UiT The Arctic University of Norway

Institute of Philosophy and First Semester Studies

The Agent-Based Model of Pluralistic Ignorance

Sincerity, Conformity and Truth

Joanna Aleksandra Kreft

Master's Thesis in Philosophy, May 2024

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank Michael Morreau, my supervisor, for his knowledge, wisdom, and guidance in shaping this thesis. Thank you for your reminders to take breaks and enjoy life.

This thesis would likely still be a work in progress if not for the contributions of my fellow students: Andreas, Christine, and Sivert. Thank you for your insightful comments, the valuable discussions we have had, and, perhaps most importantly, for providing the moral support that kept me grounded throughout this journey.

Torstein, thank you for your help in navigating the initial stages of coding the simulation.

Last but not least, I extend my heartfelt thanks to my dearest friends Jennifer, Ragna, Sara, and Solvei for your unwavering support and occasional but much-needed distractions from writing.

Introduction

This thesis investigates the phenomenon of pluralistic ignorance (PI), clears up some conceptual confusion regarding its definition, and explores ethical problems connected to it. One main contribution of the thesis is a new agent-based computer simulation of PI. Observations of this simulation lead to tentative recommendations about how to deal with PI in social networks.

PI is a socio-psychological phenomenon that involves a systematic discrepancy between people's private beliefs and public behavior in certain social contexts (Bjerring & et al., 2014). Agents falsely believe that members of their group believe some proposition, when in fact all (or almost all) reject it. Individuals then think their private beliefs are unique to the group, and due to the desire to conform and be a good member of the group, they hide their sincere beliefs. This creates a vicious cycle, contributing to maintaining false beliefs about other people's beliefs (meta-beliefs). Given that we form meta-beliefs based on other people's words and actions, it is easy to see that if people are not genuine in expressing their attitudes, PI will be maintained - people will, at large, continue to believe that the behavior their peers engaged in is an indication of their privately held beliefs, even if they allow for such discrepancy in their own conduct.

PI is ubiquitous. It appears in different social contexts, be it a classroom, a small social gathering, a parliament, or even whole nations (Prentice & Miller, 1996, p. 162). Moreover, it can produce many undesirable consequences, for example by slowing down the process of changing social norms that are not publicly supported by the majority (O'Gorman, 1975). Therefore, I believe it is important to learn more about how it is formed and how it can be dispelled. One practically important example of PI that is studied in the literature concerns alcohol abuse among college students. It turns out that students are less comfortable with drinking than it would appear from their actual drinking behavior. Since everyone believes that everyone else is more comfortable with drinking than they actually are, there arises a social norm or social permission to drink as much as they actually do (Prentice & Miller, 1996).

In the first chapter, I first give a historical background of how the term *pluralistic ignorance* has been developed and used. I will clarify some common conceptual misunderstandings, as well as give my theoretical assumptions and their justifications. Those

preliminary steps are necessary to develop a sound agent-based computer simulation that can give insight into how pluralistic ignorance is maintained and how it can be resolved.

In the second chapter, I present my agent-based model of PI together with two experiments that explore ways in which PI can be dispelled. The first experiment shows the effect of introducing an invariably sincere agent (a *sincerian*, as I will say) into the network of agents in the state of PI. The second experiment explores what happens if instead, a *contrarian* (a person who always testifies against the perceived majority opinion) enters a social network. The results of this experiment show that both sincerians and contrarians can change the behavior of insincere agents and, under certain conditions, help them to share their beliefs sincerely.

The third chapter explores the ethical consequences of the results of the experiments, focusing on the conflict that arises between epistemic and moral values when we consider ways of dispelling PI.

In the conclusion, I offer a rationale for dispelling PI by considering its effect on epistemic practices. That people live their lives authentically and share their beliefs sincerely is, I argue, an important prerequisite for argumentative discussion and the collective search for truth.

1 Pluralistic Ignorance: Historical Background and Preliminary Discussion

This chapter gives a historical overview of the development of the concept of pluralistic ignorance (PI). It clarifies conceptual confusion about what PI is and what it is not and gives a formal definition of the phenomenon. It investigates also if PI can arise in the group of rational agents, and what, if anything, can be done to dispel it.

1.1 Historical Background

The term Pluralistic ignorance (PI) was first coined by Floyd Henry Allport, an American psychologist of the twentieth century. He first used the term in the 1931 book he co-wrote with his student Daniel Katz "Students Attitudes: A Report of the Syracuse University Research Study". In this work, Allport and Katz describe the results of their study at seven different colleges at Syracuse University. The term "pluralistic ignorance" was used in relation to the curious result of a questionnaire showing that students belonging to fraternities and sororities tend to report exclusive attitudes to non-members. When asked about it, the students motivate those attitudes not by personal feelings or convictions, but rather by concern about public opinion, i.e. the reputation of their fraternity/sororities. However, public opinion should be an aggregate of each individual attitude, and since the majority of fraternity/sorority members at least claim to be privately in favor of the inclusion of non-members, one can ask what can motivate those reported responses (Allport & Katz, 1931, pp. 150-152).

Allport and Katz offer two possible explanations for their findings. One of them is that the students simply do not recognize their own biases and are able to shield themselves from the accusations of prejudice, by simply shifting the blame to the "public opinion". However, if we want to grant the students self-awareness and a sincere lack of proclivity to exclude non-members based on personal feelings another explanation is required (Allport & Katz, 1931, p. 152). It is possible that the students are unaware of the private or personal feelings of the other group members and therefore form inaccurate beliefs about the public opinion within the group. Moreover, whenever they are to express their feelings towards the non-members they do so in the context of that imagined opinion. This, as Allport and Katz suggest, influences

the expressions of students' personal attitudes, and an inaccurate expression of group opinion becomes universally accepted. They call it pluralistic ignorance (Allport & Katz, 1931, p. 152).

Allport has previous to that study, investigated the mechanism of the impression of universality, which he believes plays a role in the cases of PI. The impression of universality takes place when an individual reacts to the signals from the environment in such a way as if they were coming from a much greater number of individuals. Allport writes “A number of references have been made to the attitude assumed by the individual when he knows that he is in the presence of a large company.” and “In terms of behaviour, we may say that the individual reacts to stimuli which he actually receives as if they were coming from an enormously greater number of individuals” (Allport, 1924, p. 305). Therefore, even if a small number of people conceal their true feelings and give false evidence of a certain attitude to an individual, who may then use his mental imagination to simulate what a public opinion is on that matter, then it is not necessary for that individual to survey a representative sample of a population to have an impression of knowing accurately the opinion of the majority (Allport, 1924, pp. 305-306). In the context of PI, an individual may only dare to question public opinion privately. The impression of individuality may explain how PI can form in a crowd even if the beliefs about the proposition in question were only explicitly stated by a limited number of individuals.

Many psychological phenomena useful for understanding individuals in the state of PI, were discovered through the studies of crowd behavior. Allport suggests that “the public” may be understood as “an imagined crowd in which (as he believes) certain opinions, feelings, and overt reactions are universal.” (Allport, 1924, p. 308) To help maintain the subjective experience of the reality of the group the press, rumor, and social projection play a role. That is, individuals not only receive the signal from their peers about their beliefs to form meta-believes but in addition are bombarded by the media with implications of widespread opinions and beliefs.

Social projection explains why individuals suppose that the attitude of those who cannot be observed is favorable to the propositions uttered by the speaker (Allport, 1934, p.306). One idea is the epistemic credibility given to the speaker if he is listened to by many people, as well as the immediate supportive reactions to the surrounding neighbors. Social

projection happens when the words are accepted by the hearing individual himself and taken as evidence that others just must think alike. This explanation contrasts, however, with what happens within individuals in the state of PI. They hear the speaker; they privately disagree and yet they do not infer that others may feel the same.

The way that Allport uses the term pluralistic ignorance suggests that he does not consider it a genuine social phenomenon, but rather a situation where a single person forms a wrong opinion about what the majority of a group that he belongs to. After the term was coined, it however came to be understood as a genuine social phenomenon arising in a group as a whole. This has to be explained.

In 1986 sociologist Hubert J. O'Gorman wrote already that PI "refers to erroneous cognitive beliefs shared by two or more individuals about the ideas, feelings and actions of others" (O'Gorman, 1986, p. 333). O'Gorman wants to make sure that this group phenomenon is clearly distinguished from personal distortion made by an individual in her perception of others. He does not think that the psychological mechanism that distorts the available information about others to the individual, is neither necessary nor sufficient to explain PI (O'Gorman, 1986). O'Gorman, therefore, indirectly questions Allport's and Katz's suggestion that PI can be neatly explained by reference to the mechanism of the impression of universality.

Why is it important not to treat PI simply as a psychological error on the side of an individual? The answer is that it requires a potentially incorrect assumption that PI must entail perceptual or cognitive distortion at the psychological level. However, there may exist environmental and social conditions that help to produce and sustain those false ideas within an individual. For example, an environment can be shaped in such a way as to give disproportionate attention to people expressing certain attitudes than others, which gives the impression that certain behaviors are more common (and therefore more widely supported) than they actually are (O'Gorman, 1986).

I think that understanding PI as a group phenomenon is the right one and that is the framework I will use in this thesis. PI understood in this way helps to answer the questions about the cultural validation of false ideas. False beliefs about other people's beliefs, feelings, or attitudes in general, because they are identified as true and mistakenly thought to be

shared, can often be a basis for the legitimization of various social practices, collective actions, and subcultures (O'Gorman, 1986).

Deborah A. Prentice and Dale T. Miller (1996) further draw the connection between PI and the perpetuation of social norms by unwitting actors as well as suggest a social mechanism on why they might do just that. According to them, PI arises when widespread behavioral adherence to social norms causes a discrepancy between public actions and private sentiments. Those norms can be either implicit or explicit. Implicit norms are general and regard appropriate social behavior, whereas explicit norms are usually concerned with the conduct of members of a particular group. Prentice and Miller suggest that no matter its content, what is needed to create PI is that the norm is powerful enough to induce people to act in ways that do not correspond to their private thoughts and feelings. What does it mean that the norm is powerful enough? Prentice and Miller offer two possible explanations for an individual to disregard their private attitudes in favor of an imagined public opinion. The first one is the fear of embarrassment. The other one is the desire to be a good member of the group. They believe that the latter is the root reason for most cases of PI (Prentice & Miller, 1996).

Prentice and Miller (1996) point out that the explanation of PI by fear of embarrassment may have something to do with methodological limitations. Fear of enablement is easier to consistently induce in people and, therefore, easier to capture in a laboratory as opposed to inducing the desire to belong to the group by the study participants who barely know one another. This fact together with the approach that the best way to understand something is to reliably reproduce it in the laboratory, was a reason why many social psychologists focused on the fear of embarrassment-based explanation of PI. It is not completely impossible to induce some type of group identification within the participants in the study, but, unlike fear or embarrassment, the desire to be a good member of the group is not just a response to stimulus induced by a specific social situation but develops and persists over time.

Prentice and Miller (1996) took, therefore, a different approach than their peers and investigated the phenomenon of PI in the real world on already existing groups. Similarly to Allport and Katz, Prentice and Miller surveyed the population of campus students and asked

them about alcohol use on campus. They have found evidence of PI being present in the group understood as undergraduate students at Princeton University. Students rate themselves as less comfortable consuming alcohol than they believe the average student is. Moreover, it is clear to everyone that drinking too much alcohol is not good for a person. Therefore, it may be quite difficult to adjust one's attitudes to match with the perceived public opinion. Nonetheless, the study shows that faced with the dissonance between their private beliefs and perceived yet false perception of public opinion, male students are driven to adjust their attitudes towards drinking, while at the same time, female students tend to alienate themselves and display lower levels of overall comfort. This suggests gender differences in the effect of PI on people (Prentice & Miller, 1996).

The final historical development of academic investigation of PI is the change from exclusively seeing it as a phenomenon of erroneous perception of social norms or public attitudes and shifting the attention to beliefs. Following Jens Christian Bjerring, Jens Ulrik Hansen, and Nikolaj Jang Lee Linding Pedersen, I will characterize PI in terms of the discrepancy between people's private beliefs and publicly reported beliefs (Bjerring & et al., 2014) unless specifically stated otherwise. As we have seen before, other attitudes can and have been used to characterize PI (usually norms and public opinion), but it is not hard to switch from the language of social norms to the language of belief. Moreover, I think that conclusions awarded through using the belief framework, are easily translated into the language of social norms. In my investigation, nothing substantial will be lost due to this move. Talking about beliefs rather than norms, which I welcome, is that it shifts the focus from the behavior of people in a group to testimonies that they report to others and makes it an explicitly epistemic phenomenon.

In this work, I will be referring to beliefs about other people's beliefs as meta-beliefs. The apparent epistemic error in individuals in the state of PI is not about the propositional content of their beliefs, but rather those meta-beliefs. That is, in the case of college drinking, the students would still be in a state of PI even if they believe that excessive drinking is healthy, as long as their meta-beliefs were incorrect. This marks a distinctive problem of social epistemology, where the object of inquiry about knowledge is concerned with the content of other people's minds.

The investigation of PI as a genuine social and epistemic phenomenon can, therefore, involve answering questions like “Why do people report their private insincerely in the public context?”, “What is the best way to form beliefs about other people's beliefs?”, “How can we change erroneous meta-beliefs within group members?”.

From this historical overview, I highlight the following conceptual points. Firstly, PI has something to do with individuals in the group forming erroneous assumptions about other people’s privately held beliefs. Secondly, the majority of people in a group form those assumptions and therefore it is important to investigate how the social context influences people to do so. Thirdly, people in the state of PI not only form erroneous beliefs about others' attitudes but also act as if they support those attitudes. Fourthly, they do so, because of the desire to be a good group member. Finally, it is possible to investigate this phenomenon focusing on one specific form of attitude - belief, making PI a phenomenon about privately held beliefs and public testimonies. Those considerations will be my guideline in the next section to give a formal definition of PI.

1.2 Pluralistic Ignorance: Formal Definition

Imagine you are a member of a group that centers around watching Quentin Tarantino movies. The director is well-estimated and known for the use of the top violent imagery in his movies. You are naturally squeamish, and therefore do not enjoy his movies very much, but think that the people in a group are cool and you would love to continue to be a member of this group. The first step in this situation for it to be PI is that not only you, but all (or basically all) members of the group feel the same: they do not think that Tarantino’s movies are fun to watch. However, that is not enough. If all of the members simply just hated Tarantino movies but would like to continue to meet, they would at some point speak up and discover that maybe they should change what movies they see together during their meetings. What is stopping them from doing just that, is the fact that they all believe that everyone else does think that Quentin Tarantino’s movies are fun. Because of that, they act contrary to their privately held beliefs. If they did not do that, everyone would quickly learn that their meta-beliefs about other group members were wrong, and PI would be unable to persist. The last point characterizing the situation in the Tarantino movies club is that its members form meta-beliefs based on the actions of the other members. They see people coming to their meeting, sharing their testimony “Django Unchained was fantastic!”, etc. and they get convinced that

those actions show support for the proposition that Tarantino's movies are fun. What it means, is that there should be some evidence (be it in other's behavior or their testimonies) that leads to the formation of erroneous meta-beliefs. Two last points make sure that the social character of PI is considered, just like Hubert J. O'Gorman suggested. PI is not simply a state in which, because of some mysterious psychological mechanism, it simply happens so that all group members hold false beliefs about other group members' beliefs. Jens Christian Bjerring together with Jens Ulrik Hansen and Nikolaj Jang Lee Lindling Pedersen has defined PI (taking into account all of those considerations) in the following way:

Pluralistic Ignorance - formal definition

The members of the group that is in a state of PI:

1. all privately believe some proposition ϕ;

2. all believe that everyone else believes ¬ ϕ;

3. all act contrary to their private belief that ϕ (i.e., act as if they believe ¬ ϕ); and where,

4. all take the actions of the others as strong evidence for their private beliefs about ϕ (Bjerring, et al., 2014, p. 2458)

It is important to add here that I understand the word "action" in point 4. in a broad way and includes sharing your testimony, facial expression, and unconscious movement. There are many ways based on which we form beliefs about other people's beliefs.

Allport writes: "The relation of audience and speaker is in itself a complex phenomenon. The individuals respond to the direct stimulation of the spoken sounds. Meanwhile, the overt components of their responses are serving as contributory stimuli to one another enhancing the effect of the speaker's words" (Allport, 1924, p. 303). This passage suggests that there is an interplay between what a speaker says, how they act, and how a listener responds to that.

Erik Funkhouser (2017) suggested that belief can be understood as a signal. Signaling is a type of communication, which involves information transfer that results in a change of receiver's behavior. In my model of PI, that is exactly what we assume is going on. Agents are verbally communicating their beliefs and the receivers form meta-beliefs about it. Those

meta-beliefs about other agents play a role in influencing how an agent will interact with others in the future. However, PI only happens if at least some of the agents communicate their beliefs insincerely. It is not easy to lie convincingly. A receiver of a signal may be therefore made aware of the lie, if a sender slips inconsistencies into it, verbal or non-verbal communication. In this work, I assume that a receiver of a signal can detect with a certain probability when another agent is sharing their belief insincerely (Balbuzanov, 2019). The stronger one believes in the proposition, one publicly denies the easier it will be to discover the lie. This framework allows observers to figure out if someone is not sincerely expressing their beliefs. This allows for the formation of correct meta-beliefs in the observer even when the testimony shared by a sender was insincere.

Here is the updated version of Bjerring's et al. definition of that takes into account understanding beliefs as signals.

Pluralistic Ignorance - formal definition update

The members of the group that is in a state of pluralistic ignorance.

1. all privately believe some proposition ϕ;

2. all believe that everyone else believes ¬ ϕ;

3. all act contrary to their private belief that ϕ (i.e., act as if they believe ¬ ϕ); and where,

4. all receive a signal from the other agents that indicates their support of ϕ and take it as strong evidence of other's private beliefs about ϕ.

To further sharpen our understanding of PI, in the following subsections I will mention different phenomena similar to PI, yet crucially different. Examining those differences will paint a more nuanced picture of what PI is and what it is not.

1.2.1 Bystander Effect

The bystander effect is a situation when people fail to initiate help to a person in need, fearing physical harm, embarrassment, involvement with law enforcement, and other unknown dangers. It was first discovered and academically considered by John M. Darley and Bibb Latane in 1969 in connection with the murder of Kitty Genovese, where allegedly 38 people witnessed the attack and none of them interfered. Darley's and Latane's study indicates that

the reason why when none of the crowd members intervened in the emergency is not due to apathy or their personality traits, but rather because of the emergent group dynamics (Latane & Darley, 1969). When people do not communicate, they may think that someone is already on their way to help or fail to see themselves as responsible for helping due to the diffusion of blame among many individuals.

Latane and Darley (1970) think there is a connection between bystander effect and PI. According to them, bystanders at first fear the embarrassment of overreacting, and therefore opt to remain cool and collected. But since everyone else in the crowd does the same, no one takes the emergency seriously.

I do not think that the bystander effect is the case of PI for several reasons. First of all, as Darley and Latane suggest, the effect only happens when people do not communicate with one another and are unclear about what behavior the others have already deployed ("Maybe someone has already initiated the help!"). However, inferring people's beliefs based on other's actions and testimony is crucial for the definition of PI. Moreover, it seems that bystanders start to question their private attitude about ϕ. They may think that they are overreacting or overthinking the situation, rather than that everyone else is mistaken. Therefore, one person publicly intervening usually dispels the bystander effect, while the same cannot be said for the cases of PI.

1.2.2 Minority Confirmation

There are situations when people who hold a certain view ϕ, are right about public opinion. The majority of the people can act as if they supported the proposition ϕ because they sincerely do so. When members of society perceive that they are the minority on a specific issue, they are less likely to state their sincere beliefs due to the fear of isolation (Chang & Kim, 2019). For example, imagine a situation where you enter a discussion where everyone expresses their enthusiasm towards genetically modified food. Based on that, you form a belief that all of those people think that GM food is safe, but you think it is harmful. If you do not speak up then with your sincerely held beliefs, it is not the case PI, as long as your meta-beliefs about other members are correct. Minority confirmation happens when a minority has a correct perception that they hold different beliefs than the majority, and that is the reason why they do not share their beliefs sincerely (or at all). They may be in a very psychologically

similar state as people in PI, but this is not the case of PI, since the minority hold a correct belief about meta-beliefs.

1.2.3 Social Learning / Information Cascade

Social learning can be defined as "a change in understanding that goes beyond the individual to become situated within wider social units or communities of practice through social interactions between actors within social networks." (Reed et al., 2010, p.6) When people interact with one another, and consequently public opinion is changing, that is the case of social learning. Let us now imagine that you hold a belief that ϕ. Then, you start to interact with society, and everyone tells you that they believe that ¬ϕ is true. If you in the face of that evidence (other people's testimony) change your belief from ϕ to ¬ϕ and start behaving in a way that shows support towards ¬ϕ, it is not a case of PI, but social learning. In this situation, you have changed your beliefs so they conform to the perceived beliefs of other members of the group, so you no longer act in public contrary to your private beliefs (which would be the key feature of PI).

1.2.3 Asymmetric Loss

In his talk about pluralistic ignorance, Kevin Zollman mentioned one more phenomenon similar to pluralistic ignorance. He called it "asymmetric loss" and described it as a situation when people hold a certain private belief but refrain from sharing it in public because they do not think there is a point in doing so. They may even believe that their view is shared by the majority, but due to social tact, they do not express that belief in public (CenterforPhilosophyofScience, 2021). For example, some members of the group may agree that the spouse of one of the group members is boring, but they do not feel the need to mention it. This is not PI, because this self-censoring is not motivated by an incorrect belief about other people's beliefs based on their behavior.

1.3 Rationality of Pluralistic Ignorance

There are a few puzzles from a psychological standpoint when we think about why an agent might find herself playing a part in PI. Why does she act contrary to her private beliefs? Even if we accept the answer that she acts on the desire to be a good member of the group, we can still ask why she does not realize that since she is capable of acting contrary to her private

beliefs, her peers may do the same. This led some philosophers to conclude that being in the state of PI is irrational, i. e. people in the state of PI commit some cognitive errors.

The rationality of PI is important to me because only if agents are rational, it is possible to make use of the tools of formal epistemology, e.g. game theory and Bayesian belief updating (Bjerring & et al., 2014). Since the aim of my thesis is to study the phenomenon of PI using computer simulation through agent-based modeling, the negative answer to the question about the rationality of agents in the state of PI may threaten the basic assumptions of my research.

In this section, I will present the view of Bjerring et al. who believe that agents in the state of PI are rational, and that of Daniel Grosz who denies it. I will then offer my answer to Grosz, to further support Bjerring & et al.'s conclusion and protect PI from the charge of irrationality.

Bjerring & et al. (2014) support the view that agents in the state of PI are rational. They make use of the formal definition of PI that I have presented in the previous section and try to investigate if any assumptions or combination of assumptions must necessarily involve irrationality. There is no problem with the first statement of the definition. Simply believing some proposition ϕ does not result in any inconsistencies. What needs a deeper analysis is the relation between assumptions 2. and assumptions 3. and 4. Only if basing belief about other people's beliefs on their actions is rational, and only if they act as if they believed ¬ ϕ, it is rational to believe that they indeed believe ¬ ϕ. Bjerring & et al. ask if it is rational to act contrary to one's belief. Therefore, to argue that PI can arise in groups of rational agents, we need only to show that it is possible for assumptions 3. and 4. to be true and rational.

First, it will be beneficial to understand what is meant by rationality. One way to think about it is in terms of structural rationality which is concerned with the relationship within a set of propositions that a person might hold. Often it is assumed that structural irrationality is concerned not with the substance of propositions but rather their form, for example, a set of attitudes can be maintained without any contradictions. Given that definition, there is nothing in Bjerring's definition of PI that inherently requires an agent to hold mutually exclusive beliefs. Therefore, it seems that agents in a state of PI can at least be structurally rational.

Bjerring & et al (2014) all use a different definition of rationality. They believe that rationality has two normative dimensions: pragmatic (concerned with the goodness of the consequences) and epistemic (promoting truth). It is notoriously hard to find a way to combine the reasons, or different types of values to get one clear directive "all things considered". To argue, however, that it is *possible* for PI to arise in the group of rational agents, we need to simply make clear how it could happen in just one set of circumstances.

Let us look deeply at the parts of Bjerring & et al. definition that we have under question:

"3. all act contrary to their private belief that ϕ (i.e., act as if they believe ¬ ϕ); and where,

4. all take the actions of the others as strong evidence for their private beliefs about ϕ (Bjerring, et al., 2014, p. 2458)".

Starting with the 3. point we need to show that it can be rational to act contrary to one's private belief. Since doing so does not promote the truth, to be rational all things considered, the pragmatic advantages of comforting to the perceived beliefs of others should outweigh the advantages gained from epistemic considerations. Since conforming to public opinion can fulfill a person's desire to be a good member of the group and acting contrary to it risks rejection of the group and other negative social consequences, it is clear to me that in many cases the pragmatic gains outweigh the epistemic ones.

When it comes to the rationality of basing the evidence on the actions of others, we need to consider that there are two types of evidence that agents can have about the proposition ϕ. The first is observational evidence - a behavior and testimony of other agents. The second one is introspective evidence, that is the knowledge that the person herself acts contrary to her own belief ϕ. Bjerring & et al. think that in many instances of PI, it is rational to give more weight to the observational evidence, because of the lack of any other observational evidence that would explain why other agents act as they do. There are, therefore, often good epistemic reasons to give more weight to observational evidence, rather than to rely on introspective evidence.

Given that sometimes at least there are good pragmatic reasons to act as defined in point 3. of the definitions, as well as good epistemic reasons to act as given in point 4. of the

definition, it seems that it is possible that in certain cases PI can arise in a group of all-things-considered rational agents.

On the other hand, Grosz (2020) argues that PI is, in fact, an irrational phenomenon. His problem is concerned with point 4. of Bjerring & et al.'s definition, and therefore mostly argues against the *epistemic* rationality of agents in the state of PI. He thinks that introspective evidence serves as an undercutting defeater of observational evidence. An undercutting defeater is "reason [ ] to question whether your evidence or reasons or grounds for a belief actually indicate that the belief is true" (Bergmann, 2006, p. 159). Grosz, therefore, suggests the introspective evidence that you believe ϕ and act as if ¬ϕ, undercuts the belief that others in fact do believe that ϕ when they act as if ¬ϕ. The claim is that you have an insight into the reasons that people distort their public statements. Since you are afraid to speak your truth, others may also be afraid to be sincere. Therefore, you should at the minimum remain ignorant about the meta-beliefs of others. I, however, think this demand is too strong.

Relying on introspective evidence to infer the content of other people's minds is giving too much weight to this type of evidence. The best indicator we have about the content of other people's minds is what they say and how they act. Assuming that others' psychology closely resembles our own is often problematic. We often lack empathy to understand that people think differently than we do. It is not irrational to epistemically trust someone's testimony when they were in general reliable and truthful. Doubting them, simply because our personal experience is different, may even be morally unjustifiable. Immanuel Kant writes about the importance of trust in treating a person like a means in itself, instead of a means to an end. According to Kant "it is true and prudent that I never again trust someone who has once cheated me, for he is corrupt in his principles. But to trust no *other* human being because *one* has cheated me is misanthropy" (Kant, 2007, 7:205).

My reformulation of Bjerrings & et al. is even more supportive of the all-things-considered rationality of agents in the state of PI. An individual receives a signal from another agent that indicates their support for ϕ. That signal includes agents' actions, and testimony, but also a background context for those behaviors. PI seems to appear when people act in public contrary to their private attitudes when the social context seems to support the sincerity of those publicly expressed attitudes. In this thesis, I will therefore operate under the assumption that PI can arise in a group of all-things-considered rational agents.

1.4 Dispelling Pluralistic Ignorance

PI is usually an unwelcome social phenomenon, producing many undesirable consequences, for example by slowing down the process of changing social norms not publicly supported by the majority (O'Gorman, 1975), or raising a level of discomfort in people who struggle to adjust their private beliefs to what they think public expect of them (Prentice & Miller, 1996). Therefore, it is important to learn more about how we can dispel it. In the following subsections, I will explore certain ideas on how to do just that.

1.4.1 Public Service Announcements

Public Announcements are considered an important part of most communication campaigns. There are a variety of approaches that such a social advertisement can employ, often trying to evoke emotions (Dillard & Peck, 2000). The goal is to persuade the audience and challenge their attitudes or beliefs.

Prentice and Miller's study suggested a way to dispel PI. Their analysis indicates that the best strategy to influence people's drinking habits was to make students aware that their perception of other students' attitudes towards drinking is wrong, as well as to explain that other students also tend to erroneously judge the general attitude towards drinking within their social group. It seems like public service announcements that explain the phenomenon of PI together with information about other people's sincere beliefs and meta-beliefs can help to dispel it (Prentice & Miller, 1996).

1.4.2 Social Learning

Even though social learning is not the case of PI, as I argued in section 1.2.3 of this chapter, it can still play a role in dispelling PI. When people change their opinions about their private beliefs based on other people's actions and testimonies, that is the case of social learning. In the case of PI, it can happen that individuals will take other people's actions and behavior as evidence that their own private beliefs are wrong. When this happens, they will update their beliefs so that they match the perceived beliefs of the majority. If most members of the group do that, it dispels PI, because all of a sudden, people's perception of the majority beliefs becomes correct.

Again, Prentice and Miller's (1996) study found some evidence that it is what naturally happens with time with male students on campus. In the first interview when male students

had to report how comfortable they were with drinking, they rated themselves significantly below the average Princeton undergraduate. However, as time passed, and they were invited to the second interview, that difference was no longer there. Male students matched their private attitudes with the perceived attitudes of the public, which consequently brought them a higher level of overall comfort over time. Interestingly enough, female students did not show this tendency and remained in the state of PI through both interviews, consequently showing a lower overall comfort over time.

In a situation when a majority of the population privately upholds some unethical social norm or holds a false belief, social learning may be a beneficial way to dispel PI. Since at the same time, individuals in society all act as if they did not support that belief or action, it makes it possible that due to social learning, they will all also privately reject the unethical social norm or update their belief to be closer to the truth. Here, social learning seems to be a corrective mechanism to people's attitudes.

However, in more typical cases of PI, it seems that what we want is for people to understand that they are in the state of PI. We want them to realize that their perception of the general attitudes of people is incorrect. It can be so because the social norms held in private may be more progressive than those supported in public. Or the beliefs that are held in private are true, while those that seem to be supported in public are not. Social learning would then be detrimental to the change for the better in social behavior.

When we think about the problems associated with social learning and PI, it is also important to consider Prentice and Miller's discovery about gender differences in the reaction to being in the state of PI. Social learning can bring comfort, dealing with the cognitive dissonance of publicly supporting different beliefs than those supported in private. However, this solution does not seem to be available to most women. All of those considerations lead me to a conclusion, that social learning is not a preferable way to dispel PI.

1.4.3 Tight vs Lose Society Ties

People in a society can be closer or farther together, for example, people may share their beliefs only with a limited number of people (close friends and family), or act less restrictive with the information they share. Anthropologists have classified societies into "tight" and "loose". However, there is no clear definition of what exactly it means to be one or another

(Pelto, 1968). The differences between the two are expressed by various criteria: how social norms are expressed, is deviant behavior easily tolerated, what are the values of the members, what are their personality traits, and how does economic and social life influence group members? In this thesis, I will only make a distinction between societies with a denser connectivity within social networks (tight societies) together with the continuum to more loosely connected structures (loose societies). Using my model, it can be explored if the number of connections between individuals influences the formation of PI and its potential to get rid of PI.

1.4.4 Contrarians and Sincerians

Contrarians are agents who adopt the choice opposite of the prevalent choice of others whatever that choice may be. In the context of PI, those would be the people who after forming an assumption of majority beliefs, they keep saying the opposite of that, no matter what their privately held beliefs are. It has been shown that contrarians can have a positive effect on group dynamics preventing group polarization (Galam, 2004). Those unique features of contrarians make it interesting to check what effect they may have on the formation of PI and its dissipation.

Another type of person that may have a similar effect on group dynamics as contrarians in the case of PI are sincerians. In the case of PI, sincerians are those agents who state their sincere beliefs no matter what their perception of the beliefs of others is. PI could not emerge in the population with agents acting in this way, because people would always be properly informed about the beliefs of others, It is, therefore, interesting to investigate what effect sincerians' behavior may have on dispelling PI.

2 Agent-Based Model of Pluralistic Ignorance

The goal of this thesis is to develop a model of pluralistic ignorance (PI), which then can be implemented in an agent-based computer simulation. In this chapter, I will explain agent-based modeling and why I chose this methodology. I will also discuss the limitations of those techniques. Afterward, I will explain each part of my model used to simulate the behavior of agents in society and discuss the results of two experiments done using the model that checked two ways of dispelling PI.

2.1 Models in Science

2.1.1 What Are Models?

Models are ubiquitous and useful scientific tools. Some may even claim that creating a model is the main activity of the scientist. John von Neumann claims that "sciences do not try to explain, they hardly even try to interpret, they mainly make models." (von Neumann, 1995, p. 628). That seems to paint a rather bleak picture of what science is able to achieve, though it is hard to deny that there is no shortage of scientific models of different forms within different contexts: the Bohr model of an atom, the AD-AS model in economics, the DNA double helix model, etc. Some of the models make use of complex mathematical formulas, some are a picture together with theoretical explanations. It is notoriously hard to come up with a definition that explains what all of those things have in common. John von Neumann gives the following definition of a model, which I think is close to capturing the essential attributes of a scientific model: "By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that is expected to work - that is, correctly to describe phenomena from a reasonably wide area." (Gelfert, 2016, p. 628). From that definition we can draw the following conclusions: models use mathematics to say something about the phenomenon that can be observed in the external world, and they do not do it with a 1:1 precision, but rather to some not properly defined degree. But what about the models that do not rely on mathematical formulations? Scale models increase or decrease some features of the target system, to make them easier to handle and understand, while analogous model in addition, involves a change of medium (for example, using a water flow as a stand-in for the flow of money in the Philips machine) (Gelfert, 2016, p. 3).

It may be that there is no unified principle that can explain the nature of all models. What those different examples of models may have in common, is not their essential characteristics, but their function.

2.1.2 What Is the Function of Models?

It may be easier to define models by their function rather than their nature, but it is still not obvious that all models share the same function. One thing seems to be clear – they all help us to cognize the world, but they may use different means to achieve this goal.

The three functions of models that are easiest to identify are:

1. drawing inferences on their basis about the target system
2. making predictions
3. making it easier to develop new courses of action (Gelfert, 2016, p.26)

For the first two functions, it is necessary that a model represents reality in some way. What exactly that means is controversial. There may be different standards of what counts as a proper representation, and the appropriate levels of precision. The main idea, however, is that by learning and discovering some phenomenon emerging from a model, we may learn something about the target system. Moreover, models that stand in for a real object/phenomenon in the external world can generate predictions that can later be empirically tested. In that way, models supply us with a way of testing the theory from which the model emerged.

However, models can also be exploratory. They may be created without referring to any specific theory, simply to make it easier to develop such, through the exploration of many different possibilities. It is also possible that we from the start know that a model is in some relevant way false. William C. Wimsatt (1987) identified 7 different ways in which a model can be false. A model may be:

1. only applicable locally, and false when more broadly applied.
2. idealized in a way that conditions it describes never happen in nature.
3. incomplete
4. misdescribing the interactions that are included in the model,
5. give a completely misguided picture of nature.

6. purely phenomenological - it only gives a description or prediction of phenomena without stating if the variables used in the model actually exist.
7. fail to describe or predict data correctly (this is consistent with all the other points on the list, but sometimes it is the only thing that is known about the falsity of the model).

Wimsatt (1987) then comes to claim that those false models can be useful among others because they can act as a starting point to further develop models of increasing complexity and realism, suggest new predictive tests to the established model, or it may capture larger effects that can later be factored out to detect phenomena that would otherwise be too small to be discovered. Phenomenological models can also suggest relationships between the variables that can help relate them in a mechanistic way.

2.2 Methodology: Agent-Based Modelling

To study the development and possible dispelling of PI, I have decided to make use of agent-based modeling (ABM). Those models have become a popular method of inquiry in formal social epistemology of science. (Frey & Šešelja, 2018) and, more generally, to represent complex dynamic systems (Damaceanu, 2008). ABMs are a type of computer simulation that enables the investigation of phenomena from the bottom up. It usually makes use of the following components:

1. Agents, which are individual, heterogeneous, and autonomous software units.
2. Rules of behavior that govern the actions and interactions of agents defined by the programmer.
3. Explicit spatial and temporal dimensions (very common, but technically not required). (Romanowska, Wren, & Crabtree, 2021, pp. 6-7)

These models afford researchers insights into how intricate behaviors can emerge from simple behavioral rules governing individual agents and their interactions (Frigg & Hartmann, 2020). The significance of emergence lies in the fact that the interactions among agents can yield phenomena not necessarily foreseen by the agents themselves, yet still pertinent to their subsequent behaviors and achievements (Conte & Gilbert, 1995). ABMs can be likened to computational thought experiments. The assumptions made during model creation and implementation can be seen as premises in an argument, while the results serve as conclusions drawn from these extended arguments (Pöyhönen, 2017). Consequently, the outcomes of my model are contingent upon the assumptions I have made. This approach

offers the advantage of potentially revealing more general patterns rather than results confined to specific contexts.

To implement my model, I used the computer program NetLogo, a commonly employed tool for creating ABMs. After creating the mathematical model of PI I will test my assumptions by applying the modeling assumptions and creating a computer simulation in NetLogo. Simulations are a good way to test the model and develop the feeling of how all the elements are related. “By simulating a model’s behaviour in time, one may get a feeling for how sensitive a model is to changes in the initial conditions; by varying other parameters, one may be able to get a sense of what kinds of real-world scenarios or phenomena a model can represent.” (Gelfert, 2016, p.73)

2.3 Limitations

There are many limitations when it comes to using agent-based models in social science or formal epistemology. Since it is impossible to model all of the complexities of human personalities and all the factors that influence the formation of, the model will not be predictive, but mainly exploratory. However, even when created without the ambition to predict the behaviors of people, their usefulness of highly idealized agent-based models in the area of formal epistemology has still been put in question. In the following subsections, I will take a look at the limitations concerned with idealization and also investigate the usefulness of exploratory models in formal social epistemology.

2.3.1 Idealization

Modeling always involves theoretical trade-offs. With the model of PI, I am not interested in giving accurate empirical predictions of agents’ behavior in society but rather in identifying the fundamental mechanisms that motivate the system under investigation (Gelfert, 2016, p. 65). It would be impractical and overly complex to attempt to model every conceivable parameter influencing the actions of an agent. In pursuit of understanding, idealized models intentionally simplify or distort certain intricate aspects of reality for the sake of clarity (Frigg & Hartmann, 2020). Often, this simplification is essential for grappling with systems too intricate to fully dissect (Frigg & Hartmann, 2020), particularly in the context of deciphering human behavior within society.

In crafting and implementing my model, I adopted an Aristotelian approach to idealization. This involved envisioning which societal properties were extraneous to my inquiry and omitting them when constructing a model based on reality. The objective of such idealizations is not to precisely replicate the behavior of real systems, but rather to elucidate truths about how those systems would behave if the disregarded parameters were absent in reality. This methodology allows for the examination of specific parameters in isolation (Frigg & Hartmann, 2020). Consequently, it enables us to discern whether these parameters contribute to observed behaviors in reality and, if so, in what manner. This approach is rooted in a counterfactual understanding of causation, facilitating the examination of whether the presence or absence of certain parameters correlates with particular behaviors.

2.3.2 Exploratory Models

Highly Idealized Agent-Based Models allow us to offer simple explanations of complex phenomena. It may give a false sense of understanding a problem at hand. It has been suggested that some of the findings using such models are not as robust as expected if we make changes in the relevant parameter space. Moreover, the results of some of those studies cannot be replicated if some of the underlying assumptions about models are changed. Most crucially, if the model lacks empirical calibrations, it is hard to take seriously the normative conclusions we may draw from it (Frey & Šešelja, 2018). Frey and Šešelja (2018) suggest that there is no epistemic value in merely showing what is possible rather than what it is actual, which is what exploratory models are for. If showing possibility is not enough to justify learning something from highly idealized formal models, is there any role they may have? According to Frey and Šešelja highly idealized agent-based models can be useful as a "method for providing evidence for historical and philosophical hypotheses concerning scientific inquiry" (Frey & Šešelja, 2018, pp. 407-408). They can provide novel hypotheses underlying particular historical episodes, they may force us to update our beliefs about normative aspects of scientific inquiry, they can confirm our methodological assumptions and help us judge different normative philosophical conceptions about science by showing the superiority of one view over the other.

I think Frey and Šešelja's view is too limiting. I believe that agent-based models, even when highly idealized, can have more to say than just confirming historical context. Here are some examples of what exploratory models can do:

(a) They demonstrate the possibility or impossibility of certain phenomena by simulating interactions among individual agents and observing emergent behavior.

(b) Agent-based models introduce new ideas or considerations into the debate by allowing researchers to explore complex systems from a bottom-up perspective, considering the behaviors and interactions of individual agents.

(c) These models examine and hence establish or challenge the validity, generality, or scope of earlier arguments by providing a more nuanced understanding of the dynamics within systems.

(d) They support or undermine earlier claims about difference-making by simulating scenarios and observing how changes in agent behavior or system parameters affect outcomes.

(e) Agent-based models modify earlier arguments to correct mistakes or enhance plausibility by incorporating more realistic representations of agent behavior or system dynamics.

(f) They provide additional arguments supporting either the premises or conclusions of earlier arguments by offering empirical evidence derived from simulation experiments.

(g) Agent-based models broaden the debate by introducing a new perspective on the problem in question, often highlighting the importance of individual-level behavior and interactions in shaping macro-level phenomena. (Aydinonat, Reijula & Ylikoski, 2021, p.384)

2.4 Agent-Based Model of Pluralistic Ignorance

In this section, I present the agent-based model of Pluralistic Ignorance, based on Kevin Zollman's theoretical model (CenterforPhilosophyofScience, 2021).

2.4.1 Networks

To understand who in society interacts with whom I need to choose a model of network structures. Understanding society as a network means that we treat agents as nodes and the relationship between agents as links. A link between two agents tells us that they can interact with one another.

The simplest way to model a social network is to use a completely random structure, where the parameter simply specifies the likelihood of forming a connection between two

nodes. That is the Erdos-Renyi network. The problem is that it is not very good at representing the likely situation in the real world.

A more realistic model of a network is a small-world model. The small-world model of a network combines the properties of regular structured lattices (these networks can be highly clustered), yet have small characteristic path lengths, like random graphs. In other words, in a small-world model, there is the coexistence of a local structure and random long-range connections (Watts & Strogatz, 1998). It corresponds to people having local clusters of people; they are close and interconnected, but each person is only 5-6 people away from anyone else (Telesford & et al., 2011).

In my program, one can choose which network to implement. Erdos-Renyi network, for example, can be used as a control, while a small-world network can serve as a more accurate representation of a social network.

2.4.2 Agents

Agents in the simulation represent individuals in society. They are individual software units, following a set of rules defined by me. Each of the agents at the start of the simulation receives variables that will influence their behavior. These are:

- P - the strength of the private belief that ϕ
- α - truth-seeking orientation
- U_1 - the value of the utility function to say that ϕ
- T - what binary testimony are they reporting
- S - what signal they are giving to others about their beliefs about ϕ
- N_1 - a proportion of others that an agent interacted with that signaled that ϕ

Let us now consider in depth what the meaning of each component is.

The strength of the private beliefs (P) indicates how strongly someone supports a certain proposition. In this model, agents will be only concerned about one proposition ϕ. If P is equal to 0 it means that an agent completely rejects ϕ. On the contrary, if P is equal to 1 it means that no evidence can sway an agent to stop believing that P, and an agent treats ϕ as if he knew that ϕ.

The truth-seeking orientation α (alpha) tells us how much an agent values the sincere exchange of beliefs. The higher it is, the more likely it is that an agent will share their beliefs with others sincerely. I call it truth-seeking orientation because if everyone in the society behaved in a completely truth-seeking way and shared their beliefs sincerely, the state of PI would be impossible to achieve. People would form their meta-beliefs based on sincere declarations, which would make their meta-beliefs true. Since their meta-beliefs are correct and they declare their beliefs truthfully, PI cannot arise.

The truth-seeking orientation is contrasted with the desire for conformity to the agent's peers (1-α). This indicates how much an agent's declaration of private beliefs is affected by what the others an agent interacted with have signaled to him through their testimony and behavior. If this value is high, an agent will simply say, what he believes the majority of people also believe. If on the other hand, this value is low, the agent will share their belief sincerely (it can just so happen that the majority of people believe the same).

The testimony T takes a binary value. It simply indicates if an agent is going to say that he believes that ϕ, or if he is going to say that he believes that not ϕ. Agents in this model are basing their testimony on their truth-seeking orientation, the desire for conformity, the actual strength of their belief that ϕ, and the signals from others about their strength of belief that ϕ.

N_1 is the proportion of others that an agent interacted with that signaled that ϕ, while N_2 is the proportion of others that an agent interacted with that signaled that ¬ϕ. In the original utility function variable $N_a(\phi)$ was interpreted as a proportion of the agent's public *declaring* that ϕ. What is usually meant by a public declaration is that an agent uttered a statement, endorsing it in some way. It communicates to the others that agent *a* believes that ϕ. We can then say that originally $N_a(\phi)$ signified what percentage of agents give a testimony that ϕ is true. The fact that an agent can be wrong about their beliefs is usually accepted and included in the models in social epistemology. However, for PI, what is most important is not the truth of ϕ, but rather the truth of belief that some other agent believes ϕ. As Bjerring (Bjerring, et al., 2014, p. 2458) suggests, agents in the state of PI take the actions of others as strong evidence for their private beliefs about ϕ. Those actions include testimony but should not be limited to just that.

Let us examine again the situation of college student drinking habits. Everyone wants to drink less, but the excessive drinking of others, makes them think that everyone likes drinking more than they actually do. Based on some signals from the environment, agents can figure out when others are lying. That could be because of their body language, tone of voice, picking up on other incentives that may motivate an insincere public declaration, or acting incongruent with the statement (e.g., saying that she believes in a Catholic God, but always skipping a Sunday mass). The higher the discrepancy between the strength of the private belief that ϕ and what an agent testified, the higher the probability there is that the insincerity is going to be detected. When the lie is detected, what is signaled to the receiver is the opposite of what was said by the sender. When the private beliefs match a sender's utterance, the signal received is equivalent to the testimony. With this interpretation of the meaning of $N_a(\phi)$, we can say that it indicates a proportion of people that an agent believes they believe ϕ.

In the model the option for a lie detector mechanism to be turned off. When that happens the signal is always equivalent to what a sender has testified:

$S_{ab}(\phi) = T_b(\phi)$,

where

$T_b(\phi)$ is a strength of their testimony ϕ, it is represented by a normal distribution centered around the agent's testimony, with a standard deviation value representing *b*'s uncertainty.

The next step after receiving the signal is updating beliefs about what the majority of one's social group believes. Another function needs to be created that marks who signaled their support for the proposition ϕ:

$N_{ab}(\phi) = \{1, if \mid S_{ab} - \phi \mid < \mid S_{ab} + \phi \mid \quad 0, if \mid S_{ab} - \phi \mid \geq \mid S_{ab} + \phi \mid$

We need this function to go from continuous values to binary values so that the signal either indicates b's support for ϕ or it did not.

And finally,

$N_a(\phi) = \sum_{b \in A \setminus \{a\}} \quad N_{ab}(\phi)$ / n

This function allows us to count how many people signaled their support for a proposition ϕ.

To determine what agents in our model are going to announce we need a function that governs how the desire for sincerity of an agent measures against their conformity to their peers. For this, I will follow Kevin Zollman (CenterforPhilosophyofScience, 2021) and adapt Aydin Mohseni and Cole Randal Williams (2021, p. 3) utility function:

$U_a(\phi) = \alpha_a P_a(\phi) + (1-\alpha_a) N_a(\phi)$,

This utility is then measured against the utility of the opposite statement:

$U_a(-\phi) = 1 - U_a(\phi)$

The agent will act so that she will maximize her utility. The part of the formula "$(1 - \alpha_a) N_a(\phi)$ " can be understood as a coordination game, where agents are trying to match their responses to one another (CenterforPhilosophyofScience, 2021). Since the desire for conformity and a truth-seeking orientation are complementary measures, this game does not, necessarily, influence the agent's expectations about the truth of ϕ (though of course, the way agents update $P_a(\phi)$ can, and often does, depend on the public announcements of others).

This equation offers an answer to why PI remains stable. If the agent's truth-seeking orientation is low and she believes that ϕ, but people around her keep declaring that -ϕ, she will start declaring -ϕ as well. And if everyone else in the group acts like this, they will be in a state of PI. That does not explain, however, how this strange situation came about. If initially, no one declared anything, the truth-seeking orientation would prompt people to declare their sincere beliefs, stopping PI on its track.

2.4.3 Dispelling Pluralistic Ignorance

The new model I present shows not only a more nuanced way of understanding why PI can be stable but also allows us to investigate ways in which a group can break from the state of PI. Our formulas describing agents' behavior show that PI remains stable when agents' desire for conformity is high compared to their truth-seeking orientation. The other element influencing the stability of PI is how well can agents hide their sincere beliefs from their peers. If the signal they are sending is in harmony with their testimony, other agents may never pick up on the fact that that testimony does not sincerely represent that agent's beliefs. Consequently, agents will continue to hold false beliefs about other agent's beliefs.

The question then remains how is it possible to break out of the PI? One way to do it is for some reason agents' truth-seeking orientations become higher than their desire for conformity. It can be motivated by being tired of pretending to believe something that one does not believe, wanting to influence a change in the world, etc. My model also allows people to realize that they are in the state of PI, by picking up on the signal given off by agents and in that way "breaking the spell". Since testimony is not the only way in which we form our beliefs about other people's beliefs, the model allows for a wider range of possibilities for the evolution of society in the state of PI. Even when public announcements continue to be insincere, agents can still be able to break out of the state of PI. As their beliefs become more and more in line with reality, even the agents with a strong desire for conformity will be compelled to share their sincere beliefs.

However, it may still be too small of an effect to dispel PI. I will present now how I modeled some other elements that can prevent from PI arising and/or dispelling it.

Contrarians

When a contrarian speaks, they want to say the opposite of what a majority seems to believe. Therefore, their utility functions look different than that of a regular agent:

$U(-\phi) = N_a(\phi)$

$U(\phi) = 1 - U(-\phi)$

A contrarian will say ϕ if the utility function of ϕ has a greater value than the utility function of ¬ϕ. Otherwise, he will say ¬ϕ.

Sincerians

Sincerians are the agents who always share their beliefs sincerely, no matter what they think the majority believes. Therefore, they can be understood as agents whose truth-seeking orientation is equal to one. Their utility functions look then like that:

$U_a(\phi) = P_a(\phi)$

$U_a(-\phi) = 1 - P_a(\phi)$

A sincerian will say ϕ, if the utility function of ϕ has a greater value than the utility function of ¬ϕ. Otherwise, they will say ¬ϕ.

Modeling Public Service Announcements

What could the effect of a campaign that demonstrates people's sincere beliefs about ϕ be? Since they would not be able to tell exactly which agent was sincere and which agent lied, we cannot affect the N_a component of the utility function. However, they seem to be willing to act and talk in a way that shows their sincere beliefs. That's why I think that the effect of Public Service Announcements is an increment in the value of α - their truth-seeking orientation.

Network

To check the network effect on the formation and dispelling of PI you run the simulation on different network shapes and with different probabilities of forming connections between agents. This latter case, allows us to vary population network density - the average relation between any two actors in the population (Burt, 1980, p.82)

2.5 Experiments on the Simulation

2.5.1 Experiment 1

I designed two experiments using the model of PI I have implemented in NetLogo. Both explore how to get from the society in the state of PI to a society where all agents share their beliefs sincerely. In the first experiment, I investigate what happens when we introduce an agent who is always sincere about her beliefs (sincerian) to a group of agents in the state of PI. A sincerian has an α = 1, meaning her truth-seeking orientation is high. The experiment starts with the insincere agents when their α = 0. In the course of the experiment, I will raise their α to see what effect it has on reaching the state where all agents share their beliefs sincerely. I will also check how the number of insincere agents in society affects how quickly (or at all) the PI gets resolved. For simplicity's sake, the lie detection mechanism is disabled.

Simulation set-up

The simulation starts with x-number of insincere agents. Their parameters look as follows:

- alpha (α); the value of truth-seeking orientation is not stated, because I will vary it in the course of the experiment starting from 0.
- set N 0; this number signifies the percentage of others that the agent interacts with and believes that they support the proposition, The insincere agent believes, therefore, that everyone around them supports “not ϕ”.
- P 1; this is an agent’s strength of belief in ϕ, a sincere agent privately believes that ϕ
- T 0; this is a binary variable (it can be either 0 or 1), a sincere agent at first says to the others that “not ϕ”. Notice that it is the opposite to what she sincerely believes.
- v 1; this is a tally of interactions with the other agents. It is equal to 1 here, to show that an agent experienced a minimal history of hearing others claiming that ϕ is untrue.
- I 0; this is just a technical variable necessary to update N, it raises whenever an agent hears another agent claiming that ϕ is true.
- The v, which is a tally of interactions is equal to 1, to show that there was a minimal history of hearing other people claiming that P is false. I am going to vary alpha in the course of the experiment, so its value is not given.

Alpha and P are constant throughout a run of simulation. N, T, v, and I are updated with each simulation tick. A tick represents a single step or increment of time within the simulation's execution. During one tick in this experiment, each agent shares their belief with another randomly chosen agent that is linked with her in the network. Agents then update their parameters according to the testimonies they have received.

The parameters for a sincere agent look as follows:

- alpha 1.0
- N 0
- P 1
- T 1
- v 0
- I 0

Since a sincere agent’s truth-seeking orientation is equal to 1, the value of N, v and I do not change the value of their utility function. That is because a sincere agent never takes into account the beliefs of people around them when she considers how to testify.

Testing conditions

I ran the simulation 50 times for each setup of parameters. If after 10000 ticks, all of the agents did not become sincere the run was marked as divergent, otherwise, the run was considered successful. When a run was successful, it meant that all agents shared publicly what they believed privately.

For each population size, I ran 50 simulations varying the alpha until I got 50 successful runs with the same parameters. I did it for the following conditions:

- 2 insincere agents, 1 sincerian
- 3 insincere agents, 1 sincerian
- 4 insincere agents, 1 sincerian
- 5 insincere agents, 1 sincerian

My starting hypothesis for this experiment was as follows:

1. The lower the alpha of insincere sheep, the slower or impossible it is that all 50 runs are successful.
2. The more insincere agents there are, the slower the transition is or impossible it is that all 50 runs are successful.

Results

When a simulation started with 2 insincere agents with alpha equal to 0 and 1 sincerian all 50 runs of the simulations ended up successful. When one more insincere agent was added to the network, the situation changed dramatically (Figure 1). With the alpha of insincere agents equal to 0, the percentage of divergent runs becomes much higher, reaching 82%. However, just a little raise of alpha resulted in a huge improvement in the percentage of successful runs. With alpha equal to 0.2 all 50 runs were successful.

This trend continued when adding more insincere agents to the network. When the simulation started with 4 insincere agents and 1 sincerian, the number of divergent runs for alpha = 0, was 92 % (Figure 2). The improvement in a number of successes was raised with the higher alpha, reaching 100 % for alpha = 0.3.

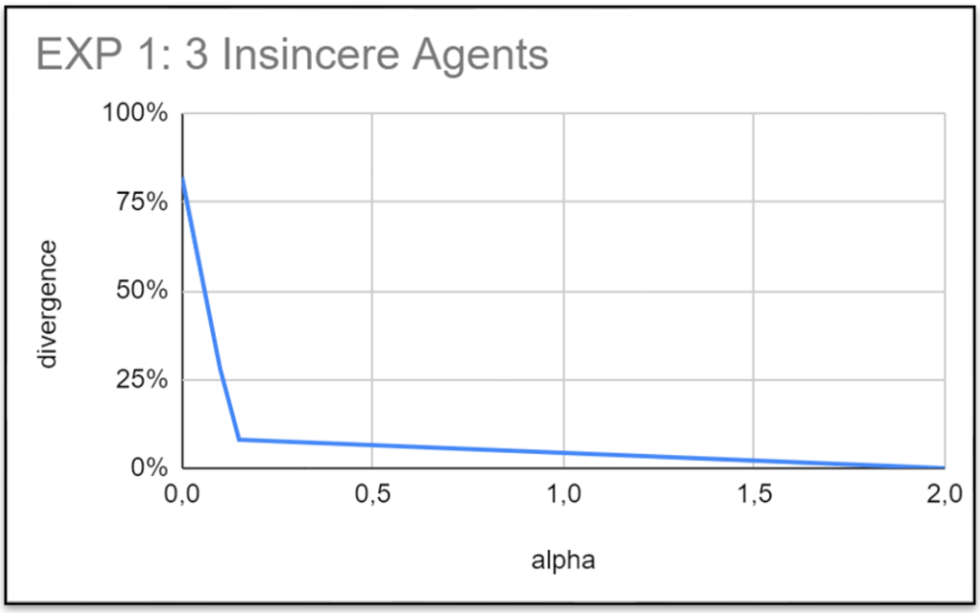

Figure 2 EXP 1: 3 Insincere Agents

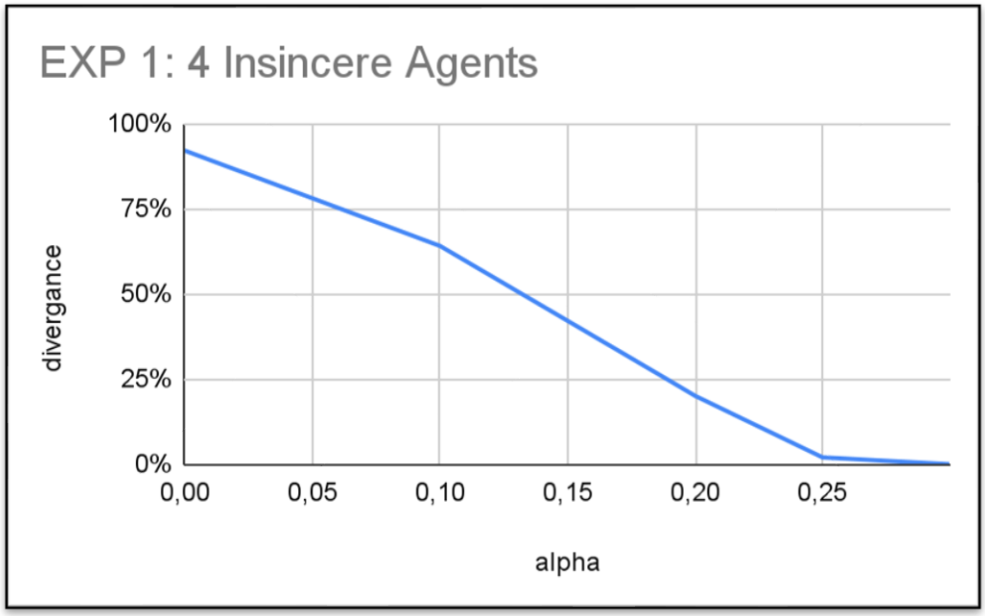

Figure 1 EXP 1: 4 Insincere Agents

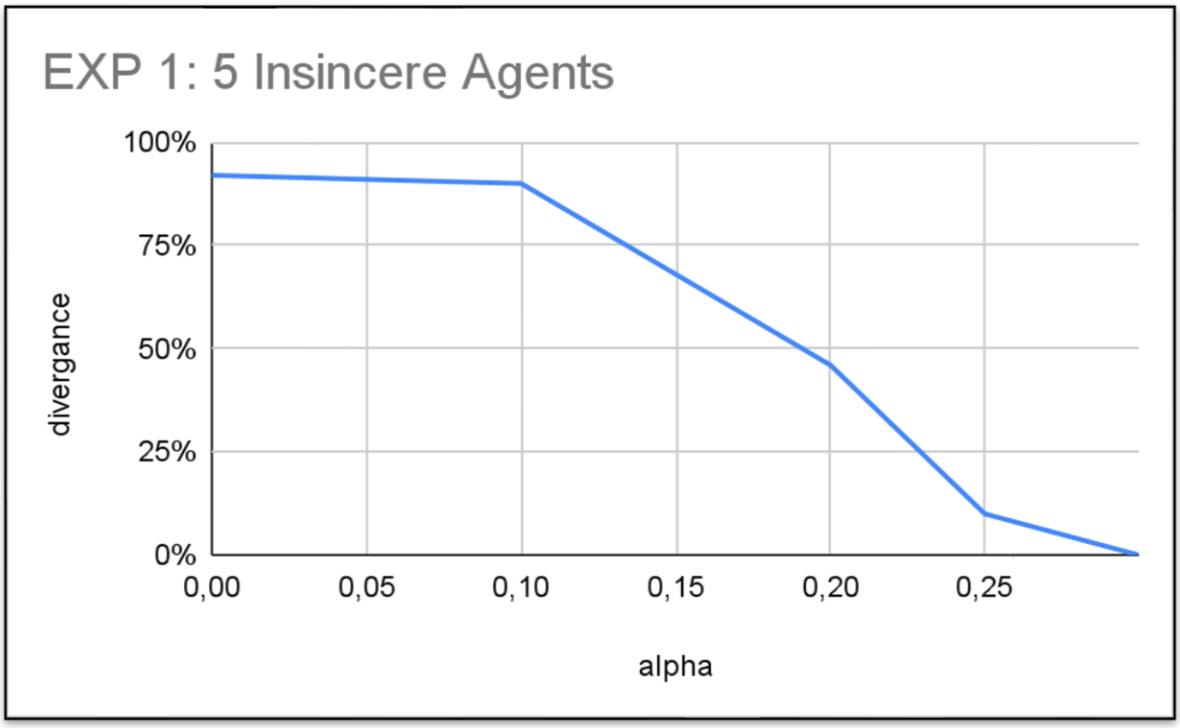

Figure 3 EXP 1: 5 Insincere Agents

The behavior of the agent was very similar in a condition with 5 insincere agents (Figure 3). Just like for a condition with 4 insincere agents, when alpha was equal to 0, the divergence rate was 92% and only when alpha was equal to 0.3 did, I get 100% successful runs. The difference between this condition and the condition with 4 insincere agents was that with in-between values of alpha, the society with 5 insincere agents performed worse than the society with 4 insincere agents.

In this next graph (Figure 4), we can see the dependence of the percentage of successful runs when the alpha of insincere agents is equal to 0, and the ratio of sincere to insincere agents. We can quickly notice that the rate of successful runs of the simulation drops significantly when we have more than two agents per sincerian.

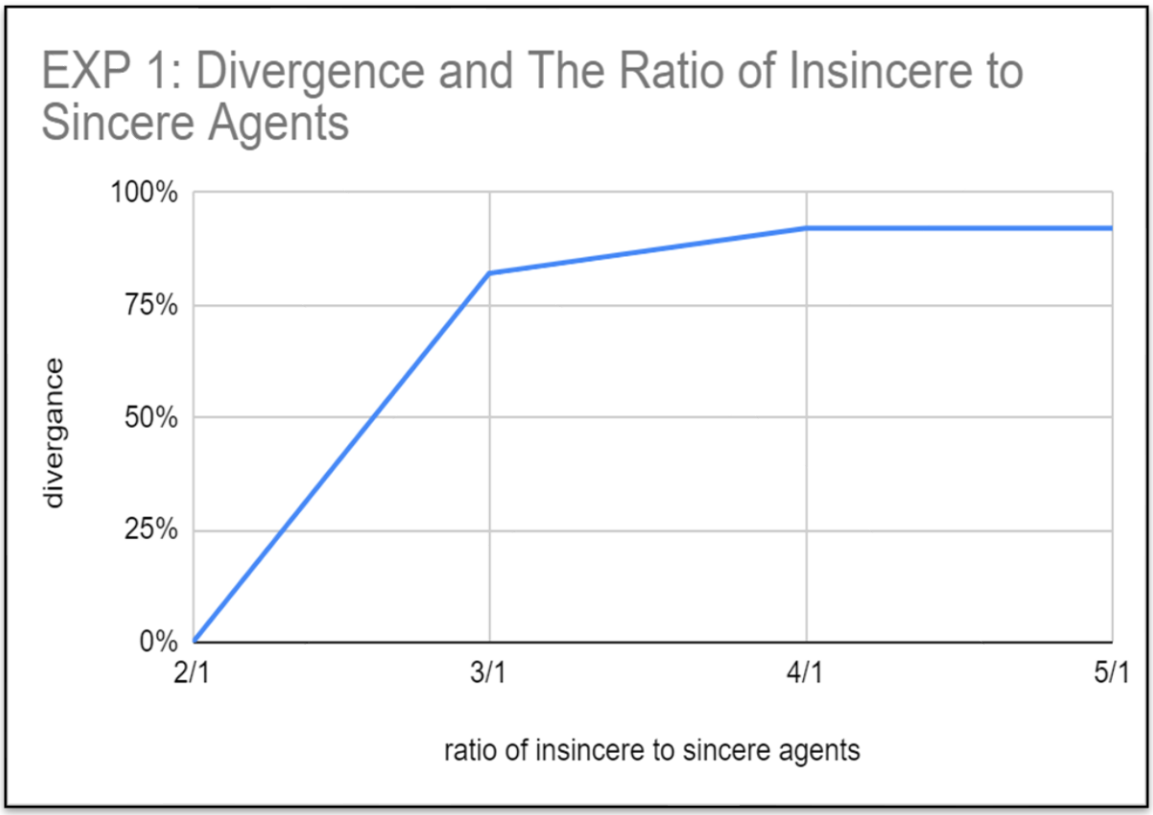

Figure 4 EXP 1: Divergence and The Ratio of Insincere to Sincere Agents

So far, I have only talked about the divergence percentage during 50 runs of the simulation for each condition. However, we can also see the amount of ticks it took for all agents to share their private beliefs sincerely. Interestingly, it seems that very often a run, if it becomes successful, it does so after very few ticks. In the graph below (Figure 5) I collected the data for all the successful runs of simulation from all setups of parameters. It shows how big the percentage of successful runs terminated at the given number of ticks.

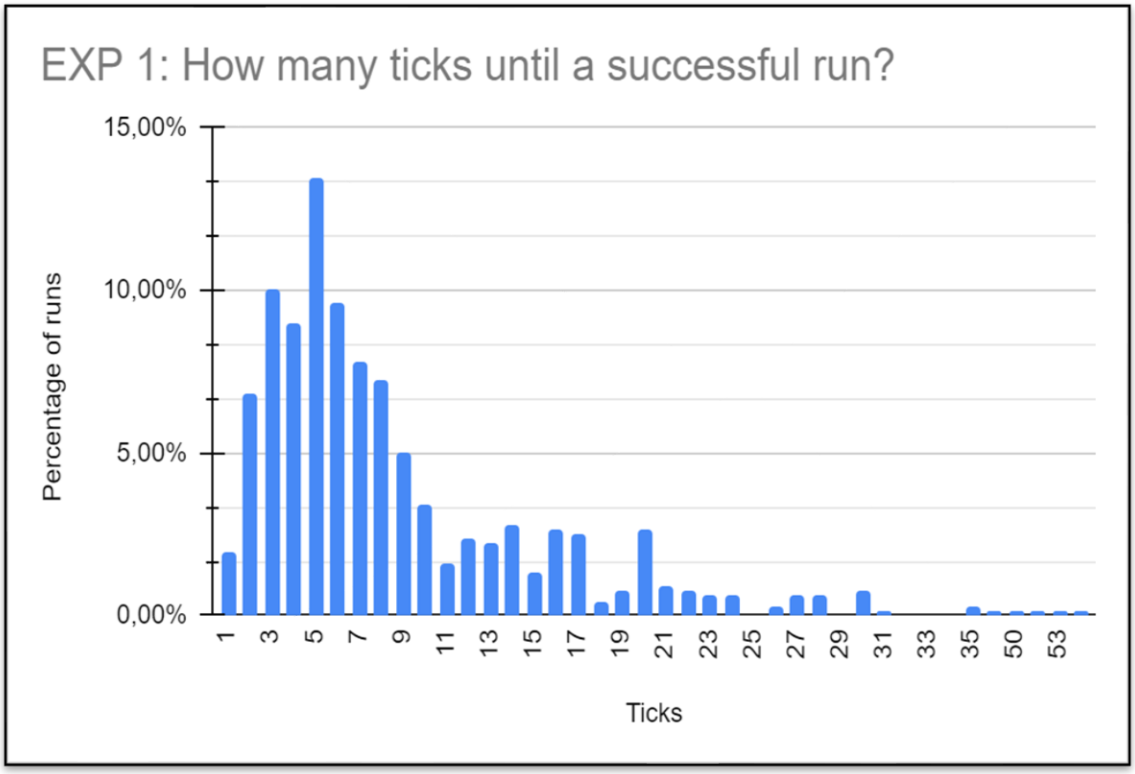

Figure 5 EXP 1: How many ticks until a successful run?

As we can see the stability is most often reached at the 5th tick of a simulation, and usually happens by the 10th mark. The 5th tick looks like a point of optimum balance where enough interaction has happened for agents to update their parameters in such a way as to sincerely share their beliefs. The simulation is constructed in such a way that it is random which two agents happen to interact with one another. It seems to be important that the very first interaction that happens involves a sincerian being listened to by at least one insincere agent. Such interactions given our parameters when alpha is equal to 0, give an insincere agent a 50:50 chance to start reporting her belief sincerely. This is because, after the interaction with a sincere agent, the utility to say that ¬ϕ is equal to the utility to say that ϕ. I coded the program in such a way, that in those circumstances it is a 50:50 chance of an agent reporting their beliefs sincerely. If an insincere agent starts to repeat sincerely, it can produce a chain reaction that affects other insincere agents. This is confirmed by visual tracking of the unfolding of the simulation. Whenever one insincere agent flips and starts to report sincerely, the rest follows. There was never a divergent run during which only some agents started to report sincerely.

Discussion

The assumptions made in our hypothesis were partially confirmed by the simulation. The rate of divergence of the simulations is affected both by the parameter alpha of insincere agents, as well as the number of insincere agents per sincerian. The higher the alpha, the more likely it is that a run will be successful. The more insincere agents per sincerian, the more likely it is that the run will be divergent. However, when the runs are successful it does not appear to be a significant difference in how many ticks it takes for it to happen between different conditions. It seems that in general if stability arises, it arises quite early during simulation no matter the parameters. It is also confirmed by visual observation of the unfolding of simulation runs that agents usually switch to correctly report their private beliefs close to one another, and when one agent switches to report correctly it is unlikely that the run will be divergent. Furthermore, in the conditions I have tested, it seems that the parameter alpha does not have to be very high for agents to start to report sincerely. In the conditions that start with 2 insincere agents and 1 sincerian, the desire to conform to other people can be all that matters (alpha = 0), and yet 100% of runs came out successful. Even when more insincere agents are introduced, the alpha equal to 0.3 (so the desire to conform – 0.7, is still very much higher than the desire to be sincere – 0.3) is enough to guarantee a run to be successful. It is worth pointing out the chaotic nature of the simulation. In conditions with parameters that make the divergence rate high, it still happens occasionally that some runs are successful, sometimes even after a very few ticks.

Are there any conclusions that can be drawn from the model to say something about actual interactions between people in the state of PI? It seems that an early intervention is important if we want to get people out of PI. If they speak too much between one another, one contrasting testimony is unlikely to make the difference. It also seems to be the case that PI cannot persist when agents at least minimally care about reporting their beliefs sincerely. The desire to conform can be quite high and yet some amount of care about not lying offers a good protection against the persistence of PI. It also seems important that when a sincere agent is introduced to a group in a state of PI, she is listened to by the highest number of agents as soon as possible. That makes it more likely that others will change their testimony and start to affect other people, making a chain reaction that leads to a disruption of PI.

Code

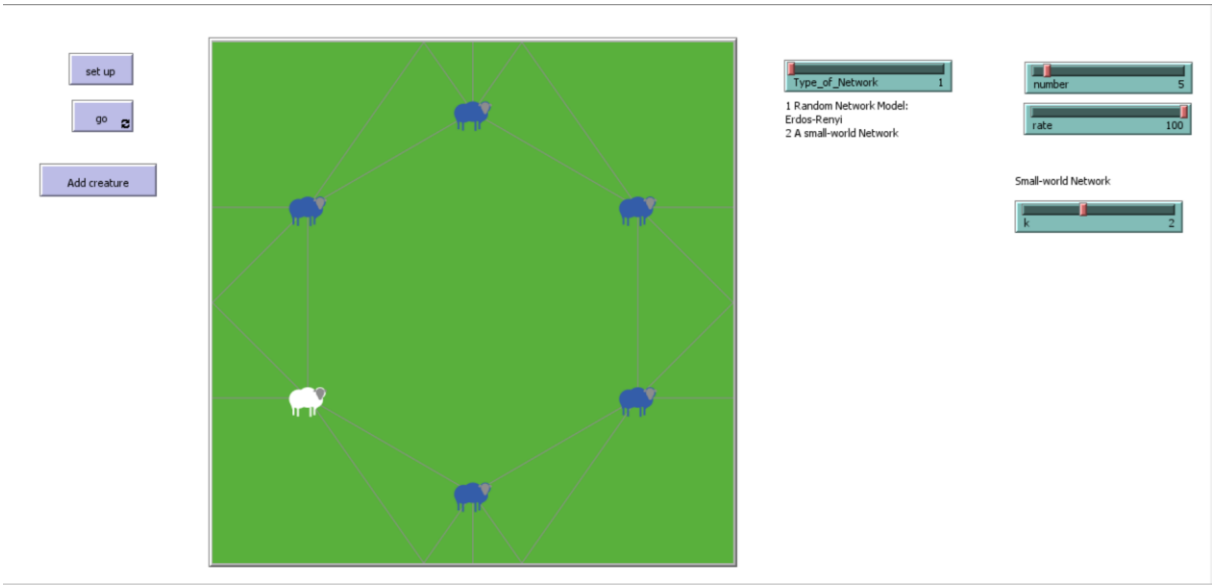

Figure 6 Interface of Experiment 1 Simulation

```
breed [sheeps sheep]

breed [creatures creature]

turtles-own [

  U; utility function to say P

  U2; utility function to say not P

  alpha; truth-seeking orientation

  N; proportion of agents that a sheep interacted with saying P

  P; strength of a private belief that p

  d; lie detection probability

  S; personal signal

  T; testimony

  v; tally of interactions

  I; how many said P=1

]

globals [

  temp; store if the agent signaled P=1

  ; number the number of agents in the simulation

  ; rate the likelihood of connection

]
```

```
; Add a sincerian

to addcreature

  create-creatures 1 [

    set color white

    set size 3

    set shape "sheep"

  ]

  layout-circle turtles 15

  ask creatures [

    set alpha 1.0; choosing the number between 0 and 1

    set N 0; this number will grow when encountering agents that signal
that they believe that P

    set P 1

    set T 1

    set v 0; current number of interactions with others

    set I 0;

  ]

  ask turtles [

    ask other turtles [

      if random 100 < rate and not link-neighbor? myself [

        create-link-with myself

      ]

    ]

  ]

end

to setup

  clear-all

  reset-ticks

  ask patches [set pcolor green]

  ; Random network
```

```
if Type_of_network = 1 [
  ; Create agents
  create-sheeps number [
    set color blue
    set size 3
    set shape "sheep"
  ]
  ; Connect connections between agents with a probability of rate%
  ask turtles [
    ask other turtles [
      if random 100 < rate and not link-neighbor? myself [
        create-link-with myself
      ]
    ]
  ]
  ; Make it visually nicer that agents are all in a circle
  layout-circle sheeps 15
]
; Small world network, more realistic
if Type_of_network = 2 [
  ; Create agents
  create-sheeps number [
    set color blue
    set size 3
    set shape "sheep"
  ]
  layout-circle sheeps 15
  ask sheeps [
    if random 100 < rate [
      create-links-with other sheeps in-radius ((30 * pi / number)) [
```

```
          ;; Connect with neighbors
          set color blue ;; Optional: color the links
        ]
      ]
      ;; Create a few long-range connections
      repeat k [
        let target one-of sheeps with [self != myself and not link-
neighbor? myself]
        if target != nobody [
          create-link-with target [
            set color red
          ]
        ]
      ]
    ]
  ]
  ; Assign the starting value to the parameters
  ask sheeps [
    set alpha 0.25; choosing the number between 0 and 1
    set N 0; this number will grow when encountering agents that signal
that they believe that P
    set P 1
    set T 0;
    set v 1; current number of interactions with others
    set S 0
    set I 0
  ]
end
to go
  ; Modeling interactions between agents, agents receive a signal from
another agent about their belief that P and update their utility functions
accordingly
```

```
  ask turtles [

    let interacting-turtle one-of link-neighbors ; choosing one sheep that
the sheep is connected to

    ifelse T = 1 [ set temp 1] [set temp 0]; calculating what was signaled
by a turtle

    if interacting-turtle != nobody [ ; in case there was no links

      ask interacting-turtle [

        set v v + 1; updating the number of sheep encounters

        set I I + temp

        set N I / v; updating what fraction of interaction was in favor of
p

        set U (alpha * P) + ((1 - alpha) * N) ; updating the utility
function according to new N

        set U2 1 - U

        ifelse U > U2 [ set T 1] [set T 0]  ; deciding what to say

        if U = U2 and random 100 < 50 [set T 1]

              ]

    ]

  ]

  ; Show when a sheep is insincere

  ask turtles [

    if T = 1 and P = 0 [set color red]

    if T = 0 and P = 1 [set color blue]

    if T = 1 and P = 1 [set color white]

    if T = 0 and P = 0 [set color white]

  ]

  tick

  if all? turtles [color = white] [

    show ticks

    print "STOP"

    stop

  ]
```

end

2.5.2 Experiment 2

The second experiment is also designed to investigate the conditions needed for PI to be resolved. Instead of a sincerian, this time we introduce a contrarian to the network of insincere agents. A contrarian is a person, that does not have a desire to sincerely present their belief. In that way, the contrarian has an $\alpha = 0$, such that her truth-seeking orientation is low. However, their utility functions differ from the other agents. She wants to testify the opposite of what she thinks the majority of agents believe. That is why a contrarian testifies that ϕ, when $N_a(\neg\phi)$ is bigger than $N_a(\phi)$. Just like in the previous experiment, I start when the alpha of insincere agents is equal to 0. During the experiment, I'll adjust the alpha parameter to observe its impact on achieving a state where all agents sincerely share their beliefs. Additionally, I will investigate how the number of insincere individuals in a society influences whether and how fast PI resolves. For simplicity's sake, the lying detector mechanism is disabled.

Simulation set-up

The simulation starts with x-number of insincere agents. Their parameters look the same as in the previous experiment.

- alpha y;
- set N 0;
- P 1;
- T 0;
- v 1;
- I 0;

Like with the previous experiment, I am going to vary alpha in the course of the experiment, so its value is not stated. Alpha and P remain constant throughout each simulation run. However, N, T, v, and I are updated after every simulation tick. During each tick of this experiment, every agent shares their belief with another randomly selected agent linked within the network. Subsequently, agents adjust their parameters based on the testimonies they receive.

The parameters for a contrarian look as follows:

- alpha 0
- N 0
- P 1
- T 1
- v 0; a contrarian enters the network and reacts to the environment without the preconception or history of interactions
- I 0

Just like insincere agents, contrarians update their N, T,v, and I after every simulation tick. The value of T, however, is determined according to a different utility function:

$U(\phi) = N_a(\neg\phi)$

$U(\neg\phi) = N_a(\phi)$

And T = 1 if $U(\phi) > U(\neg\phi)$

Testing conditions

Testing conditions correspond to the testing conditions from the previous experiment. I ran the simulation 50 times for each parameter setup. The conditions of success are defined slightly differently than in the previous experiment. If after 10000 ticks, all of the agents, except a contrarian, did not become sincere the run was marked as divergent, otherwise, the run is successful. The contrarian does not have to report successfully because when she manages to affect the network in such a way that they start to report sincerely their belief in a proposition that the contrarian privately believes in as well, it is given by the definition of the contrarian that she will start to report insincerely.

For each population size, I ran 50 simulations varying the alpha of insincere agents until I got 50 successful runs with the same parameters. I did it for the following conditions:

- 2 insincere agents, 1 contrarian
- 3 insincere agents, 1 contrarian
- 4 insincere agents, 1 contrarian
- 5 insincere agents, 1 contrarian

My starting hypothesis for this experiment was as follows:

1. A contrarian will have a similar effect as a sincere sheep in this context. It will allow agents to break out of PI, in a similar context as a truth-seeking agent.
2. The more insincere agents there are, the slower the transition is or it becomes impossible for all 50 runs to be successful

Results

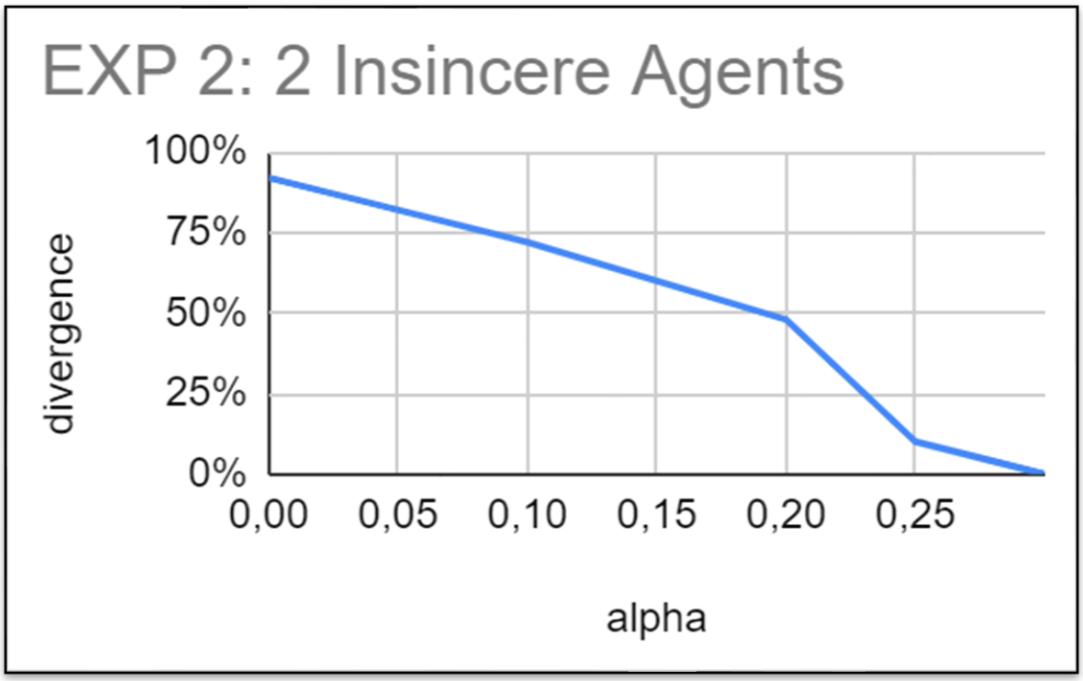

Figure 7 EXP 2: 2 Insincere Agents

When a simulation started with 2 insincere agents with alpha equal to 0 and 1 contrarian, the divergence rate was high – 95% (Figure 7). At alpha equal to 0.25 we got almost a perfect success rate, where 90 % were successful, and at alpha equal to 0.3 all runs were successful.

When one more insincere agent was added to the network, the situation remained fairly similar (Figure 8). With the alpha of insincere agents equal to 0, the percentage of divergent runs got slightly higher, 96%. At alpha equal to 0.2, the divergence rate got significantly better, 20 %, and at alpha equal to 0.37 all runs were successful.

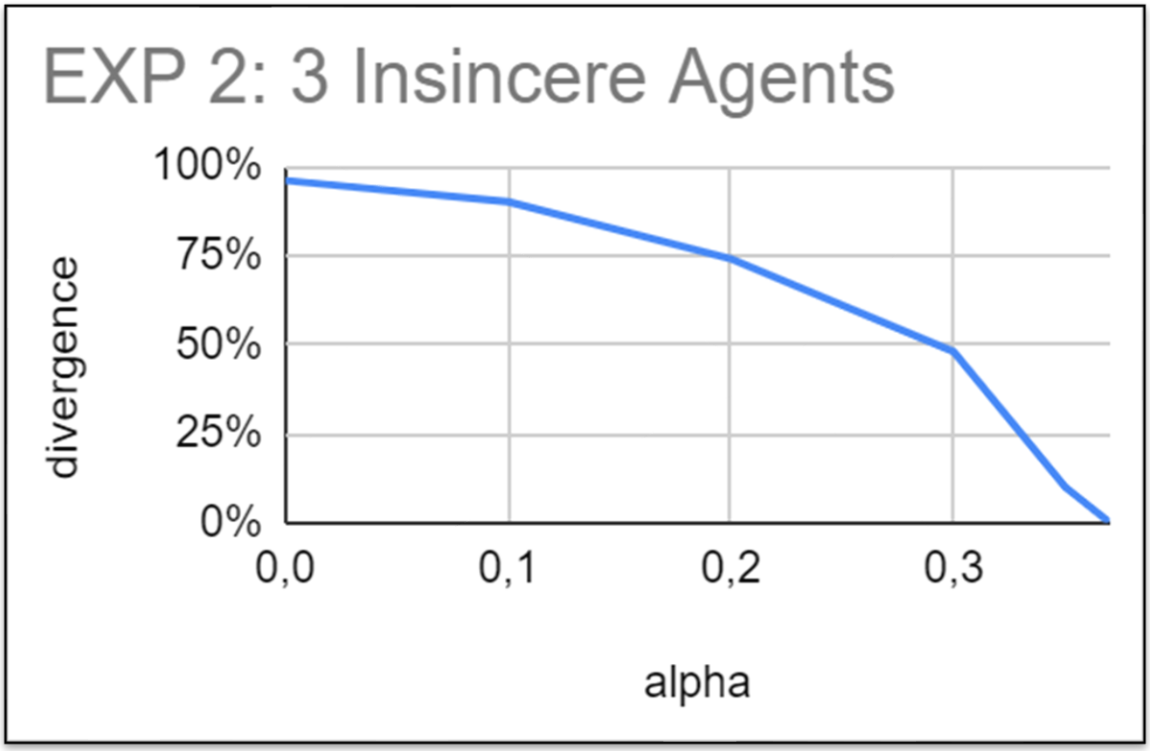

Figure 9 EXP 2: 3 Insincere Agents

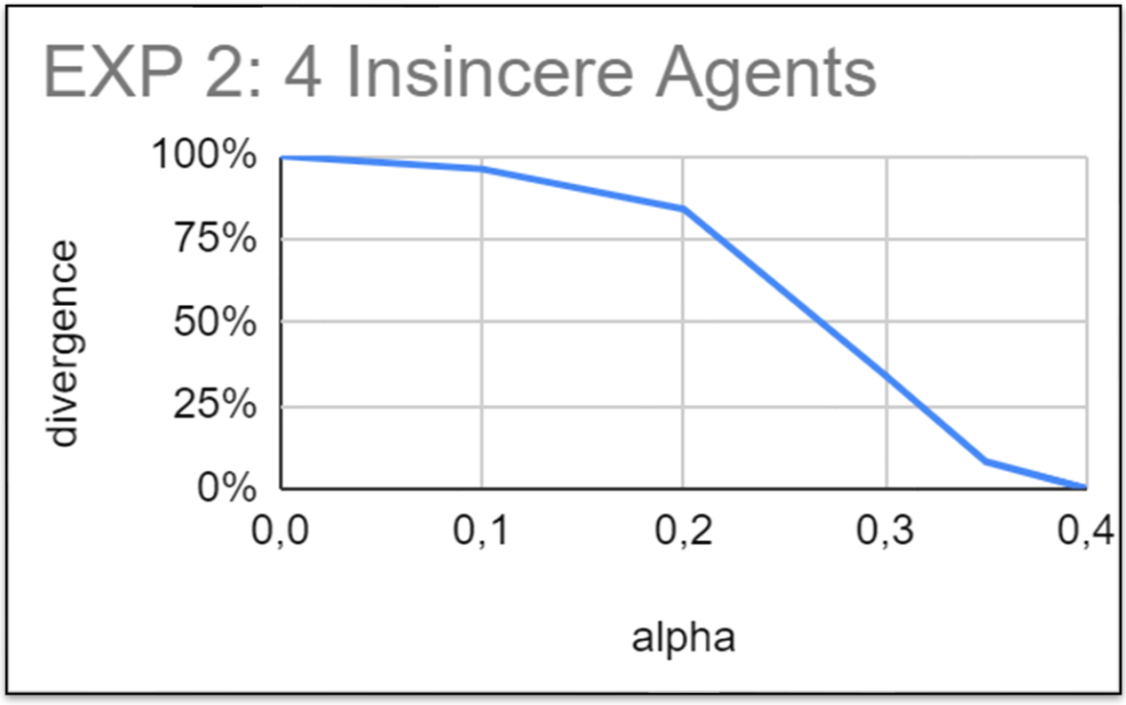

Figure 8 EXP 2: 4 Insincere Agents

The trend of new agents making it more difficult to break off PI continued when the simulation started with 4 insincere agents and 1 contrarian (Figure 9). The number of divergent runs for alpha is 100%. The improvement in the number of successful runs happened with the higher alpha, reaching 100 % for alpha = 0.4.

The agent's behavior did not change much when we added one more insincere agent to the network (Figure 10). Just like for a condition with 4 insincere agents, when alpha was equal to 0, the divergence rate was 100 % and 50 successful runs happened when alpha was equal to 0.4.

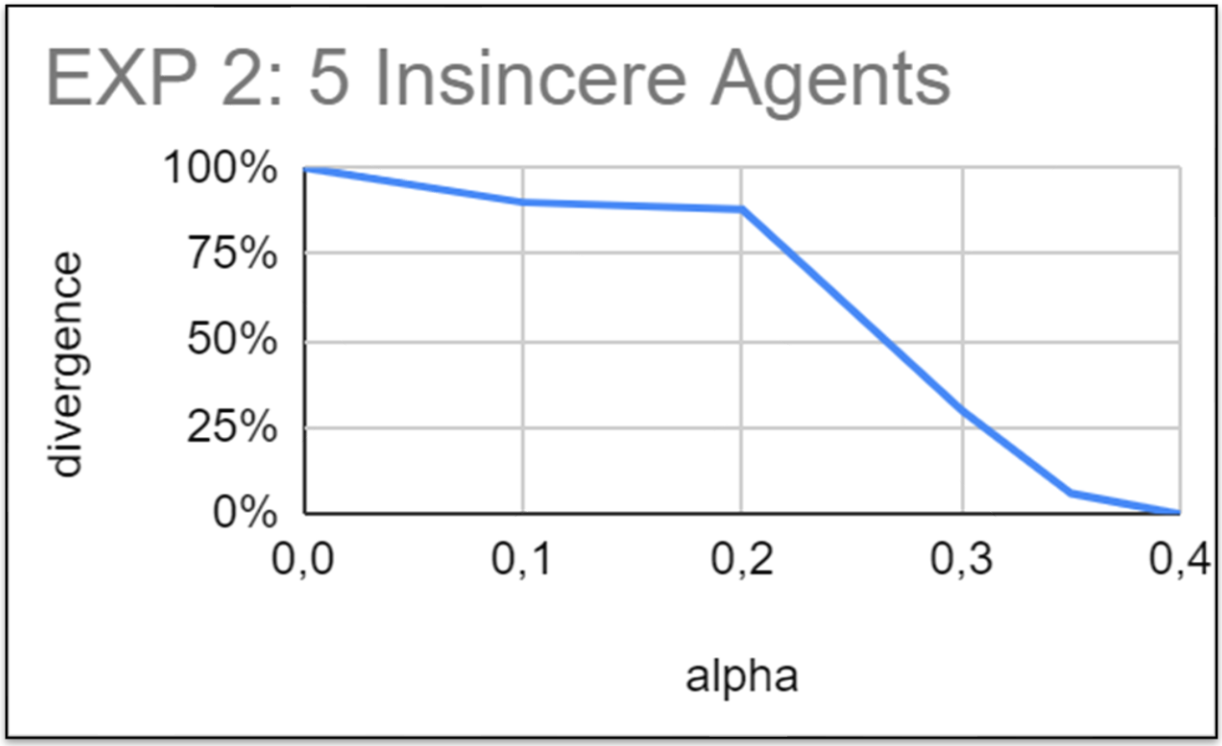

Figure 10 EXP 2: 5 Insincere Agents

In this next graph (Figure 11), we can see the dependence of the percentage of successful runs when the alpha of insincere agents is equal to 0, and the ratio of sincere to insincere agents. The rate of divergent runs of the simulation rises linearly until it quickly reaches 100 %.

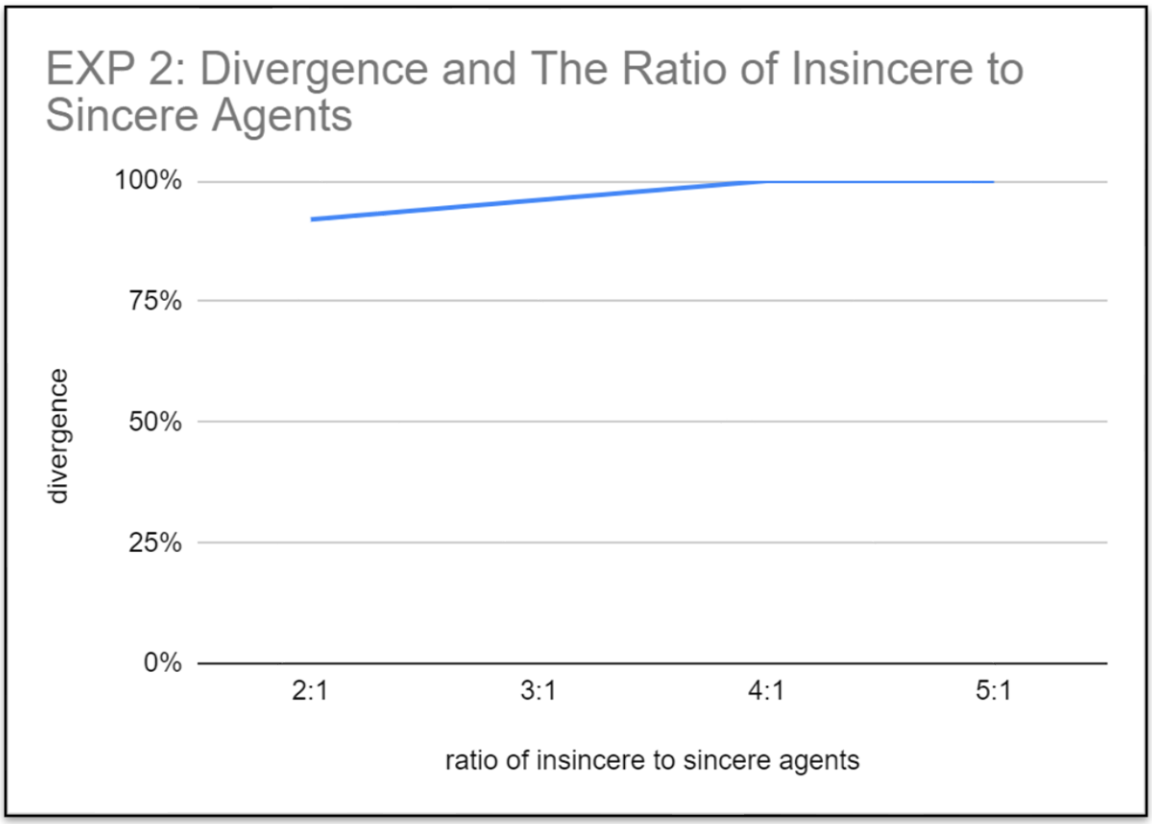

Figure 11 EXP 2: Divergence and The Ratio of Insincere to Sincere Agents

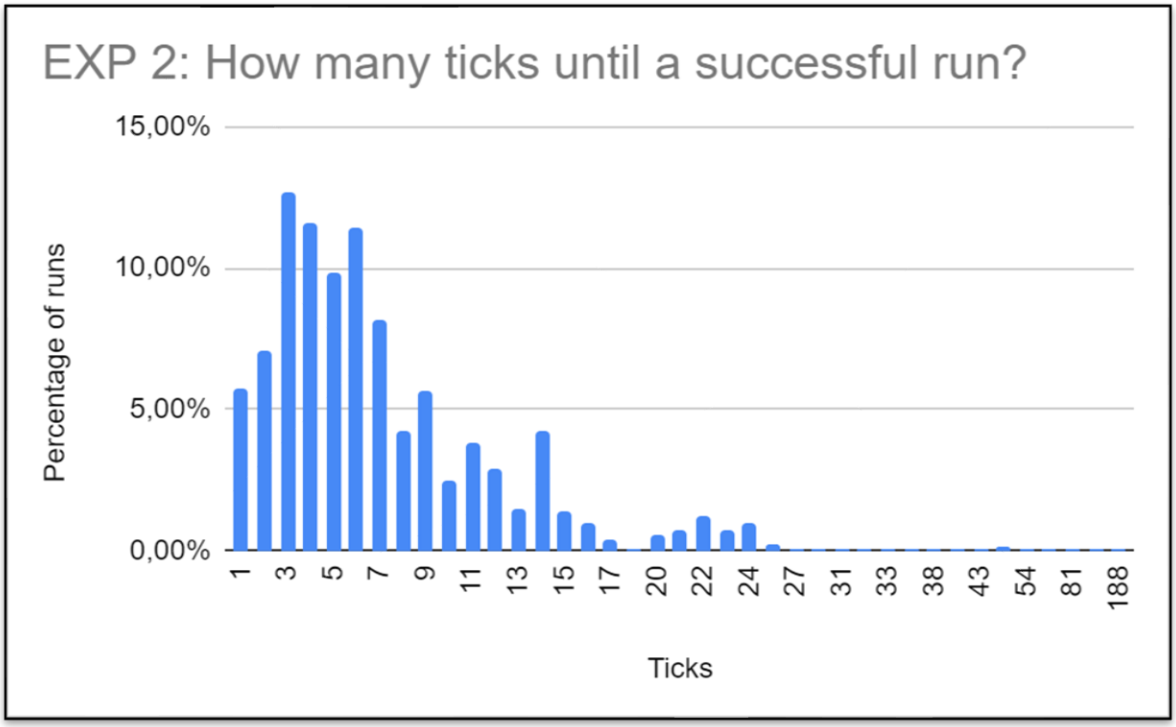

Figure 12 EXP 2: How many ticks until a successful run?

In the next graph (Figure 12) I collected the data from all of the successful runs of the experiment and showed how big the percentage of each successful run terminated at a given number of ticks. Like with the first experiment, it seems that a run is very unlikely to be successful after a relatively small number of ticks (around 15).

Discussion

The first hypothesis was not confirmed in this experiment. A contrarian does not have as powerful of an effect on the network as a sincerian. It is because the contrarian reacts to the other agent's changing its testimony based on the current stated publicly majority opinion. That means that the effect of her actions can be undermined by her following actions. A sincerian is consistent with his testimony allowing it to affect the other agents in a straight-forward way. However, a society of only insecure agents will not get themselves out of the state of PI, unless their alpha is at least equal to or larger than 0.5. That means that contrarians do help in the resolution of PI. However, when there are more than 3-4 insincere agents in the network and only one contrarian, alpha seems to be quite high for a successful run to happen consistently.

The second hypothesis was partially confirmed by the experiment. The more insincere agents were introduced to the experiment, the higher alpha was needed to get 50 successful runs one after another. The rate of divergence was also higher for the given values of alpha, the more insincere agents there were in a network. However, when the PI did get resolved, there did not seem to be a significant dependence on alpha on how many ticks it took for it to happen. It seems that in general if stability arises, it arises quite early during simulation no matter the parameter setup of the simulation. Like with the previous experiments through a visual observation of the unfolding of simulation, it was possible to notice that agents usually switch to report correctly their private beliefs close to one another, and when one agent switches to report correctly it is likely that the run will be successful.

This experiment suggests that contrarian behavior helps the network to get out of the state of pluralistic ignorance. However, it does so less efficiently than a sincerian. Moreover, a contrarian herself will be constantly fluctuating, sometimes machine mismatching their public statements, which means she will be displaying the behavior of people in the state of PI (notably however, for different reasons than a desire to conform).

Code

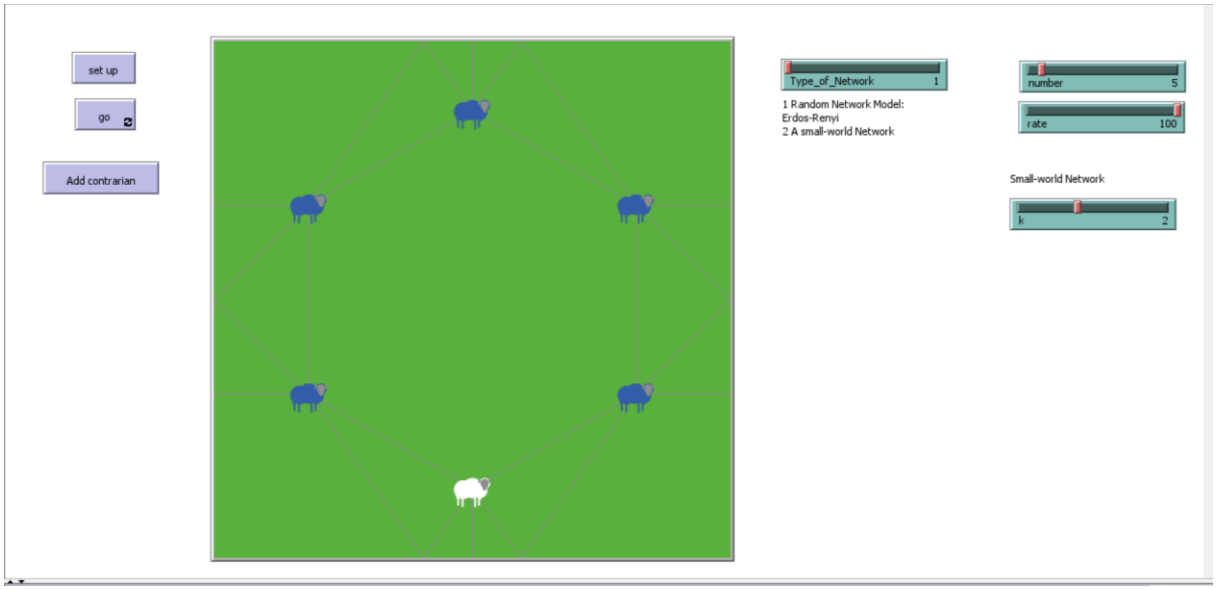

Figure 13 Interface of Experiment 2 Simulation

```
breed [sheeps sheep]; agents, nodes of the network

breed [creatures creature]

turtles-own[

  U; utility function to say P

  U2; utility function to say not P

  alpha; truth-seeking orientation

  N ; proportion of agents that a sheep interacted with saying P

  P; strength of a private belief that p

  d; lie detection probability

  S; personal signal

  T; testimony

  v; tally of interactions

  I; how many said P=1

]

globals[

  temp; store if the agent signaled P=1

  ]
```

```
; add a contrarian

to addcreature

 create-creatures 1 [set color white set size 3 set shape "sheep"]

 layout-circle turtles 15

 ask creatures

   [set alpha 0; choosing the number between 0 and 1

    set N 0; this number will grow when encountering agents that signal that
they believe that P

    set P 1

    set T 1

    set v 0; current number of interactions with others

    set I 0;

    ]

  ask turtles [

    ask other turtles [

      if random 100 < rate and not link-neighbor? myself [

        create-link-with myself

      ]

    ]

  ]

end

to setup

clear-all

reset-ticks

ask patches [set pcolor green]

;random network

 if Type_of_network = 1 [

;create agents

  create-sheeps number [set color blue set size 3 set shape "sheep"]

;connect connections between agents with a probability of rate%

ask turtles [
```

```
    ask other turtles [
      if random 100 < rate and not link-neighbor? myself [
        create-link-with myself
      ]
    ]
  ]
;make it visually nicer that agents are all in a circle
layout-circle sheeps 15
  ]
; small world network, more realistic
  if Type_of_network = 2
  [ ;create agents
  create-sheeps number [set color blue set size 3 set shape "sheep"]
  layout-circle sheeps 15
    ask sheeps [
      if random 100 < rate [create-links-with other sheeps in-radius ((30 *
pi / number)) [ ;; Connect with neighbors
      set color blue  ;; Optional: color the links
      ]]
    ;; Create a few long-range connections
      repeat k [
      let target one-of sheeps with [self != myself and not link-neighbor?
myself]
      if target != nobody [
      create-link-with target [set color red]
      ]
    ]
  ]
  ]
; assign the starting value to the parameter
ask sheeps
```

```
  [set alpha random-float 1.0; choosing the number between 0 and 1

   set N 0; this number will grow when encountering agents that signal that
they believe that P

   set P 1

   set T 0;

   set v 1; current number of interactions with others

   set S 0

   set I 0

  ]

end

to go

 ; modeling interaction between agents, agents receive a signal from another
agent about their belief that P and updates its own utility functions
accordingly.

  ask turtles [

    let interacting-turtle one-of link-neighbors ; choosing one sheep that
the sheep is connected to

    ifelse T = 1 [ set temp 1] [set temp 0]; calculating what was signaled
by a turtle

    if interacting-turtle != nobody [ ; in case there was no links

        ask interacting-turtle [

        set v v + 1 ; updating the number of sheep encounters

        set I I + temp

        set N I / v ; updating what fraction of interaction was in favor of
P

        set U (alpha * P) + ((1 - alpha) * N) ; updating the utility function
according to new N

        set U2 1 - U

        ifelse U > U2 [ set T 1] [set T 0]  ; deciding what to say

        if U = U2 and random 100 < 50 [set T 1]

      ]

    ]

  ]

; how contrarians react to the new information
```

```
  ask creatures [

   ifelse U > U2 [ set T 0] [set T 1]  ; deciding what to say

   if U = U2 and random 100 < 50 [set T 0]

  ]

; visual representations of agents being insincere, in the simulation it
visually indicates if the network is in the state of pluralistic ignorance

 ask turtles [

    if T = 1 and P = 0 [set color red]

    if T = 0 and P = 1 [set color blue]

    if T = 1 and P = 1 [set color white]

    if T = 0 and P = 0 [set color white]

    ]

tick

;stop the simulation when sheep are no longer in the state of PI

   if all? sheeps [color = white] [

    show ticks

    print "STOP"

    stop

  ]

end
```

2.6 Model of Pluralistic Ignorance - Code

In this section, I present the code for the basic model of PI on which the experiments were based. Different experiments can be set up using this model that due to the length limitation of this thesis could not be included here.

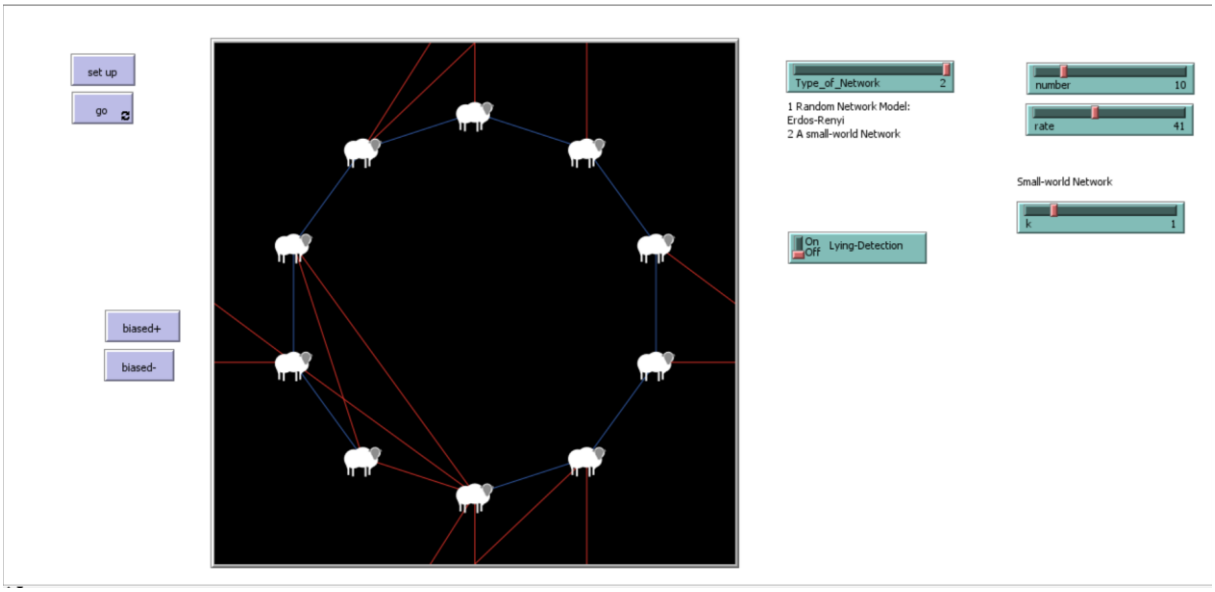

Figure 14 Interface of the Basic Model

```
breed [sheeps sheep]; agents, nodes of the network

turtles-own[

  U; utility function to say P

  U2; utility function to say not P

  alpha; truth-seeking orientation

  N ; proportion of agents that a sheep interacted with saying P

  P; strength of a private belief that p

  d; lie detection probability

  S; personal signal

  T; testimony

  v; the tally of interactions

  I; how many said P=1

  ]

globals[

  temp; store if the agent signaled P=1

]

;make agents change their minds but from some outside sources and not from
talking to one another

to biased+ ;stronger belief

  ask sheeps [
```

```
    if P <= 0.95 [set P P + 0.05]]
end
to biased- ; weaker belief
  ask sheeps [
    if P >= 0.5 [set P P - 0.05]]
end
;set up the simulation
to setup
  clear-all
  reset-ticks
;random network
 if Type_of_network = 1 [
;create agents
  create-sheeps number [set color white set size 3 set shape "sheep"]
;connect connections between agents with a probability of rate%
ask sheeps [
    ask other sheeps [
      if random 100 < rate and not link-neighbor? myself [
        create-link-with myself
      ]
    ]
  ]
;make it visually nicer that they are all in a circle
layout-circle sheeps 15
  ]
; small world network, more realistic network
  if Type_of_network = 2
  [ ;create agents
  create-sheeps number [set color white set size 3 set shape "sheep"]
  layout-circle sheeps 15
```

```
  ask sheeps [

      if random 100 < rate [create-links-with other sheeps in-radius ((30 *
pi / number)) [ ;; Connect with neighbors

      set color blue  ;; Optional: color the links

      ]]

    ;; Create a few long-range connections

      repeat k [

      let target one-of sheeps with [self != myself and not link-neighbor?
myself]

      if target != nobody [

        create-link-with target [set color red]

      ]

    ]

  ]

  ]

; assign the starting value to the parameters

ask sheeps

  [set alpha random-float 0; choosing the number between 0 and 1

   set N 0; this number will grow when encountering agents that signal that
they believe that P

   set P random-float 1.0

   ifelse P >= 0.5 [ set T 1] [set T 0]; without feedback from others, we
say what we believe is the case

   set v 0; current number of interactions with others

    ]

end

to go

 ; modeling interaction between agents, agents receive a signal from another
agent about their belief that P and updates its own utlity functions
accordingly

  ; what happens when agents can detect lies

  if Lying-Detection = TRUE

 [ ask turtles [
```

```
    let interacting-turtle one-of link-neighbors ; choosing one sheep that
the sheep is connected to

    if interacting-turtle != nobody [ ; in case there was no links

;when an agent does believe the proposition but says that he does not, the
probability of detection rises the stronger he believes it

      ifelse T = 0 and P > 0.5 [

        set d 2 * (P - 0.5)

        ifelse random-float 1.0 < d [ set S 1] [ set S 0]

] [ ;when an agent does not believe the proposition but says that he does,
the probability of detection rises the stronger he disbelieve it

         ifelse T = 1 and P < 0.5 [

           set d 2 * P

    ifelse random-float 1.0 < d [ set S 0] [ set S 1]

  ] [

    set S T

  ]

]

      ifelse S = 1 [ set temp 1] [set temp 0]; keeping tall how the interaction
affect the receiving sheep

        ask interacting-turtle [

        set v v + 1 ; updating the number of sheep encounters

        set I I + temp

        set N I / v ; updating what fraction of interaction was in favor of
P

        set U (alpha * P) + ((1 - alpha) * N) ; updating the utility function
according to new N

        set U2 1 - U

        ifelse U > U2 [ set T 1] [set T 0]  ; deciding what to say

        if U = U2 and random 100 < 50 [set T 1] ; when your credence is right
in the middle, randomly chose what to say

      ]

    ]

   ]

  ]
```

```
  if Lying-Detection = False

  [

    ask turtles [

    let interacting-turtle one-of link-neighbors ; choosing one sheep that
the sheep is connected to

    ifelse T = 1 [ set temp 1] [set temp 0]; calculating what was signaled
by a turtle

    if interacting-turtle != nobody [ ; in case there was no links

        ask interacting-turtle [

        set v v + 1 ; updating the number of sheep encounters

        set I I + temp; keeping track of all interactions when someone
testified that P

        set N I / v ; updating what fraction of interaction was in favour of
p

        set U (alpha * P) + ((1 - alpha) * N) ; updating the utility functions
according to the new N

        set U2 1 - U

        ifelse U > U2 [ set T 1] [set T 0]  ; deciding what to say

        if U = U2 and random 100 < 50 [set T 1]  ;when your credence is right
in the middle, randomly chose what to say

      ]

    ]

   ]

  ]

; visual representations of agents being insincere, in the simulation it
visually indicates if the network is in the state of pluralistic ignorance

 ask turtles [

    if T = 1 and P < 0.5 [set color red]

    if T = 0 and P > 0.5 [set color blue]

    if T = 0 and P <= 0.5 [set color white]

    if T = 1 and P >= 0.5 [set color white]

    ]

  tick

end
```

3 Sincerity, Conformity, and Truth

Pluralistic Ignorance (PI) is a complex social phenomenon occurring in different contexts. It is usually considered undesirable because it prevents people from expressing themselves sincerely and involves them forming false beliefs about others. Therefore, we can analyze PI in two dimensions: epistemic and moral. Agents in the state of PI do not live authentically. They either need to pretend they have beliefs they disagree with, or they need to conform to social norms that they think are wrong. Moreover, in doing so they are lying to one another which perpetuates the state in others. Epistemically, the wrongness comes from the fact that false meta-beliefs (beliefs about other people's beliefs) spread within society. That undermines the foundation for the honest discussion and exchange of information which can have dramatic epistemic consequences. On the other hand, the desire for conformity is not, in itself, wrong. Agreement in a social sphere can be a beneficial state for the society, where social interaction goes smoothly, and norms are clear and not contested. In this chapter, I am going to explore what kinds of epistemic and more virtues and vices are involved in perpetuating as well as breaking out of the state of PI. Since PI is a social phenomenon those virtues and vices will be considered in terms of their role in the person that possesses them, but, more importantly, to a social network an individual belongs to.

3.1 Moral and Epistemic Vices

Epistemic virtues are those characteristics of a person that bring her closer to realizing the epistemic value. There are different accounts of what constitutes that value. I will be relying on the view that truth is the ultimate epistemic value, and other values are derived from it. The moral virtue that in normal circumstances benefits from discovering/getting closer to the truth, which is also relevant in the discussion about PI is sincerity. The corresponding epistemic vice is then dishonesty.

Moral virtues, on the other hand, are those characteristics of a person that bring her closer to realizing the moral value. For simplicity's sake, I will call this value moral goodness. The moral virtues then relevant to PI that also promote achieving moral goodness are authenticity and sincerity. The first virtue is living your life as you want it, whereas sincerity is involved with not lying to others.

3.2 Character Profiles in Pluralistic Ignorance

Based on the results of the experiments from the second chapter, I will now have a look at 3 character profiles of agents in a social network and see what moral and epistemic character they have.

The first character is “a sheep”. This is an agent whose desire for conformity is stronger than the desire for sincerity. When she believes that the other agents believe the opposite of what she believes in, she will conform and share her beliefs insincerely. Therefore, when placed in the network of pluralistic ignorant agents she will perpetuate the phenomenon. In this context she ends up living inauthentically, either conforming to social norms that she thinks are wrong or stopping herself from honestly sharing her beliefs or opinions. Moreover, her lack of authenticity prevents others from forming true beliefs about her beliefs. Whenever a person states a proposition: P, there is an implicit “I believe that…” that precedes it. In that way even if P is false, a person is still sincere (and therefore does not lie) when she says P if she believes that P is indeed the case (Table 1). Therefore, a person’s sincerity does not depend on the truth of the proposition P. A sheep will therefore often be insincere and lie about her beliefs, especially when the majority of people in her network state publicly that they believe the opposite of what she does.

Statement	Condition for truth	Condition for falsehood
God exists.	God’s existence	God’s inexistence
X says: (I believe that) God exists.	X believes that God exists.	X doesn’t believe that God exists.

Table 1 Sincerity and Truth

The second character is a “sincerian”. This character always shares her beliefs sincerely no matter what other agents in her network are doing. In this way she shares her beliefs sincerely, preventing others from forming false meta-beliefs. Not only that but if her beliefs match the private beliefs of a sheep, she has the potential to influence sheep to be sincere and live more authentically, breaking out of the state of PI. Moreover, sincerian lives authentically since her beliefs and opinions match her actions.

The third character is a contrarian. She does not care about sharing her beliefs but rather always says the opposite of what she perceives a majority opinion is. It can coincide that she shares her private beliefs publicly but only when it goes against the majority opinion. Therefore, she is flippant about the truth. She also does not seem to live authentically: even though she constantly tries to defy other members of the network, she fully relies on them to decide what belief, opinion, or social norms she will publicly endorse. On the other hand, when she enters society in a state of PI, she is sometimes able to make other sheep share their beliefs sincerely, breaking out of PI.

Character	Ethics – sincerity, authenticity	Epistemology – true meta beliefs (beliefs about other's beliefs)
Sheep	Individual **X** Group **X**	Individual **X** Group **X**
Contrarian	Individual **X** Group ✓	Individual **X** Group ✓
Sincerian	Individual ✓ Group ✓	Individual ✓ Group ✓

Table 2 Ethical vs Epistemic Virtues

3.2.1 Mandevillian Intelligence of Contrarians

Epistemically a sheep and a sincerian have an unambiguous effect. The former has undesirable effects on both individual and group levels, while the latter is desirable on both of those planes. Only when it comes to a contrarian a mismatch on individual and group levels arises (Table 2). That is the case of Mandevillian Intelligence. Smart (2017) defines it as "a specific form of collective intelligence in which individual cognitive vices (i.e., shortcomings, limitations, constraints, and biases) are seen to play a positive functional role in yielding collective forms of cognitive success".

	Individual vice	Individual virtue
Collective vice		Reverse-Mandevillian Intelligence
Collective virtue	Mandevillian Intelligence	

Table 3 Mandevillian Intelligence

In the case of a contrarian, her individual vice is that she is flippant about their testimony and therefore often ends up lying to others and being insincere. However, in some contexts when they enter a social network with other people in a state of PI, they can make them break out of it, allowing for others to live more authentically and share their beliefs sincerely. It is important to notice, however, that this situation is highly context-dependent. For example, a society with only contrarians will not be uniformly sincere. Moreover, a contrarian can make people change what they testify publicly and be more sincere, but as soon as the majority of people she interacts with do so, she will react to it by changing her own testimony. Therefore, she will exhibit the behavior of a sheep in a state of PI, even if motivated by different desires.

3.3 Utilitarian Perspective

Let us now consider the effect of the moral and epistemic character of the agents on the utility generated by their actions. According to John Stuart Mill, the Utility Principle states that "actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of the happiness" (Mill, 1879, pp. 9-10). Since we are taking on the utilitarian perspective when we have to consider two cases:

1. Publicly supporting ϕ brings utility to the group.
2. Publicly support ϕ does not bring utility to the group.

I assume that a contrarian, a sincere sheep, and a sheep believe ϕ, but that sheep testify that not P.

Let us begin with the first scenario, in which public support of ϕ is beneficial for the group, bringing happiness to its members. It can be that ϕ is true, so supporting ¬ϕ helps us navigate this world better. It can be that public support of not ϕ, makes people behave in a morally desirable way. For example, Bernard Williams thinks that belief in the truth of utilitarianism corrupts people because "a utilitarian must always be justified in doing the least bad thing which is necessary to prevent the worst thing that would otherwise happen in the circumstances (including, of course, the worst thing that someone else may do) - and what he is thus justified in doing may often be something which, taken in itself, is fairly nasty" (Williams, 1972, p. 96). In that case, it is best to not share their beliefs in utilitarianism sincerely. Another example is a soccer team. If players at least publicly state that they are confident they are going to win the match and they are confident that others believe they are going to with the match, it can help the team to perform better even when individuals alone are not sure about their chance of winning.

Character	Utility
Sheep	✓
Contrarian	**X**
Sincerian	**X**

Table 4 Utility, scenario 1

In this scenario being a contrarian or a sincerian can disrupt PI, which can have negative consequences for the group. n this framework, only when acting as if ϕ is the case, is not beneficial, contrarians and sincerian bring in a desirable effect of disrupting PI.

Character	Utility
Sheeple	**X**
Contrarian	✓
Sincerian	✓

Table 5 Utility, scenario 2

3.4 When Sincerity, Truth, and Conformity Are in Conflict

Conformity does not have to be an inherently detrimental character trait. It can "foster harmony between individuals and secure their interpersonal, physical, and mental well-being" (Laursen & Veenstra, 2023, p. 74). Humans are social animals and therefore, we have a need for interpersonal connection. Conformity brings benefits for an individual because it makes it easier to maintain close relationships and foster a sense of belongingness. Moreover, it helps to develop good social skills and ability for perspective-taking. Conformity allows as well for lower levels of conflict as well. In a group setting, conformity serves two roles: group formation and group stability. The group is established based on something that is shared by all of its members. Therefore, all good members of the group are engaged in identity signaling, that is those behaviors that are designed to convey the information about what are defying characteristics of members of that particular social group. Moreover, when people act in a similar way it creates the appearance of consensus, which stabilizes the group. Conformity is also a source of cohesion within a group. Individuals are generally more inclined to dedicate their time and resources and make sacrifices, for groups where members share common traits, compared to those with noticeable differences among members (Laursen & Veenstra, 2023).

However, it can conflict with sincerity if to conform a person needs to lie. Moreover, sincerity can also lead to falsity to spread. In that case, conformity and sincerity can together work against promoting the truth. Let us analyze the behavior of our agent types in the two following conditions:

1. ϕ is false. Conforming to ¬ϕ brings group stability.
2. ϕ is true. Conforming to ¬ϕ brings group stability.

We assume that all agent types privately believe that ϕ is the case.

What happens if ϕ is false and agreement on the truth value of ϕ brings social cohesion? Sheep are in a state of PI, they are insincere about their private belief and therefore, paradoxically, they support true belief in the public sphere. Lack of the opposing voices and a high conformity rate allows for high levels of social cohesion. Social cohesion and truth are therefore in conflict with sincerity. The behavior of sheep promotes and supports only the former values.

Contrarians bring divergent opinions up in public. This can have a beneficial effect on sheep by allowing them to share their beliefs sincerely. However, it also brings the possibility of conflict to the group. Non-conformity signals marginal status and a lack of commitment to a group (Laursen & Veenstra, 2023). A contrarian will always try to testify in a way that goes against the perceived majority belief. It can just so happen that she sometimes shares their beliefs sincerely, and under some conditions, she can make sheep honest and, in that way, make a belief in P common knowledge. In the case when ϕ is false, however, contrarian will have the effect of promoting a false belief, while at the same time risking disrupting social cohesion of society.

Sincerians care deeply about sincerity and always present their beliefs honestly. Her behavior can influence sheep to share their beliefs sincerely. That means that in the scenario where ϕ is false, a sincerian promotes the value of sincerity. However, when a sincerian holds a false belief, her action will contribute to the support of felicity in a social network. Moreover, while the effect of her actions is being realized in a social network, there will be a short period of instability, which may bring conflict to the group.

	Stability of a group	Publicly stated truth
Sheeple	✓	✓

Contrarian	**X**	**X**
Sincerian	Periods of instability	**X**

Table 6 Stability of a group and truth - scenario 1

What happens if ϕ is true, and agreement on the truth value of ϕ brings group stability? Sheep are in a state of PI, they are insincere about their private belief and therefore, they support a false belief in the public sphere, even when they all privately support the true belief. Lack of the opposing voices and a high conformity rate allows for high levels of social cohesion. Group stability is then in conflict with sincerity and truth. The behavior of sheep promotes and supports only the former value.

As usual, contrarians bring divergent opinions to the public. It influences sheep to share their sincere and true beliefs in the public sphere. However, since a contrarian changes her testimony depending on the perceived majority belief, she risks disrupting the stability of a social network.

A sincerian ha a similar effect to a contrarian. She shares her true and sincere beliefs, influencing sheep to do the same. In that way it allows sheep to break out of a state of PI. However, there will be some periods of instability when different beliefs about ϕ are promoted in social networks. This period of instability, however, tends to end quickly and agents in a social network end up publicly and privately agreeing on the truth of ϕ. This is a situation in which sincerity, social cohesion, and truth are promoted in society.

	Stability of the group	Publicly stated truth
Sheep	✓	**X**
Contrarian	**X**	✓
Sincerian	Periods of instability	✓

Table 7 Stability of a group and truth – scenario 2

In the given scenarios the benefit of conformity does not depend on the truth of ϕ. This may be easier to accept if we instead think of ϕ as social norms or values. When the attitudes do not have a truth-value content, it may be beneficial to hold them, when cultivating those specific attitudes helps in maintaining social cohesion, since it does not carry with itself a risk of believing something false.

3.5 Limitations

The validity of the results of the experiments hinges on the validity of the model. Since the model is very general and concerned only with agents where two desires are considered: the desire to speak truthfully and the desire to conform. Other desires may be at play depending on which attitude ϕ we are concerned with. Those desires can modify the behavior of the agents in significant ways which would make the model inapplicable. It is also quite certain that a society may have different 3 character types other than sheep, contrarian, and sincerian. The model does not allow to say anything of the behavior of agents of that sort. However, it is possible that certain attitudes are only concerned with the two desires represented in the model. That is when the model is most useful to show the trends in a social network.

3.6 Further Studies

The investigation from the experiments and the analysis of the previous section suggests that there can be an adaptive reason why pluralistic ignorance may arise in a social network. If the benefits of social cohesion are higher than the benefits of promotion of the truth belief, or if public agreement on social norms, opinions, and values is more beneficial than living authentically and sincerely showing one's attitudes, agents will tend to conform to what the majority is saying. It still, however, does not explain why, even when they share their beliefs privately, agents would end up publicly the opposite. My model can be used to further investigate a question of that sort. In this section, I will suggest some ways to test the hypothesis of the origin of PI.

3.6.1 New Members

The model I created can be used to investigate the results of adding new members to the aging society. An old society contains members publicly and privately promoting a certain attitude P. When new members enter such a society, while they privately oppose ϕ and their desire for conformity is high enough, they will start publicly supporting ϕ. If we then start slowly removing the aging members of society, we may create a situation where we are left with only a new member who is privately supporting ¬ϕ yet publicly supporting ϕ.

3.6.2 Learning from Outside Sources

Another suggestion for a test using the model is a situation where people learn from outside sources, leading them to change their attitude about ϕ. A social network may at first consist of agents agreeing on the truth/value of ϕ. If they each individually learn from sources outside of the social network, and change their attitudes about ϕ, while still publicly conforming to what they used to, agents in the network will end up in a state of PI.

Conclusion

In this thesis, I clarify the concept of pluralistic ignorance (PI), distinguishing it from other phenomena, and adapting the model of PI into an agent-based computer simulation. By the use of the simulation, I have investigated how to dispel PI. In the previous chapter, I have presented possible conflict areas between the values of sincerity, truth-promotion, and conformity. In conclusion, I now take a step back and explain why we should care at all about the possibility of being in the state of PI.

Many benefits of dispelling PI have little to do with epistemic issues. Agents in a state of PI have to deal with stress, isolation, and cognitive dissonance. They may be pushed to change their behavior and publicly expressed attitudes even when there are no good reasons for it. Moreover, sometimes PI prevents us from updating social norms in line with new moral insights because people will rather conform to old norms due to the fear of being a bad member of the group. Agents in the state of PI are also prevented from living their lives authentically. It can be argued that they chose that lot themselves, however, the analysis of individual behavior without taking into account group dynamics does not seem to capture the full picture of how to understand responsibility for being in the state of PI. As we have seen through the experiments on the model, an agent with the same personal characteristics can behave either sincerely or insincerely based on the social network she ended up in. Somewhat paradoxically, we cannot discard the effect of others on our ability to live authentically. PI also creates a dynamic in which people lie to one another (share their meta-beliefs dishonestly), which creates a negative feedback loop - people are lying because people are lying because people are lying… These on their own are good reasons to dispel PI. However, I would like to focus on the epistemic aspect of PI and how it prevents society from discovering truths.

One of the reasons it is desirable for society to engage in the pursuit of knowledge is the instrumental value of having true beliefs. Knowledge allows for progress in a society, in different domains ranging from science, technology, medicine, and arts. On the individual level, knowledge allows for informed decision-making. It helps individuals lead their lives, choose their careers, and care for their health better. Moreover, a free exchange of ideas is a trademark of an open society, where everyone is equally valued. The shared pursuit of understanding and knowledge allows for moral progress and makes society more just.

PI underminers the sincere search for truth. PI can only arise in a society where the majority of members have a greater desire for conformity than to share their beliefs truthfully. How this desire manifests in individual behavior depends on the interactions of agents in the network. This suggests that PI is a sign that there is something about the structure of the interactions of individuals in society that devalues sincerity. This, consequently, has a direct effect on the pursuit of knowledge by society. Therefore, I will argue that PI is undesirable from the epistemic standpoint. Even if it in some cases allows for promoting the truth publicly, when no one does so privately, this happens only because of a lucky accident.

It is difficult to know when we are right about something. There are no phenomenological mechanisms that alert us if we happen to believe a false proposition. Moreover, in contemporary society, our ability to assert knowledge often relies on the testimony of others for seemingly indisputable facts, such as our birthdate, the existence of Australia, the identities of our parents, or the shape of the Earth. Without trust in others' testimony (epistemic trust), we would not be able to celebrate achievements stemming from the collective efforts of society over millennia. No single person possesses all the resources necessary to independently acquire the vast array of knowledge that we collectively hold (McLeod, 2021). Epistemic trust enables individuals to benefit from the wealth of cultural knowledge, a crucial support system given that our understanding of the world depends on each other (Hardwig, 1985). For those reasons, it is particularly important that people assert their beliefs sincerely. Without it, the foundation for our collective effort to gain knowledge is broken.

Jürgen Habermas believes that the process or argumentation is a rule-governed practice. There are certain rules and norms that all of the participants of the discussion. The principles of sincerity and accountability state that every participant of the discussions needs to assert only what she genuinely believes and always either to justify upon what she asserts or to provide reasons for not offering a justification. Sincerity is therefore recognized by Habermas as a precondition for genuine argumentation characterized as a search for truth and organized in a way that better arguments compete for universal approval (Finlayson & Rees, 2023). The recognition that sincerity is important for many, if not all, epistemic practices is the main premise for arguing that we should try to dispel PI.

Firstly, PI prevents the spread of knowledge to new members of a social network. If people privately believe a true proposition, PI does not allow for that truth to spread to new agents. Sometimes it may just so happen that people end up in the state of PI, where proposition ϕ is true and publicly supported as true, while no one privately supports it. This situation does not allow new members to learn the sincere beliefs of the majority of the population, which is a necessary first step to realize that there is falsity spreading in society.

John Stuart Mill argues that to silence one person who differs in her beliefs from all mankind is not more justifiable than that person silencing mankind (Mill, 2001, p.18). In the case of PI, this silencing is often self-imposed or manifests itself in the form of insincerity, but it does not make it any less pernicious. If the new member silences herself due to the expectation of conformity, the whole society loses. Mill further argues that: "the peculiar evil of silencing the expression of an opinion is, that it is robbing the human race; posterity as well as the existing generation; those who dissent from the opinion, still more than those who hold it. If the opinion is right, they are deprived of the opportunity of exchanging error for truth: if wrong, they lose, what is almost as great a benefit, the clearer perception and livelier impression of truth, produced by its collision with error "(Mill, 2001, p.19).

Moreover, in the case where PI allows truth to be promoted publicly in society, it only happens by accident. Since each person publicly supports ϕ only to conform to the perceived majority belief, no one has a reason to believe ϕ that hinges on relevant evidence in the external world. The only reason for public support ϕ is perceived public support for ϕ. That is not to say, that public support for ϕ is never good evidence for the truth of ϕ. What is important, is that people have reached the conclusion that ϕ is true independently. Then we can reason, that since many people have weighted their independent evidence that ϕ (through experiments, eyewitnesses, looking up information on the internet/library) and decided that ϕ is the case, the probability that ϕ is the case is high. Since public support for ϕ by people in a state of PI hinges only on what other agents testify, their assertions are not independent. Of course, a discussion is an important way of figuring out the truth together. Our errors in thinking are corrigible. We are capable of rectifying our mistakes not only through experience but also through discussion on how the experience should be interpreted. Mill claims that "Wrong opinions and practices gradually yield to fact and argument; but facts and arguments, to produce any effect on the mind, must be brought before it." (Mill, 2001, p.

21). It suggests that even when people learn from sources outside of their social network, a crucial component of epistemic practice – argumentative discussion, is still missing in a society in the state of PI.

Even if the ideas promoted publicly turn out to be useful in some way even though they are false, it is still not a good reason to maintain PI. The problem is that in the state of PI people treat a certain belief as unquestionable in the public sphere. Since everyone is publicly conforming to perceived majority voices, no one publicly disputes not only its truth value but also the utility of society believing it. Once again following John Stuart Mill "The usefulness of an opinion is itself a matter of opinion: as disputable, as open to discussion, and requiring discussion as much as the opinion itself "(Mill, 2001, p. 24).

Moreover, even when ϕ is true and promoted publicly as true, individuals in the state of PI will privately support the falsity. Public support for ϕ does not, therefore, guarantee that people will reap the benefits of actually believing a true proposition. If people, when they do not think they are being observed and judged, act in a way that supports ¬ϕ they are still prone to making bad decisions and suffering from them. For example, let us imagine that everyone publicly says they believe that climate change is one of the biggest existential threats and in public acts in a way to prevent global warming. However, in their private household, they consistently burn garbage, use energy uncarefully, and buy new electronics every year. We can see that the benefit of actually believing that climate change is one of the biggest existential threats is not fully there if people only publicly show their support for the truth of that proposition.

Another danger of PI, even when it results in a public support of a true proposition is that people then self-censor and do not challenge the publicly held ideas. Our presumptions that an opinion is true should be based on the fact that it was contested and has not been refuted, rather than it should not be contested because it is presumed to be true (Mill, 2001, p. 21). Since a public opinion that ϕ results from the conformity of members of society, this mechanism of constantly evaluating the truth of opinions circulating in society, that Mill describes, is not in place. Conformity is not a good way to reach a social agreement on truth. A fruitful discussion involves people sharing their reasons for believing a certain proposition. If a person is only saying something to conform to the rest of the group, she cannot fully

participate in the discussion. If a whole society is in PI such discussions are unlikely to occur. Moreover, we need public discussions where we can epistemically trust agents to share their beliefs sincerely. We need scientists to work on different research questions and a way to disseminate the truths of their findings. We need a good education system. And we sometimes need people to disagree with one another. Cognitive diversity refers to the range of different perspectives, approaches, and thinking styles that individuals bring to a particular problem or situation. Conformity can bring people to a consensus too quickly, while disagreement can lead to people maintaining their opposing beliefs and having a chance to explore the solution space more broadly.

The important results of my experiments show that voices opposing the public consensus (be it by a sincerain or a contrarian) are epistemically beneficial for a social network because they contribute to other people showing their beliefs sincerely, which then allows everyone to form true meta-believes about others. Therefore, one of the empirical recommendations of this thesis is, to allow for divergent opinions to be voiced and listened to in society. Moreover, we should encourage people to testify sincerely even when they are afraid of the social cost of doing so. Public understanding that cognitive diversity has benefits for society as a whole (in the form of epistemic discoveries) should guide our norms when it comes to social discourse.

Contrarians' behavior, however, seems to contradict the main premise of supporting the dispelling of PI. They are not primarily concerned with sharing their beliefs sincerely. Nonetheless, there is a two-fold positive effect that contrarians have on epistemic practices. Firstly, if a society is in a state of PI, contrarians, under certain conditions that this thesis has explored, can help insincere people to open up and share their beliefs sincerely. Secondly, they help to challenge the ideas that due to conformity of the other members of the social network have not been challenged. Therefore, they support the sincerity of agents in a social network, at the same time combating non-conformity and contributing to a broader scope of public discussion by exploring ideas that in the public sphere have been left unchallenged.

Work Cited

Allport, F. H. (1924). *Social Psychology*. Boston, Houghton.

Allport, F. H., & Katz, D. (1931). *Students Attitudes: A Report of the Syracuse University Research Study*.

Aydinonat, N. E., Reijula, S., & Ylikoski, P. (2021). Argumentative landscapes: the function of models in social epistemology. *Synthese*, *199*(1), 369-395.

Balbuzanov, I. (2019). Lies and consequences: The effect of lie detection on communication outcomes. *International Journal of Game Theory*, *48*(4), 1203-1240.

Bergmann, M. (2006). *Justification without awareness: A defense of epistemic externalism*. Clarendon Press.

Bjerring, J. C., Hansen, J. U., & Pedersen, N. J. L. L. (2014). On the rationality of pluralistic ignorance. *Synthese*, *191*, 2445-2470.

Burt, R. S. (1980). Models of network structure. *Annual review of sociology*, *6*(1), 79-141.

CenterforPhilosophyofScience. (2021, February 17). "Conformity, Social Networks, and the Emergence of Pluralistic Ignorance" by Kevin Zollman [Video]. YouTube. https://www.youtube.com/watch?v=Cf_Lt2onyT8&ab_channel=CenterforPhilosophyofScience

Chang, B., & Kim, J. N. (2019). Silent minority? Willingness to express opinions of motivated public depending on the perceived group size in the context of GM food controversy. *Public Relations Review*, *45*(5), 101786.

Conte, R., & Gilbert, N. (1995). Introduction. In N. Gilbert & R. Conte (Eds.), *Artificial Societies: The Computer Simulation of Social Life* (1st ed.). Routledge. https://doi.org/10.4324/9780203993699

Damaceanu, R. C. (2008). An agent-based computational study of wealth distribution in function of resource growth interval using NetLogo. *Applied Mathematics and Computation*, *201*(1-2), 371-377.

Dillard, J. P., & Peck, E. (2000). Affect and persuasion: Emotional responses to public service announcements. *Communication Research*, 27(4), 461-495.

Finlayson, J. G., & Rees, D. H. (2023). Jürgen Habermas. In E. N. Zalta & U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy (Winter 2023 Edition). Retrieved from https://plato.stanford.edu/archives/win2023/entries/habermas/

Frey, D., & Šešelja, D. (2018). What is the epistemic function of highly idealized agent-based models of scientific inquiry?. *Philosophy of the Social Sciences*, *48*(4), 407-433.

Frigg, R., & Hartmann S. (Spring 2020 Edition). Models in Science. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2020/entries/models-science/

Funkhouser, E. (2017). Beliefs as signals: A new function for belief. *Philosophical Psychology*, *30*(6), 809-831.

Galam, S. (2004). Contrarian deterministic effects on opinion dynamics: "The hung elections scenario." Physica A: Statistical Mechanics and its Applications, 333, 453-460. https://doi.org/10.1016/j.physa.2003.10.041

Gelfert, A. (2016). *How to Do Science with Models*. Springer.

Grosz, D. (2020). The irrationality of pluralistic ignorance. *Episteme*, *17*(2), 195-208.

Hardwig, J. (1985). Epistemic dependence. *Journal of Philosophy,* 82 (7):335-349.

Kant, I. (2007). Anthropology, history, and education. Cambridge University Press.

Latané, B., & Darley, J. M. (1969). Bystander" apathy". *American Scientist*, *57*(2), 244-268.

Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help*? New York , NY : Meredith Corporation.

Laursen, B., & Veenstra, R. (2023). In defense of peer influence: The unheralded benefits of conformity. Child Development Perspectives, 17(1), 74-80. Chicago.

McLeod, C. (Fall 2021 Edition). Trust. In Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/fall2021/entries/trust/

Mill, J. S. (1879). *Utilitarianism* (7th ed.). Longmans, Green, and Co.

Mill, J. S. (2001). *On liberty*. Batoche Books.

Mohseni, A., & Williams, C. R. (2021). Truth and conformity on networks. *Erkenntnis*, *86*, 1509-1530.

O'Gorman, H. J. (1975). Pluralistic ignorance and white estimates of white support for racial segregation. *Public Opinion Quarterly*, *39*(3), 313-330.

O'Gorman, H. J. (1986). The discovery of pluralistic ignorance: An ironic lesson. *Journal of the History of the Behavioral Sciences*, *22*(4), 333-347.

Pelto, P. J. (1968). The differences between" tight" and" loose" societies. *Trans-action*, *5*, 37-40.

Pöyhönen, S. (2017). Value of cognitive diversity in science. Synthese, 194(11), 4519-4540.

Prentice, D. A., & Miller, D. T. (1996). Pluralistic ignorance and the perpetuation of social norms by unwitting actors. In *Advances in experimental social psychology* (Vol. 28, pp. 161-209). Academic Press.

Reed, M. S., Evely, A. C., Cundill, G., Fazey, I., Glass, J., Laing, A., Newig, J., Parrish, B., Prell, C., Raymond, C., & Stringer, L. C. (2010). What is social learning? *Ecology and Society, 15*(4), r1. Retrieved from http://www.ecologyandsociety.org/vol15/iss4/resp1/

Romanowska, I., Wren, C. D., & Crabtree, S. A. (2021). *Agent-based modeling for archaeology: Simulating the complexity of societies*. SFI Press.

Smart, P. R. (2017). Mandevillian intelligence. Synthese. https://doi.org/10.1007/s11229-017-1414-z

Telesford, Q. K., Joyce, K. E., Hayasaka, S., Burdette, J. H., & Laurienti, P. J. (2011). The ubiquity of small-world networks.

von Neumann, J. (1995). Method in the physical sciences. In F. Bródy & T. Vámos (Eds.), T*he Neumann Compendium*. World Scientific.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature, 393(6684), 440-442.

Williams, B. (1972). *Morality. An Introduction to Ethics*. Cambridge University Press.

Wimsatt, W. C. (1987). False models as means to truer theories. *Neutral models in biology*, 23-55.

Appendix

Experiment 1 - Data

Data

2 insincere agents	
alpha = 0	
run	ticks
1	5
2	13
3	8
4	6
5	8
6	9
7	17
8	2
9	8
10	3
11	5
12	9
13	30

14	8
15	3
16	6
17	4
18	3
19	3
20	3
21	1
22	50
23	5
24	3
25	4
26	2
27	5
28	7
29	35
30	19
31	51
32	11
33	4

34	4
35	10
36	20
37	91
38	8
39	5
40	7
41	24
42	4
43	7
44	13
45	5
46	3
47	12
48	2
49	7
50	12

3 insincere agents				
	alpha = 0	alpha = 0.1	alpha = 0.15	alpha = 0.2

run	ticks			
1	---	4	6	5
2	---	7	23	4
3	---	---	13	5
4	22	---	14	6
5	---	11	8	21
6	---	5	6	17
7	---	16	2	8
8	2	---	3	5
9	20	---	---	4
10	---	2	3	6
11	---	4	24	4
12	---	5	16	8
13	---	39	7	3
14	---	4	3	7
15	---	---	3	6
16	---	---	9	10
17	---	---	6	5
18	---	7	3	7
19	---	8	7	7

20	---	31	12	5
21	5	8	---	10
22	---	---	3	1
23	---	5	1	5
24	3	8	3	3
25	22	3	6	18
26	---	26	3	19
27	---	2	10	9
28	---	6	12	6
29	---	9	16	7
30	---	3	7	7
31	---	9	7	6
32	5	53	9	7
33	13	5	10	4
34	---	---	4	2
35	9	6	8	3
36	---	11	2	3
37	----	12	21	9
38	----	15	14	5
39	---	---	14	6

40	---	3	---	6
41	---	2	16	5
42	---	30	9	6
43	---	7	15	7
44	----	---	11	7
45	----	---	2	8
46	---	8	6	13
47	---	5	9	7
48	---	7	---	3
49	---	---	2	6
50	---	---	5	3

4 insincere agents					
	alpha = 0	alpha = 0.1	alpha = 0.2	alpha = 0.25	alpha = 0.3
run	ticks				
1	---	---	---	4	5
2	---	---	8	27	6
3	----	---	8	5	6
4	---	5	6	5	4
5	---	---	7	20	3

6	---	---	---	6	6
7	---	---	20	5	2
8	---	7	7	2	3
9	---	---	5	14	4
10	---	---	5	5	4
11	13	17	13	16	8
12	---	---	---	7	3
13	---	14	10	6	2
14	---	8	6	8	5
15	---	---	---	2	4
16	12	---	6	10	4
17	---	---	11	4	10
18	---	---	3	16	3
19	---	---	20	4	3
20	---	9	16	8	3
21	---	---	12	7	6
22	---	---	3	9	6
23	---	---	21	23	5
24	---	---	---	5	3
25	---	---	9	4	2

26	---	27	7	7	2
27	----	2	6	---	3
28	---	---	11	10	6
29	---	---	10	9	3
30	---	13	5	5	8
31	---	---	12	4	6
32	13	20	4	6	5
33	---	---	5	8	6
34	---	---	---	8	3
35	---	6	15	5	3
36	---	---	4	5	3
37	---	---	16	6	4
38	---	7	11	12	10
39	---	---	---	6	3
40	---	10	12	14	4
41	---	---	8	6	2
42	9	28	14	4	3
43	---	---	6	7	7
44	---	---	---	5	4
45	---	9	9	10	3

46	---	6	7	9	6
47	---	---	7	7	3
48	---	9	---	3	3
49	---	19	7	10	4
50	---	---	---	7	7

5 insincere agents					
	alpha = 0	alpha = 0.1	alpha = 0.2	alpha = 0.25	alpha = 0.3
run	ticks				
1	---	---	5	12	10
2	---	---	---	5	2
3	---	---	29	10	10
4	---	---	---	---	6
5	---	---	---	4	3
6	---	---	13	16	7
7	11	---	---	---	7
8	---	---	10	14	4
9	---	---	---	12	6
10	---	---	7	6	5
11	---	---	---	7	10

12	---	---	14	9	4
13	---	---	---	16	3
14	---	11	5	6	3
15	---	---	6	20	7
16	---	---	---	7	8
17	---	---	---	7	5
18	---	---	16	9	10
19	---	---	9	10	8
20	---	---	16	6	10
21	---	---	10	9	7
22	---	---	4	15	4
23	---	---	17	12	6
24	---	---	7	4	7
25	---	---	---	8	7
26	---	---	---	14	9
27	---	8	9	3	9
28	10	---	---	11	2
29	---	---	---	6	4
30	---	---	41	11	7
31	---	---	---	8	3

32	---	---	7	---	4
33	---	---	9	8	6
34	---	---	3	7	12
35	---	---	8	41	4
36	---	---	7	3	7
37	---	---	---	14	5
38	3	---	--	8	9
39	---	---	20	5	4
40	---	---	7	11	3
41	---	45	--	3	2
42	---	---	---	26	7
43	---	---	---	---	7
44	10	---	---	17	3
45	---	8	17	16	5
46	---	---	---	---	9
47	---	7	---	8	6
48	---	---	---	8	3
49	---	---	18	13	13
50	---	---	4	16	6

Experiment 2 - Data

condition 1 contrarian, 2 insincere sheep					
	alpha = 0	alpha = 0.1	alpha = 0.2	alpha = 0.25	alpha = 0.3
1	---	---	---	1	19
2	---	---	3	---	3
3	---	---	1	3	3
4	---	1	8	3	6
5	---	4	7	---	1
6	---	---	5	1	13
7	---	---	4	1	2
8	---	---	---	81	1
9	---	1	22	1	10
10	---	2	7	12	4
11	---	---	---	2	5
12	---	---	1	4	1
13	---	23	---	1	8
14	---	2	---	1	9

15	---	---	8	33	1
16	---	---	8	2	1
17	---	---	1	10	8
18	1	---	---	---	1
19	---	4	---	12	9
20	1	---	8	4	1
21	---	---	---	27	1
22	---	---	14	3	1
23	---	---	---	5	16
24	---	4	---	4	2
25	---	---	5	4	2
26	---	---	6	53	3
27	---	---	1	19	2
28	---	---	---	12	6
29	---	---	10	38	2
30	---	1	---	2	1
31	---	---	---	2	5
32	---	1	---	1	10
33	---	---	---	5	2
34	---	---	41	3	7

35	---	---	---	8	2
36	---	---	---	5	5
37	---	---	3	---	23
38	---	2	---	15	2
39	---	2	5	5	4
40	---	---	---	188	1
41	---	---	13	4	1
42	---	---	14	9	7
43	---	---	14	1	21
44	---	---	---	54	15
45	---	---	---	---	3
46	---	---	1	3	7
47	---	1	3	2	5
48	8	5	---	1	2
49	---	---	---	3	9
50	1	---	---	15	6

condition 1 contrarian, 3 insincere						

sheeps						
	alpha = 0	alpha = 0.1	alpha = 0.2	alpha = 0.3	alpha = 0.35	alpha = 0.37
1	---	---	---	2	2	4
2	---	---	---	2	2	3
3	---	2	---	5	4	4
4	---	---	---	2	2	9
5	---	---	---	9	13	7
6	---	---	3	---	23	6
7	---	---	15	3	14	3
8	---	---	---	6	3	2
9	---	9	21	5	2	10
10	---	---	---	3	1	3
11	---	---	---	3	4	9
12	1	---	---	---	2	2
13	---	---	---	2	4	2
14	---	3	---	3	6	3
15	---	---	---	---	1	4
16	---	---	2	6	1	4
17	---	---	2	---	16	9

18	---	---	7	10	---	2
19	---	---	---	2	5	2
20	---	---	---	57	8	2
21	8	---	---	---	5	4
22	---	---	---	3	2	10
23	---	---	---	---	2	22
24	---	---	8	5	4	4
25	---	---	---	8	2	3
26	---	---	---	2	4	2
27	---	---	---	30	6	6
28	---	---	8	10	7	3
29	---	---	5	2	8	5
30	---	---	---	---	1	7
31	---	---	5	2	4	1
32	---	---	----	3	20	3
33	---	---	2	6	2	7
34	---	---	---	2	4	2
35	--	---	---	1	3	4
36	---	---	---	4	16	32
37	---	---	---	4	3	5

38	---	---	---	----	6	3
39	---	---	---	7	6	4
40	---	---	3	4	4	149
41	---	---	---	---	3	14
42	---	---	---	22	4	4
43	---	---	2	4	8	10
44	---	---	---	7	16	6
45	---	---	---	12	4	2
46	---	---	---	4	2	31
47	---	---	---	---	11	9
48	---	---	---	2	2	3
49	---	10	---	14	5	5
50	---	26	---	21	5	5

condition 1 contrarian, 4 insincere sheep						
	alpha = 0	alpha = 0.1	alpha = 0.2	alpha = 0.3	alpha = 0.35	alpha = 0.4
1	---	---	---	3	3	5

2	---	---	---	3	3	7
3	---	---	---	4	9	2
4	---	---	15	5	14	4
5	---	---	---	---	3	4
6	---	---	---	4	3	4
7	---	---	13	---	3	5
8	---	---	10	---	4	3
9	---	---	---	7	---	4
10	---	---	---	3	---	3
11	---	---	---	6	2	4
12	---	---	---	3	3	7
13	---	---	---	---	3	11
14	---	---	---	---	5	2
15	---	---	---	6	5	9
16	---	10	---	11	4	4
17	----	---	---	9	6	11
18	---	---	---	3	4	5
19	---	---	---	53	5	3
20	---	---	---	6	5	5
21	---	---	---	---	3	3

22	---	---	---	11	10	8
23	---	---	---	---	4	4
24	---	---	---	4	10	43
25	---	---	---	---	5	37
26	---	---	---	---	3	3
27	---	---	5	---	6	3
28	---	---	22	---	3	7
29	---	---	---	3	2	6
30	---	---	---	6	10	3
31	---	---	---	5	3	5
32	---	---	---	14	9	5
33	---	---	---	7	4	3
34	---	---	---	---	6	3
35	---	---	---	10	23	3
36	---	---	---	12	---	4
37	---	---	---	---	4	4
38	---	---	---	3	3	3
39	---	---	---	5	---	3
40	---	---	---	---	4	4
41	---	6	17	---	6	15

42	---	---	---	4	4	2
43	-----	---	3	---	38	6
44	---	---	---	23	5	6
45	---	---	---	5	9	5
46	---	---	---	---	6	8
47	---	---	---	12	7	4
48	---	---	---	7	9	9
49	---	---	---	7	6	4
50	---	-----	9	4	2	3

condition 1 contrarian, 5 insincere sheep						
	alpha = 0	alpha = 0.1	alpha = 0.2	alpha = 0.3	alpha = 0.35	alpha = 0.4
1	---	---	---	---	9	6
2	---	---	---	8	5	6
3	---	---	---	6	7	5
4	---	---	---	14	8	6
5	---	---	---	---	3	3

6	---	---	---	22	3	6
7	---	---	---	8	3	7
8	---	14	---	---	8	13
9	---	---	17	---	3	3
10	---	7	---	9	5	11
11	---	---	---	---	7	5
12	---	---	---	11	8	4
13	---	---	---	5	5	4
14	---	---	---	11	3	3
15	---	---	---	22	6	6
16	---	10	---	4	---	5
17	---	---	---	---	9	4
18	---	---	8	3	7	4
19	---	---	---	---	---	6
20	---	---	---	5	6	5
21	---	---	---	5	---	8
22	---	---	---	---	3	6
23	---	---	---	9	2	3
24	---	---	---	11	10	8
25	---	---	---	9	4	10

26	---	---	---	5	6	19
27	---	---	---	11	5	4
28	---	---	---	6	13	3
29	---	---	---	8	6	4
30	---	---	---	---	9	3
31	---	---	---	7	15	7
32	---	---	6	15	20	2
33	---	---	---	3	9	4
34	---	---	---	4	4	10
35	---	---	---	5	43	3
36	---	---	---	7	3	10
37	---	---	---	5	4	3
38	---	---	---	---	3	5
39	---	9	---	---	5	4
40	---	---	---	11	5	6
41	---	---	---	7	4	4
42	---	---	---	---	4	4
43	---	---	---	---	3	6
44	---	---	---	4	4	4
45	---	24	5	---	3	4

46	---	---	12	11	2	6
47	---	---	---	11	8	9
48	---	---	---	7	7	2
49	---	---	4	7	3	7
50	---	---	---	---	4	3

.