roberts, 2017), to monetary gain and fun, amongst others (Hrčková et al., 2019).
Online disinformation campaigns use the digital infrastructure provided by third party social media platforms to spread or advertise its content. As a result, these social media platforms have a mediating role between disinformation campaigns and their target(s) (Bessi et al., 2015). As a result, (mainly) US tech companies hold a lot of power in this domain. Additionally, messaging boards like reddit or 4chan are used as testing or breeding grounds of disinformation campaigns (Morgan, 2018).

In this section I provide a brief overview of disinformation research based on three perspectives:

- **Computer Science:** The focus is on the detection of properties of disinformation, either of content itself (feature detection) or particular patterns of its engagement (i.e. likes, lifespan of interaction, spread over a network). This perspective asks *"What does it look like?"*
- **Psychology:** Aims at understanding the psychological traits or mechanism that make individuals vulnerable to disinformation. This perspective asks *"Why does it work?"*
- **Societal:** Discusses the effects disinformation has on a societal level, i.e. on politics, polarization, and behaviour. This perspective asks *"What does it do?"*

### 2.1.1 Computer Science Perspective - What does disinformation look like?

The computer science perspective is motivated by the underlying assumption that if disinformation can be detected automatically it can be removed, surpressed or labeled before it reaches a large audience. One set of studies tries to identify features that allow them to predict whether a piece of content is false (or a rumour) or not. Such features may be content-based, network-based or platform specific (Qazvinian, Rosengren, Radev, & Mei, 2011), or based on the user who posted the content (Liu, Nourbakhsh, Li, Fang, & Shah, 2015). Machine-learning approaches are also used to train algorithms on large annotated datasets (i.e. Wang (2017), Mitra and Gilbert (2015) Papadopoulou, Zampoglou, Papadopoulos, and Kompatsiaris (2019)) yet these approaches are labour-intensive and subsequent algorithms may be easily fooled and perform well only on the datasets they were trained on (Gröndahl, Pajola, Juuti, Conti, & Asokan, 2018).

When studying how users engage with misinformation, Zollo and Quattrociocchi (2018) claim that the spread of misinformation is driven by confirmation bias and enabled by echo chambers that are formed on the web. The spread of the disinformation is therefore determined by the size of the echo chambers. Del Vicario et al. (2016) also find that echo chambers play a role and find that the cascade behaviour of conspiracy theories is different than that of science stories. Kumar and Shah (2018) also see a role played by echo chambers, suggesting that the impact of disinformation is high when individuals see the same message repeated over and over, as would occur in echo chambers. However, in their review, Tucker et al. (2018) question the claim that echo-chambers play a central role, as other research has shown that users are exposed to a variety of viewpoints on online media. Vosoughi, Roy, and Aral (2018) consider only one type of content (news) and note that false news stories spread faster and further, and suggest this might be because they are more novel. They also find that bots spread both true and false news at similar rates, suggesting that the reach of false news is due to human behaviour. Budak, Agrawal, and El Abbadi (2011) propose that looking at dynamics of spread can be the solution to stopping the spread of disinformation: if influential users can be "infected" with true information early on, this can halt the campaign.

### 2.1.2 Psychological perspective - Why does disinformation work?

Reviewing relevant literature, Kumar and Shah (2018) provide three reasons why people are vulnerable to disinformation: People are unable to tell if information is false, particularly if it has been well-crafted, echo-chambers result in frequent exposure which increases belief, and mechanisms like confirmation bias, naïve realism and social norms lead people to seek confirmation of their own beliefs and the acceptance of peers.

When it comes to an individual's ability to detect disinformation, Pennycook, Cannon, and Rand (2018) found that people who have high "bullshit receptivity" and lower analytical thinking skills are more likely to perceive fake news as being accurate.Uscinski et al. (2020) Studied belief in COVID-19 related conspiracies,

and found that belief in these was predicted by denialism, conspiracy thinking, and ideological motivations (i.e. Trump supporters were more likely to believe these conspiracies).

Related to the issue of echo-chambers and repeated information, Pennycook and Rand (2018)) found that the illusionary truth effect (statements seem more plausible when they are repeated) held for political disinformation, with a significant but small increase in perceived accuracy of fake news which participants had been exposed to previously. Notably, they showed that warnings that content was false did not decrease this effect. Closely related is the "continued influence effect" where people continue to rely on information they know to be false. Warnings and alternative explanations have been shown to reduce this effect, but cannot remove it completely Ecker, Lewandowsky, and Tang (2010).

Two cognitive biases that likely play a role in the effect of disinformation are anchoring and confirmation bias. Anchoring is the process where the first data one receives strongly influences any future beliefs or estimates about the truth. Jost, Pünder, and Schulze-Lohoff (2020) showed that this also occurs in the context of fake news, even when people are aware the information they receive is false Confirmation bias occurs when people are (unconsciouly) motivated to believe certain statements and reject others based on their previous beliefs. Taber and Lodge (2006) showed that in a political context, people are much more likely to accept arguments that support their previous beliefs and will pick apart counter-arguments. A similar process likely occurs with disinformation.

Little is know as of yet how these factors (personal skills, biases) are different across different contexts. Bessi et al. (2015) showed that on Facebook, people who were involved in conspiracy theories on one topic were more likely to be involved in other conspiracy theories as well. There is also some evidence that personal involvement increases the willingness of people to spread rumors, especially if they induce fear (Chua & Banerjee, 2018), suggesting that besides the topic, the "stakes" associated with disinformation can also influence how it is processed.

### 2.1.3 Societal perspective - What are the effects of disinformation?

Research on the subsequent effect of this exposure to the individual is limited. Based previous research on the effectiveness of political ads, Allcott and Gentzkow (2017)) claim that the exposure of disinformation based on a dataset of fake news about the 2016 US election had a negligible effect because the average American was only exposed to a handful of false stories. Guess et al. (2020) found that consumption of fake news was linked to distrust towards media and stronger feelings of polarisation. In additional experiments, they showed that a single exposure to a false story would increase the belief of the participant in the (political) claim made in that article. Though this change was significant it was also small, so it is not yet clear if there is a meaningful effect.

Disinformation may also work in less direct ways. For example, people may have wrong perceptions about public policy or events, and these may be encouraged by disinformation (i.e. exaggerating the money spent on welfare could be beneficial to the agenda of the Republican party.) (Tucker et al., 2018). Kolmes (2011) argues that a lack of belief in man-made global warming in the American public was the result of disinformation that was deliberately spread

by fossil fuel companies, but such a level of influence would be extremely hard to quantify. Indeed, the broader the perspective and the less defined the case, the harder it becomes to describe the effects of disinformation using emperical data. However, even if an effect is not immediately tangible, it does not mean it is absent. Asmolov (2018) believes that disinformation is effective not because of its actual content, but because the arguments it creates sever social ties, resulting in a polarized society. However, the direction of causality is not clear - some argue that the prevalence of echo chambers is exaggerated, and that polarization precedes large-scale online disinformation campaigns, which then make use of resulting biases (and possibly exaggerate them) (Tucker et al., 2018). With a broader perspective, the possible sollutions to the problem of disinformation also change. Lewandowsky, Ecker, and Cook (2017) reject that with better communication techniques the issue of disinformation can be solved. In their work about the "Post-Truth Era" they identify a number of trends that they believe have contributed to the issue, ranging form a decrease in trust, growing inequality, polarization, and an (online) media landscape that rewards extremism. They state *"...post-truth claims [...] do not seek to establish a coherent model of reality. Rather, they erode trust in facts and reality, to the point where facts no longer matter or are not even acknowledged to exist."*

Based on this brief review, we can make a number of observations. The first is that efforts aimed at detecting and removing disinformation would need to focus on doing so almost immediately once a piece of disinformation is introduced into an online environment, since the psychological perspective suggests that warnings or removal at a later stage would be unsuccesful. This is already challenging, since actors who deliberately spread disinformation can change their tactics to avoid detection. But even assuming that detection is feasible, mass deletion of posts might have other risks. Given that individuals who are most likely to believe in disinformation are also distrusting towards media and authority, this could also create an effect where these groups move to other (less-controlled) platforms. Another observation is that polarization and echo-chambers are likely related to the succes of disinformation, but the direction and srength of this relationship remains up for debate. This is closely related to the outstanding issue of long-term effects of disinformation campaigns, which are hard to study emperically and occur in a wide context of social, politcal and economical factors that may all determine the effect of disinformation. The fact that these mechanisms remain unclear, combined with the current knowledge on what individuals are vulnerable to disinformation, suggest that a broader, societal perspective on disinformation is needed.

## 2.2 Network Science

### 2.2.1 Why is network science relevant to disinformation?

Online disinformation raises concern because it can be spread fast and far. This is due to two reasons: 1) Individuals can (unknowingly) act as amplefiers of content in their social circle, by easily sharing content with their connections or by drawing attention to it, and 2) the physical and digital infrastructure of the internet enables and encourages a number and span of connections that was not possible before. In this sense, online disinformation campaigns harness the

social and digital infrastructure of the internet to their benefit. Both the social and digital infrastrucutres are *networks*, which motivates follwing discussion of network science. Additionally, network science has long been used to study topics closely aligned with the problem of disinformation, such as the diffusion of information over a population and the rise (and fall) of political movements (Guilbeault, Becker, & Centola, 2018). Since In the network science approach, a system is modelled as a collection of nodes and edges. Nodes are the object of interests (such as individuals or computers) and the edges indicate a connection between the nodes. The underlying assumption of network science is that these connections are fundamental in understanding the system of interest as a whole (Brandes, Robins, McCranie, & Wasserman, 2013).

A key feature of many real-world networks is that they are not random – which means that there is something driving the formation of networks, and in turn, the formation of a network can be used to achieve certain aims (Newman, 2003). An important discovery in network science was that of the small-world model by Watts and Strogatz (1998), who showed that this network structure occurs in many real-world networks and demonstrated that diseases spread much faster over these networks. Small-world networks can mathematically be described as having a mean shortest distance between nodes that scales logarithmically or slower with the number of nodes N (Newman, 2003). Barabási and Albert (1999) found many real-world networks to have a degree distribution that followed a power-law distribution and called these networks scale-free networks. They showed that these types of network are created if networks are grown following preferential attachment: new nodes attach to nodes which already have a high number of edges.

### 2.2.2 Information Diffusion, (Complex) Contagion and Network Strucutres

As shown above, several different types of networks exist, influenced by the contexts in which they grew. Nowadays, many of the networks used daily are not grown organically – they are at least partly designed. Therefore, it is relevant to consider if design choices influence information diffusion – a highly relevant process when studying disinformation. I consider two types of network characteristics. Structural characteristics relate to the network properties discussed above, such as the distribution of node degrees and the average path length. Node characteristics relate to what behaviour a node can perform, i.e. how many neighbours it can communicate to at one time, the maximum number of connections it can maintain, or when it will form or break a connection.

The structure of a network has effects on how information, or a message, is propagated over it. Bampo, Ewing, Mather, Stewart, and Wallace (2008) studied the effect of network topology on the success of a viral marketing campaign by simulating it over a random, small-world and scale-free network. The scale-free network was more sensitive to changes in the seed size and the average number of connections a message was passed on to.

Some insight on the influence of node attributes can be gained from a marketing study, which showed that the spread of a Facebook application differed significantly depending on whether users could personally invite their friends, or

if the application sent out passive notifications Aral and Walker (2011). Though personal invitations had a higher adoption rate, passive notifications were sent far more often and as a result, were more effective. Therefore, how a node can behave in a network influences diffusion. This also concerns how a node became active in the diffusion process in the first place, or how contagion works.

Long before network science appeared as a stand-alone discipline, sociologist Granovetter (1977) showed that weak ties – edges that connect individuals who otherwise have little in common – greatly increase the reach and speed of information spreading over a network. This finding was later confirmed by the small-world model of Watts and Strogatz (1998). This so called "Strength of Weak Ties" explains why information and diseases can spread rapidly across the world – one edge is enough to link a group that is otherwise isolated from another. Centola and Macy (2007) argued, however, that the strength of weak ties is dependent on what exactly is being diffused in the network. They showed that when considering complex contagion, where an individual needs to be exposed to multiple sources before being "activated" themselves, weak ties can lower diffusion. Whereas certain types of information or diseases can be passed on by a single moment of contact (simple contagion), more complex behaviours or beliefs may only be adopted once an individual sees that multiple of its connections have done so. As a result, on the same network, the pattern and size of diffusion differs depending on whether its spread is governed by simple or complex contagion.

### 2.2.3 Multiplex networks

The literature discussed so far concerns properties of a single network. However, individuals seldomly partake in only one network. Multiple networks that are related to each other can be described through multiplex networks. Multiplex networks are networks where nodes are linked to each other with more than one type of edge, and the edges of one type that connect nodes together form one layer (Lee, Min, & Goh, 2015). Multiplex networks are of interest because interaction between the different layers can create non-additive and non-linear effects on the overall network. This is also why it is important to use multiplex networks when exploring systems that include multiple types of connections in reality – if the corresponding network model only has one layer, it may not correctly describe the behaviour of the system it should represent (Lee et al., 2015).

The general network properties discussed in the introduction also apply – with some modification – to multiplex network. An additional property of interest in multiplex networks is the correlation between different layers. Correlation can be described in multiple ways, ranging from the extent of node multiplicity (i.e. how many vertices appear in different layers), to interlayer degree correlation, edge overlap, and cross-layered clustering coefficients (Lee et al., 2015).

In terms of diffusion over networks, it has been shown by Brummitt, Lee, and Goh (2012) in a case of threshold activation (a form of complex contagion, i.e. x number of neighbours need to be activated for a node to be activated), multiplex networks can show cascade effects even if the individual layers are not susceptible to global cascades. This shows that when studying cascade effects, it

is necessary to consider any multiplicity in the system of interest. The same authors later showed that heterogeneity in threshold activation rules can enhance or inhibit cascades, depending on how many nodes require the threshold to be met in all layers (Lee, Brummitt, & Goh, 2014). Sahneh and Scoglio (2014)) showed that, when considering the diffusion of competing and exclusive viruses, in a multiplex network there exists an equilibrium in which the two viruses to coexist simultaneously. In a single-layer network this is not possible: one of the two will always come to fully dominate the other.

Yağan and Gligor (2012) consider the case where the influence of two types of connections is weighted (i.e. the influence of a neighbour in one network is greater than that in the other) and analyse how this influences complex (threshold) contagion. Their result show that both average degree and the relative difference in weighted influence strongly affect the probability of a global cascade, as well as its subsequent size.

The fact that network multiplicity has such a strong effect on probability and size of cascades has substantial consequences for studying the effect of disinformation. Most studies on the spread of disinformation (or closely related phenomena) (these will be discussed in detail in the following section), model the spread of beliefs over a single layer network. However, individuals gain information from multiple sources and in the case of online disinformation, are likely exposed to mediating (or amplifying) influences from their offline ("real") network of family and friends. Therefore, I propose that in order to properly study the spread of disinformation over a network, one needs to consider a multiplex network of at least two layers representing the online network and the offline network.

## 2.3 Psychology of advertising

Fennis and Stroebe (2015) define advertising as "any form of paid communication by an identified sponsor aimed to inform and/or persuade target audiences about an organization, product, service or idea." Until the 20th century, advertisements focussed on the use of information or arguments to convince potential customers. In the early 1900s, advertisements also began using emotional appeals and the projection of ideas and beliefs. Both approaches have coexisted since then.

The psychology of advertising aims to identify how (characteristics) of advertisements affect individuals and to understand the psychological processes behind those effects (Fennis & Stroebe, 2015). The field was pioneered by Scott (1916) with his book "The Psychology of Advertising". Generally, three types of outcomes are studied: cognitive responses (beliefs and thoughts about a brand), affective responses (moods and emotions), and behavioural responses (buying a product, switching to a different brand) (Fennis & Stroebe, 2015).

Advertisers are mainly interested in changing the attitudes (which are made up of both cognitive beliefs as well as the affective state) of potential customers, as well as ensuring that a certain attitude leads to the desired behaviour. There are several theories on how attitudes change. One of the earliest was the Yale

Attitude Change approach, developed by Hovland, Janis, and Kelley (1953), which focused on factors related to the source, the form of communication, and the audience. They also proposed several stages in which a message was processed, a framework that was extended in the Information Processing Model of McGuire (1968). He suggested six stages in which a message is processed: presentation, attention, comprehension, yielding, retention and behaviour. The stage of yielding is where attitude change occurs. Both the Yale Attitude Change model and the Information Processing Model assume that this is an active and conscious process. A different theory was proposed by Greenwald (1968) in his Cognitive Response Model. He shifted away from models of learning and memorization of new information and stated that persuasion depended mainly on whether someone accepts the premise of the message when viewing it.

Dual process theories were developed to address the fact that the variables suggested by earlier models could not consistently explain the response to a message Xu (2017). The two prominent dual process models are the Elaboration Likelihood Model by Petty & Cacioppo (1986) and the Heuristic Systematic Model by Chaiken et al (1980). In general these models assume that persuasion can happen in two ways: The first is through a conscious process in which the recipient engages with new information and adjusts their beliefs accordingly. The second route is reliant on heuristics and is used when someone's motivation or skills are too low to process the information fully. Building on the dual process theories, Kruglanski and Thompson (1999) proposed the unimodel, suggesting that the two types of processing were not separate processes, but variations of the same process that could be triggered by different contextual factors.

Advertisers are not only interested in convincing customers that their product is good, but hope that this is also translated in the behaviour of the customer. For a long time, attitude was seen as the most important predictor of behaviour. However, empirical studies showed that it could not be the sole explanation (Fennis & Stroebe, 2015). Fishbein and Ajzen formulated the Theory of Reasoned Action in the 1970s, which also took "subjective norms" into account (which describe social norms and the desire to comply with those norms by (not) performing a certain behaviour). They later extended this into the Theory of Planned Behaviour, which added the "perceived behavioural control" individuals had over the action they may perform (Ajzen, 1991). The three elements (attitude, subjective norm, and perceived behavioural control) predict the intention to perform a certain behaviour well, but there still remains a gap between intention and performing the behaviour itself (Fennis & Stroebe, 2015).

This theory, again, is based on a conscious process in which an intention is formed. Yet advertisers may also be interested in processes where certain behaviour is performed automatically. The simplest form of this is habitual behaviour, where past behaviour is a better predictor than someone's intention. However, advertisers may also try to prime behaviours to meet some goal (i.e. unconsciously smelling cleaning product might lead someone to formulate the goal of cleaning their house) (Fennis & Stroebe, 2015).

## 2.4 Propaganda

## 2.5 Simulation modelling

Agent-based modelling is a method that allows for the combination of many factors and mechanism that make up a complex system in order to study their interaction and resulting behaviour (Flache et al., 2017). In the social sciences it is a useful approach precisely because it allows for the combination of different disciplines in order to mimic some of the multidimensionality of the real world (Epstein, 2006).

Nowak, Szamrej, and Latané (1990) were among the first to suggest that individual psychological processes proposed in theory could be simulated in order to see if they produced the expected group behaviour. They simulated Social Impact Theory, which assumes that social influence is caused by the strength of sources, the immediacy of sources, and the number of sources supporting a certain attitude. They modelled binary opinion change on a gridded population, and showed that clusters of minority opinions could be maintained.

Deffuant, Amblard, Weisbuch, and Faure (2002) studied how extreme opinions can prevail in communities. They modelled a population in which individuals have random 1-to-1 meetings, and in such a meeting the opinion of either individual changes based on the opinion and uncertainty of the other. Extremists were modelled as having both extreme opinions and confidence (low uncertainty). If the general population has high uncertainty, extremists are able to polarise a population (i.e. bipolar extremism), or have the population convert to either extreme standpoint.

Flache et al. (2017) reviewed a large set of agent-based models on social influence. They showed that depending on what assumption is made on how individuals influence each other, the overall pattern of system behaviour changes significantly – from consensus to fragmented clusters of opinions to polarization. They stress that there are many models describing social influence, but that comparison between models is lacking, as well as empirical testing.

Ross et al. (2019) used agent-based modelling to study the phenomenon of a "Spiral of Silence", in which people with the minority opinion do not dare speak up, causing a positive feedback loop. In particular, they studied how bots could influence opinion formation by triggering a spiral of silence. They found that bots could do so, even if their influence was a fourth of that of real humans. The denser the network, the stronger this effect.

(Battiston, Cairoli, Nicosia, Baule, & Latora, 2016) Chattoe-Brown, E. (2014). Using agent based modelling to integrate data on attitude change

Notably, many of these studies tend to isolate the phenomena of interests. For example, Deffuant et al. (2002) considered several mechanisms for how two individuals exchange opinions, but leave out of scope how, why or when individuals meet to do this. On the flipside, Ross et al. (2019) only consider the influence of someone's online connections in their decision to speak up, whereas

this decision is likely influenced by many social and political factors. Such scoping decisions are understandable from the perspective of wanting to isolate mechanisms behind a phenomena, but make it hard to translate the results to the messier real world.

# 3 Research Gap

- **Dependence on scope & contextual factors:** Psychological studies showed that individual beliefs and skills impact if people fall for disinformation, network studies show that the inclusion of multiple networks can change diffusion behaviour and even enable information cascades where they were not possible, and a simulation study found that changing the assumptions on how people share opinions determines the outcomes on a system level. Yet most literature does not (explicitly) address this dependence on context and sensitivity to scoping when it comes to disinformation.
- **A variety of plausible (theoretical) explanations:** Explanations of why disinformation is so dangerous range from individual skills in media literacy, psychological biases we are all subject to, to features of the online platforms over which it is spread (i.e. bots, algorithms that encourage extremism), to broader social and economic trends such as a lack of trust in institutions and (economic) inequality. Relating to the previous point, which explanation is most appropriate is likely dependent on context. There is, however, no comprehensive work to explain what this variety of explanations means when trying to understand the problem on online disinformation.
- **Lack of a system-level, multidisciplinary/transdisciplinary approach:** The previous two points also strongly relate to this observation: There is no work that tries to comprehensively bring together work on disinformation from various disciplines to understand how we may understand the problem on a system level (Lewandowsky et al. (2017) comes closest ).

# 4 Research Approach

## 4.1 Disinformation is a post-normal problem

Post-normal science refers to the practice of science in contexts where uncertainty is high, there is no agreement on values, and decision-making involves both high stakes and time pressure. The idea of post-normal science was developed in the 1990s to address the position of science in these contexts, mainly from an environmental perspective Funtowicz and Ravetz (1995); Ravetz (1999). I argue that online disinformation should also be treated like a post-normal problem, for several reasons:

- *There are multiple plausible approaches and explanations to disinformation.* Ranging from technological solutions to theories on what psychological effects play a role it the succes of disinformation, many different

disciplines have tackeled disinformation. None of these approaches need to be wrong, instead, they are indicative of a plurality of valid perspectives.

- *The values of individuals and organisations relating to disinformation are conflicting.* In the most obvious example, clearly those who want to stop disinformation campaigns have different values than those who set them up. But governments and tech companies have also not agreed on suitable ways to address disinformation, let alone the users of social media platforms.

- *It cannot be treated with a reductionist approach where the system is divided into smaller elements that are analysed in isolation.* Of course, specific studies can increase our understanding in how, for example, cognitive biases influence the processing of fake news messages. However, given the scale and the complexity of the problem, to properly understand it, online disinformation needs to be addressed on a system level.

If online disinformation is a post-normal problem, it should be also addressed as such - it's uncertainty and complexity should be addressed upfront rather than brushed over by any proposed theory or model. Yet at the same time, researc output needs to be useful - just saying there are multiple valid explanations for the effect of online disinformation campaigns does not provide a solution. A theory of disinformation is needed that allows us to explore the impact of different contexts, theories and values, yet can also be applied to specific cases to understand why a disinformation campaign created a certain effect.

## 4.2 A Metamodel to enable theoretical exploration

I propose to develop a *metamodel* which enables theoretical exploration of the online disinformation problem. If a model is an abstraction of reality for some given purpose, a meta-model describes the next level of abstraction: It describes the structure and behaviour of a class of models and is not case-specific (Sprinkle et al. 2010). Metamodeling is commonly used in software development to reduce the resulting software's sensitivity towards change (Atkinson & Kühne, 2003). UML is an example of a metamodel - with its rules and language specific models can be created. In simulation modeling, meta-models can be used to describe how different (sub)components of a model relate or communicate to each other (Cetinkaya 2010).

A metamodel of disinformation then describes the elements describes the behaviours (functions) and the relationships between these functions in order to describe the overall system of interest, without describing the functions themselves. For example, a superstructure of disinformation likely needs a functionality that describes how people internalise information, but for the functional superstructure itself it does not matter how this actually takes place. Such an approach allows the theory to be adapted to different contexts, or be updated when new insights arise. Moreover, it invites users (i.e. policymakers) to ask "what-if" questions, providing an upfront way to test both assumptions and policies. The metamodel then becomes a tool for theoretical exploration or exposition Edmonds (2017): It lets us test how changing assumptions or rela-

tions affect the outcome space and what elements or mechanisms are crucial in describing disinformation.

### 4.3 Research Goal & Questions

The research goal then is:

*To create a metamodel that allows for the formulation of a theory of disinformation that can be adapted to specific contexts and viewpoints.*

In order to reach this purpose, and to show the use of such a theory for policy analysis, the following research questions are formulated:

1. What are the requirements of the superstructure model and its necessary scope?
2. What elements or mechanisms related to disinformation need to be included in the model regardless of context?
3. Which elements or mechanisms in the superstructure are dominant in determining the emergent behaviour of the system?
4. What policies can be formulated to lower the effect of a disinformation campaign on a system? Do these policies have the same effect across cases? Who controls the policy levers to implement them?

## 5 Methodology

Given that there is no "true" theory of disinformation, the goal of this research is to create a metamodel that allows us to explore and make sense of online disinformation and its effects on a system level. In the broad sense, this entails the design of a metamodel of online disinformation. Such a metamodel can then be used to implement case-specific simulation models.

Therefore, we take the design science approach: Our knowledge and understanding of the problem (in this case, disinformation) is generated both *during the design process*, as well as *with the application of the designed artefact* (the metamodel) (A. R. Hevner, March, Park, & Ram, 2004). In other words, this research contributes to a system-level understanding in two ways: The metamodel that will be developed allows for sensemaking and exploration of the problem, but the development of the metamodel also contributes to our understanding of disinformation, as we learn about what compontents, relations and assumptions are important.

To structure the project, I use the design science research cycles as described by A. Hevner and Chatterjee (2010). The "core" activity is the design of metamodel in the design cycle, which iterates between designing & building and evaluation. This process is fed by two other cycles: The relevance cycle and the rigor cycle. The relevance cycle links the design proces to the problem environment and provides the requirements and (institutional) context in which

the metamodel will be used. The relevance cycle closes by testing the metamodel in its environment, providing information for its evalution. The rigor cycle links the design process to existing knowledge. In this case, that entails both domain-specific literature on disinformation (as described in the literature review) as well as research on (meta)modelling practices, component-based modelling and simulation models on related problems. Linking the design process to this "knowledge base" is essential to ensure the metamodel is based on science - both in terms of its content as well as its implementation. As stated before, during the design process we will also learn more about the problem itself. The rigor cycle ensures this new knowledge is fed back into the knowledge base.

Once the boundaries and criteria of the metamodel are established, the research becomes iterative - that is, the experience obtained in creating a simulation model will feed into both the metamodel and subsequent simultation models, and exploratory analysis of a simulation model can inform areas of focus in the next case study. This is typical for design driven research, where our understanding of the system is improved through the design process A. R. Hevner et al. (2004).The *goal* of the research is to create a tool for theoretical exploration and sensemaking, and during the *process* of developing this tool we will gain the knowledge necessary to do so.

The complete research design is shown in Figure **??**

## 5.1 Defining the purpose and requirements of the metamodel

In order to design and evaluate a metamodel, its purpose and requirements need to be clear. The motivation to create the metamodel stems from what we found in the knowledge base - a lack of a systems perspective on disinformation (see the discussion in the Research Gap). The first question that then needs to be answered is: Who would benefit from such a systems perspective? This is the input from the **problem environment** to the **design process**. Policy makers, scholars, activists and platform owners alike would benefit from increased understanding. I therefore propose to start with an **actor analysis** to map interested parties, their interests and policy levers they have access to. For the metamodel to be useful, it needs overlap with an actor's policy levers - no use in presenting a metamodel to Facebook that focusses on the relationship between the creators of disinformation campaigns and economic security. The actor analysis will be done in two phases: First through **document analysis** and then by **interviews with relevant actors** identified in the first stage. Based on the outcomes of this actor analysis, the **goal and requirements** of the metamodel can be formulated.

## 5.2 First Design Cycle

Given the goals and requirements and the input from the knowledge base (the literature review), the first iteration of the design cycle can start. This begins by answering the question: ***What exactly does the metamodel consist of?*** I prpose that as a starting point, the metamodel is made up of two components:

a **conceptual model** and an **ontology**. The conceptual model describes on a high level how different mechanisms relate to each other (i.e. how individuals connect on a network and how individuals process information internally) and will be in the form of a **system diagram**. An ontology is a formal naming system to describe and categorize entities and their relationships. Given that disinformation involves many different disciplines, an ontology can serve as a shared vocabulary to enable communication between them. The ontology can be created in Webprotege[1] or OWLready[2] using the open-world assumption: The ontology should allow for future extention, and it is assumed we cannot immediately define the complete system. Creating an ontology also makes the translation of the metamodel into case-specific simulation models easier (Benjamin, Patki, & Mayer, 2006): The relations and hierarchy between key components have already been formulated, as well as the constraints of the system. Such case-specific simulation models form the the backbone of the following design and relevance cycles.

## 5.3 Simulating Case Studies

Once the basic structure of the metamodel is known, the design is improved, expanded and changed through an iterative process of simulation models created for a specific case using the metamodel. These simulation models serve a dual purpose: They drive the design cycle, and are also an example of "field-testing" done in the relevance cycle. Using simulation models to iteratively improve the metamodel design stems from the generative science approach Epstein (2006): Each cycle, the question is if - with the given metamodel - the observed effects in the case study can be recreated using a simulation model. If it is not possible, the metamodel is incomplete - it lacks complexity, or misses specifc elements. Experimentation can then show what should be added to the metamodel.

For now I propose to use three case studies, each with a different scope, timeline or application of disinformation, to cover as much of the different phenomena of disinformation given the time constraints:

- BlackLivesMatter discourse during the 2016 US election. Disinformation campaigns, probably conducted by the Russian Internet Research Agency, targeted both conservative right wing and progressive activists online, and even managed to organise actual protests. These campaigns have been widely documented (i.e. Frenkel (2018), Arif, Stewart, and Starbird (2018), Select Committee on Intelligence (2019)).
- The Corona/5G conspiracy exposure to extreme action
- Greenwashing campaign by fossil fuel companies: goals &long term effects

Data collection for the case studies will be done through mixed methods, i.e. interviews/focus groups (or a corona-proof alternative), document analysis, network analysis, and data from crawling social media posts. The simulation model itself will Likely be primarily agent-based (i.e. using NetLogo), but other modeling paradigms are possible as the metamodel or the case demands. The

[1] https://webprotege.stanford.edu/
[2] https://pythonhosted.org/Owlready/

simulation models, once implemented, need to be tested extensively on the assumptions made and the effect of using policy levers (as identified in the actor analysis) **Sensitivity analysis** and **exploratory modelling analysis** will therefore be conducted using the ema-workbench (Kwakkel, 2017).

### 5.4 Evaluation of the metamodel

- Evaluate the metamodel on the basis of the criterea established previously
- Analysis of the results obtained by the cases - what can they tell us about online disinformation on a higher level?
- Answering the "Now what?" question

## References


Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election [Journal Article]. *Journal of economic perspectives*, *31*(2), 211-36.

Aral, S., & Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks [Journal Article]. *Management Science*, *57*(9), 1623-1639. Retrieved from `https://pubsonline.informs.org/doi/abs/10.1287/mnsc.1110.1421` doi: 10.1287/mnsc.1110.1421

Arif, A., Stewart, L. G., & Starbird, K. (2018). Acting the part: Examining information operations within# blacklivesmatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, *2*. Retrieved from `http://faculty.washington.edu/kstarbi/BLM-IRA-Camera-Ready.pdf`

Asmolov, G. (2018). The disconnective power of disinformation campaigns [Journal Article]. *Journal of International Affairs*, *71*(1.5), 69-76. Retrieved from `www.jstor.org/stable/26508120`

Bampo, M., Ewing, M. T., Mather, D. R., Stewart, D., & Wallace, M. (2008). The effects of the social structure of digital networks on viral marketing performance [Journal Article]. *Information Systems Research*, *19*(3), 273-290. Retrieved from `https://pubsonline.informs.org/doi/abs/10.1287/isre.1070.0152` doi: 10.1287/isre.1070.0152

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks [Journal Article]. *Science*, *286*(5439), 509-512.

Benjamin, P., Patki, M., & Mayer, R. (2006). Using ontologies for simulation modeling. In *Proceedings of the 2006 winter simulation conference* (p. 1151-1159). doi: 10.1109/WSC.2006.323206

Bessi, A., Zollo, F., Del Vicario, M., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Trend of narratives in the age of misinformation [Journal Article]. *PloS one*, *10*(8), e0134641-e0134641. Retrieved from `https://www.ncbi.nlm.nih.gov/pubmed/26275043` `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4537127/` doi: 10.1371/journal.pone.0134641

Brandes, U., Robins, G., McCranie, A., & Wasserman, S. (2013). What is network science? [Journal Article]. *Network science*, *1*(1), 1-15.

Brummitt, C. D., Lee, K.-M., & Goh, K.-I. (2012). Multiplexity-facilitated cascades in networks [Journal Article]. *Physical Review E*, *85*(4), 045102.

Budak, C., Agrawal, D., & El Abbadi, A. (2011). Limiting the spread of misinformation in social networks [Conference Proceedings]. In *Proceedings of the 20th international conference on world wide web* (p. 665-674). ACM.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties [Journal Article]. *American journal of Sociology*, *113*(3), 702-734.

Chua, A. Y. K., & Banerjee, S. (2018). Intentions to trust and share online health rumors: An experiment with medical professionals. *Computers in Human Behavior*, *87*, 1–9.

Deffuant, G., Amblard, F., Weisbuch, G., & Faure, T. (2002). How can extremism prevail? a study based on the relative agreement interaction model [Journal Article]. *Journal of artificial societies and social simulation*, *5*(4).

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., ... Quattrociocchi, W. (2016). The spreading of misinformation online [Journal Article]. *Proceedings of the National Academy of Sciences*, *113*(3), 554-559. Retrieved from `https://www.pnas.org/content/pnas/113/3/554.full.pdf` doi: 10.1073/pnas.1517441113

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation [Journal Article]. *Memory & Cognition*, *38*(8), 1087-1100. Retrieved from `https://doi.org/10.3758/MC.38.8.1087` doi: 10.3758/MC.38.8.1087

Edmonds, B. (2017). Different modelling purposes [Book Section]. In *Simulating social complexity* (p. 39-58). Springer.

Epstein, J. M. (2006). Agent-based computational models and generative social science [Book Section]. In *Generative social science: Studies in agent-based computational modeling* (STU - Student edition ed., p. 4-46). Princeton University Press. Retrieved from `www.jstor.org/stable/j.ctt7rxj1.5` doi: 10.2307/j.ctt7rxj1.5

Fallis, D. (2015). What is disinformation? [Journal Article]. *Library Trends*, *63*(3), 401-426.

Fennis, B. M., & Stroebe, W. (2015). *The psychology of advertising* [Book]. Psychology Press.

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

Frenkel, S. (2018, August). *How a fake group on facebook created real protests.* Newspaper Article. Retrieved from `https://www.nytimes.com/2018/08/14/technology/facebook-disinformation-black-elevation`

Funtowicz, S. O., & Ravetz, J. R. (1995). Science for the post normal age. In *Perspectives on ecological integrity* (pp. 146–161). Springer.

Granovetter, M. S. (1977). The strength of weak ties [Book Section]. In *Social networks* (p. 347-367). Elsevier.

Greenwald, A. G. (1968). Cognitive learning, cognitive response to persua-

sion, and attitude change [Journal Article]. *Psychological foundations of attitudes*, *1968*, 147-170.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is "love" evading hate speech detection. In *Proceedings of the 11th acm workshop on artificial intelligence and security* (pp. 2–12).

Guess, A. M., Lockett, D., Lyons, B., Montgomery, J. M., Nyhan, B., & Reifler, J. (2020). "fake news" may have limited effects beyond increasing beliefs in false claims [Journal Article]. *Harvard Kennedy School Misinformation Review*, *1*(1). Retrieved from https://misinforeview.hks.harvard.edu/article/fake-news-limited-effects-on-political-

Guilbeault, D., Becker, J., & Centola, D. (2018). Complex contagions: A decade in review [Book Section]. In S. Lehmann & Y.-Y. Ahn (Eds.), *Complex spreading phenomena in social systems: Influence and contagion in real-world social networks* (p. 3-25). Springer International Publishing.

Hevner, A., & Chatterjee, S. (2010). Design science research in information systems. In *Design research in information systems.* Springer.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research [Journal Article]. *MIS Quarterly*, *28*(1), 75-105. Retrieved from www.jstor.org/stable/25148625 doi: 10.2307/25148625

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion* [Book]. New Haven: Yale University Press.

Hrčková, A., Srba, I., Móro, R., Blaho, R., Šimko, J., Návrat, P., & Bieliková, M. (2019). Unravelling the basic concepts and intents of misbehavior in post-truth society/desentrañando conceptos básicos e intentos de mala conducta en la sociedad de la post-verdad. *Bibliotecas. Anales de Investigación*, *15*(3).

Jost, P. J., Pünder, J., & Schulze-Lohoff, I. (2020). Fake news-does perception matter more than the truth? *Journal of Behavioral and Experimental Economics*, *85*, 101513.

Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2020). Political astroturfing on twitter: How to coordinate a disinformation campaign. *Political Communication*, *37*(2), 256–280.

King, G., Pan, J., & Roberts, M. E. (2017). How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American political science review*, *111*(3), 484–501.

Kolmes, S. A. (2011). Climate change: A disinformation campaign [Journal Article]. *Environment: Science and Policy for Sustainable Development*, *53*(4), 33-37. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/00139157.2011.588553 doi: 10.1080/00139157.2011.588553

Kruglanski, A. W., & Thompson, E. P. (1999). Persuasion by a single route: A view from the unimodel [Journal Article]. *Psychological Inquiry*, *10*(2), 83-109.

Kumar, S., & Shah, N. (2018). False information on web and social media: A survey [Journal Article]. *arXiv preprint arXiv:1804.08559*.

Kwakkel, J. H. (2017). The exploratory modeling workbench: An open source toolkit for exploratory modeling, scenario discovery, and (multi-objective) robust decision making. *Environmental modelling & software*, *96*.

Lee, K.-M., Brummitt, C. D., & Goh, K.-I. (2014). Threshold cascades with response heterogeneity in multiplex networks [Journal Article]. *Physical Review E*, *90*(6), 062816.

Lee, K.-M., Min, B., & Goh, K.-I. (2015). Towards real-world complexity: an introduction to multiplex networks [Journal Article]. *The European Physical Journal B*, *88*(2), 48. Retrieved from https://doi.org/10.1140/epjb/e2015-50742-1 doi: 10.1140/epjb/e2015-50742-1

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition*, *6*(4), 353–369.

Liu, X., Nourbakhsh, A., Li, Q., Fang, R., & Shah, S. (2015). Real-time rumor debunking on twitter. In *Proceedings of the 24th acm international on conference on information and knowledge management* (pp. 1867–1870).

McGuire, W. J. (1968). Personality and attitude change: An information-processing theory [Journal Article]. *Psychological foundations of attitudes*, *171*, 196.

Mitra, T., & Gilbert, E. (2015). Credbank: A large-scale social media corpus with associated credibility annotations. In *Ninth international aaai conference on web and social media*.

Morgan, S. (2018). Fake news, disinformation, manipulation and online tactics to undermine democracy [Journal Article]. *Journal of Cyber Policy*, *3*(1), 39-43. Retrieved from https://doi.org/10.1080/23738871.2018.1462395 doi: 10.1080/23738871.2018.1462395

Newman, M. E. (2003). The structure and function of complex networks [Journal Article]. *SIAM review*, *45*(2), 167-256.

Nowak, A., Szamrej, J., & Latané, B. (1990). From private attitude to public opinion: A dynamic theory of social impact [Journal Article]. *Psychological review*, *97*(3), 362.

Papadopoulou, O., Zampoglou, M., Papadopoulos, S., & Kompatsiaris, I. (2019). A corpus of debunked and verified user-generated videos. *Online information review*.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news [Journal Article]. *Journal of experimental psychology: general*, *147*(12), 1865.

Pennycook, G., & Rand, D. G. (2018). Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking [Journal Article]. *Journal of personality*.

Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1589–1599).

Ravetz, I. (1999). What is post-normal science [Journal Article]. *Futures-the Journal of Forecasting Planning and Policy*, *31*(7), 647-654.

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 [Journal Article]. *European Journal of Information Systems*, *28*(4), 394-412. Retrieved from https://doi.org/10.1080/0960085X.2018.1560920

https://www.tandfonline.com/doi/full/10.1080/0960085X.2018.1560920
doi: 10.1080/0960085X.2018.1560920

Sahneh, F. D., & Scoglio, C. (2014). Competitive epidemic spreading over arbitrary multilayer networks [Journal Article]. *Physical Review E*, *89*(6), 062817.

Scott, W. D. (1916). *The psychology of advertising: A simple exposition of the principles of psychology in their relation to successful advertising* [Book]. Boston: Small, Maynard.

Select Committee on Intelligence. (2019). *Russian Active Measures Campaigns and Interference in the 2016 U.S. Election Volume 2: Russia's Use of Social Media with Additional Views* (Tech. Rep.). United States Senate. Retrieved from https://www.intelligence.senate.gov/sites/default/files/documents/Report$_V$olume2.pdf

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, *50*(3), 755–769.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., ... Nyhan, B. (2018). *Social media, political polarization, and political disinformation: A review of the scientific literature* (Report). Hewlett Foundation. Retrieved from https://eprints.lse.ac.uk/87402/1/Social-Media-Political-Polarization-and-Political-D

Uscinski, J. E., Enders, A. M., Klofstad, C., Seelig, M., Funchion, J., Everett, C., ... Murthi, M. (2020). Why do people believe covid-19 conspiracy theories? *Harvard Kennedy School Misinformation Review*, *1*(3).

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online [Journal Article]. *Science*, *359*(6380), 1146-1151. Retrieved from https://science.sciencemag.org/content/sci/359/6380/1146.full.pdf

Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks [Journal Article]. *Nature*, *393*(6684), 440.

Xu, Q. (2017). Dual process models of persuasion [Journal Article]. *The international encyclopedia of media effects*, 1-13.

Yağan, O., & Gligor, V. (2012). Analysis of complex contagions in random multiplex networks [Journal Article]. *Physical Review E*, *86*(3), 036103.

Zollo, F., & Quattrociocchi, W. (2018). Misinformation spreading on facebook [Book Section]. In *Complex spreading phenomena in social systems* (p. 177-196). Springer.