The Reliability of Scientific Communities: a Logical Analysis

MSc Thesis *(Afstudeerscriptie)*

written by

Hanna Sofie van Lee
(born July 12, 1991 in Nieuwegein, Netherlands)

under the supervision of **Dr Sonja Smets**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
September 28, 2015	Dr Jakub Szymanik (chair)
	Dr Roberto Ciuni
	Prof Dr Vincent F. Hendricks
	Prof Dr Fenrong Liu
	Dr Sonja Smets

Institute for Logic, Language and Computation

Abstract

In the history of science, it has often occured that an entire community of scientists believes in a theory that is later proven to be wrong. For example, in 1915, Einstein and De Haas published a paper on the *Einstein-De Haas effect*. During the years after, experimental results showing that the effect was incorrect were ignored by the scientists in their field. Only ten years later it got accepted by the entire community that the results of the Einstein-De Haas experiment were false. There are many possible explanations for such a collective failure of a scientific community. Bayesian analyses of Kevin Zollman suggest that specific network structures can repair false beliefs more easily than others, and that varying the weights of beliefs (i.e., ensure the diversity of opinions) can also positively affect the reliability of scientific communities.

This thesis investigates the truth-tracking abilities of scientific communities from a logical perspective such that it can highlight the higher-order reasoning abilities of agents. The thesis starts with a contribution to the most relevant philosophical debates on truth and the social dimensions of science and knowledge. Then, a summary of other research on the relationship between the network of epistemic communities and their truth-tracking abilities will be given. Next, a Multi-agent Dynamic Evidence-based Logic will be introduced and it will be shown how to apply this to analyse the subjects under study. The final part of the thesis gives an overview of different conclusions that a logical analysis can give on the reliability of scientific communities. The main conclusion of this thesis is that the truth-tracking ability of scientific communities is greatly affected by the distributions of the bias evidence and distribution of the failures of the experiments. In fact, in the settings of this thesis, these distributions affect the behaviour of the agents more dominantly than the structure of the network or the weights of the bias evidence do.

Acknowledgements

I would like to thank some people who made it possible for me to write this thesis. I want to thank my *Super* supervisor Sonja Smets. By introducing me to the topic of social epistemology during a project in January you helped me to discover a great field. During the first months of the research you were a constant support and helped me to properly direct my attention. Especially during the last week you have been the greatest help I could wish for. I realise that you have put a lot of effort in improving my thesis. Thank you, Sonja! Further I want to thank the members of the committee for their time and their interest in my thesis. And thank you to my friends of the Master of Logic at the UvA. I have learned a lot from you.

Furthermore I would like to thank my parents who encouraged me to study logic and always believed in me. Thanks mom, for helping me to write this thesis by patiently correcting my work, today and during all my studies. Thanks dad for showing me the book *Logicomix* in the first place. And to my big sister and brother I want to say: Lotte, you are great example of how to work hard and simultaneously live a happy life! Arthur, although you were away during most of my Master time, you have often given me good advice through Skype. Thank you both.s

Finally I am grateful for the support of my friends in Utrecht (special thanks to my roomies Nynke en Tim) and my fellow members of the board of USVV Odysseus '91 (Matthijs, Ernst-Jan, Elke, Willem, Nard, Roel and Linde) for forgiving me to put all my focus on my thesis instead of on you guys. Last but not least, Tim: sometimes annoyed by the time I spent on my studies, but mostly proud of what I do, having you by my side makes me the happiest girl I can be. Thank you.

"Parrots mimic their owners.
Their owners consider that a sign of intelligence."

- Marty Rubin

Contents

Chapter 1

Introduction

In the history of science, it has often occured that an entire community of scientists truthfully believes in a theory that is later proven to be wrong. For example, during the 1910s, Einstein and De Haas published a paper on the nature of magnetism. For a long time, everyone in their field believed that the results of the Einstein-De Haas experiment were correct and experimental results showing that the effect was incorrect were ignored. It was not until the 1920s that other scientists publicly argued that Einstein and De Haas's main claim was false and that a new theory on the nature of magnetism got accepted by the community. There are many possible explanations for such a collective failure of a scientific community: for example insufficient expertise (a good method of experiment was not available yet), social bias (the high status of Einstein could have mislead other scientists in the group), or money and pressure (the community could have been bothered by political or financial issues). Since the main goal of a scientific community is to track the truth, in a scientific environment it is crucial to use a reliable working method that allows the community to properly combine different pieces of evidence and be resistent to false derivations.

In this thesis, I will study the phenomenon of *social proof* in scientific communities, by analysing how different factors affect the group interactions. Such a study can focus on the social constructs or on the psychological and biological mechanisms behind human behavior, typically being based on empirical data. One can also look at group behavior from a more abstract perspective, for example one can formalize economic reasoning using logic or math. Any study on the behavior of scientific communities can be enriched by discussions from philosophy of science. In this thesis, logic and philosophy of science will be the main disciplines that are used to study the interactions within epistemic communities. I will first build a philosophical framework to stipulate the problems that surface in communities engaged in scientific research and communication. In specific, I will discuss two case-studies that provide the input and guidelines for the features I will investigate later. Additionaly, using a new multi-agent version of evidence-based logic (such as Justification Logic, [1]), we can see how different factors have an effect on the social interaction and decisions of the group. Note that I will use examples from natural science, as opposed to social science and formal science. It is important to emphasise this, because theories in social sciences are typically presented as being less definite than the 'laws' of natural science, and theories in formal science are never derived from

or tested by experiment, unlike those in natural science. I do use tools from formal science (i.e., logic) and ideas from social science (i.e., those of social epistemology) in this research, but the notions of 'experiment' and 'theory' regard those of natural sciences.

Research by Bala and Goyal in [3, 4] and by Zollman in [41, 42, 43] has shown that the network structure of an epistemic community and the strenght of the beliefs of individuals can affect the truth-tracking ability of the group. Some network structures and some behaviors can repair wrong beliefs more easily than others. These analyses use simple Bayesian models to analyse the agents' behaviors. However, as argued by Baltag et al. in [5], the agents' higher-order reasoning is not explicitly modelled in a Bayesian model. A multi-agent epistemic logic does allow agents to reason about higher-order phenomena such as other agents' minds. This thesis analyses the truth-tracking abilities of scientific communities from a logical perspective such that it can also shed light on the higher-order reasoning abilities of agents. To capture the motivation behind people's beliefs, i.e., their *justification*, I need an evidence-based logic. Unfortunately there does not yet exist a logic that includes both multi-agents and evidence management and reflects on the social structure of a group of agents. Therefore, I will combine tools of various epistemic logics to uncover the formal structure of group behavior. I will adjust the existing *Logic for Dynamic Justified Belief* as introduced in [9] to construct a multi-agent model that manages and compares all available evidence. By focussing on the semantics instead of on a the complete set of axioms, this thesis will provide models of specific situations but will not contain a presentation of a complete logical system. However, I have good reasons to believe that it will be possible in future research to transform the current logic into a well-designed system and prove that it is sound and complete. With the help of Kripke models from the new logic, I wish to learn which conditions can help to make scientific communities less susceptible to epistemic errors. For example, I will compare different network structures and vary the strenghts of agents' prior beliefs. Note that even when I will be focussing on social groups of scientists, I will not study the group knowledge (as defined in classical multi-agent epistemic logic) but rather how individual attitudes such as knowledge and beliefs are based on evidence and influenced by their neighbours.

The research in this thesis touches up the side of social epistemology, which plays a role in the redesign of epistemic institutions to improve their truth-tracking ability. Today, this topic has become even more relevant since the Internet has amplified the problems of the irrational behavior of groups due to easy and wide-spread information exchange. The more data we collect, the more complex it is to organise, process and format all the information [21, p. 8]. A logical analysis will give new insights into the results that have been presented by Bala and Goyal in [3, 4] and Zollman in [41, 42, 43], where the problem is approached more from an economical or mild-philosophical fashion by using Baysian reasoning and where the conclusions are based on a large number of trials and do not concern the details of adoption behaviour.

In chapter 2 I will describe the philosophical foundations on which the thesis is built. In chapter 3, I will summarise the current state of affairs of research on net-

work structures of epistemic communities. In specific, I will discuss Zollman's claims on the ideal settings for scientific communities. Further, I will briefly study the relevant existing logics and introduce the new Multi-agent Dynamic Evidence-based Logic in chapter 4. Consequently, in chapter 5 I will use this new logic to study the effects of network structure and epistemic behaviour of scientific communities. Finally, in chapter 6 I will summarise my findings and discuss how the logical model can be eleborated and generalised.

Chapter 2

Philosophical Framework

Before I start studying the formal dynamics of network structures, let me first set up the philosophical framework. There are three relevant (interrelated) topics that I will now discuss: the objectivity of knowledge, the relation between theory and experiment and the social dimensions of scientific knowledge. It would be beyond the scope of this thesis to include a complete discussion on each of these topics including all arguments for and against, so I will be brief. For a more complete overview of the philosophical debates, I refer the reader to the *Stanford Encyclopedia of Philosophy*-pages on social epistemology ([20]) and the social dimension of scientific knowledge ([26]). To illustrate the philosophical claims, I will describe two case studies: the discovery of the weak neutral current in the 1970s and the Einstein-De Haas experiment in 1914.

2.1 The objectivity of knowledge

One of the largest and oldest debates in philosophy concerns truth. Realists, for example, argue that the world exists objectively, i.e., independent from any observer. Believing in the cumulative character of science, realists aim for developing new theories that are improvements of old ones. For a realist, the experiment will reveal only the observable part of reality. Existing non-observables are there but might not be testable for a realist. Anti-realists, on the other hand, do not aim for describing objective mind-independent reality and put more focus on experiments than realists. For example, Van Fraassen's constructive empiricism holds that science aims to give us theories which are empirically adequate, i.e., describe and explain empirical findings ([28]). Some anti-realists argue that truth is relative to time and context. For example, Kuhn argues that science evolves through so called *paradigm* shifts: a set of concepts that constitutes all true theories, research methods, postulates, etc. of a specific domain at a certain time span ([23]).

My contribution to this debate is compromising: there might be an absolute truth (in natural science), but we can rarely be sure that we have reached it. Besides all true *a priori* propositions like 'all bachelors are unmarried', there are some *a posteriori* propositions of whose truth we can be certain, e.g. propositions of the form 'Bob gives Alice a bouquet of flowers'. However, most theories in natural science are *synthetic* and universal, i.e. their truth is derived from experiment while they claim to hold for every execution of the experiment that will ever be done.

To use the results of a set of experiments and derive a universal statement such as a scientific theory, we must use the principles of induction. Since the justification of induction requires induction, we may not derive irrevocable universal statements from experiments.[1] Hence, we can never be completely certain that we have found the absolute truth after conducting a scientific experiment. Fortunately, in this thesis I will look at fictive scientific research from a meta-perspective. Assuming that there is an absolute truth, from this perspective I can distinguish the true theory from the false. In the next section I will explain in more detail in what sense theory and experiment are related.

2.2 The relation between theory and experiment

Experiment is an essential feature of natural science. A theoretic claim is perceived as more convincing when supported by experimental results. Naive scientists treat discovery as an objective observation of the world, made with unproblematic and transparant experimental techniques. Moreover, they treat experiment as being independent from theory. In this light, it is often believed that "experiment tests theory". As argued in [29], this is no longer a tenable philosophical position. Most philosophers of science agree that there is a complex and farreaching interrelation between theory and experiment.

The history of the discovery of the weak neutral current in the 1970s clearly demonstrates the interrelation between theory and experiment. From the 1960s untill 1971, both theorists and experimenters did not believe in the existence of the weak neutral current. Before 1971, a bubble chamber called Gargamelle already gave the first empirical evidence for the existence of the weak neutral current, but in this time there were enough theoretical counterarguments to reject this evidence and ascribe the neutral current candidates to neutron background. Another experiment using different techniques also failed to convince theorists or experimenters to believe in the existence of weak neutral currents. It was only in 1971 that, under a different interpretation, these experiments were used to actually confirm the existence of the weak neutral currents.

In mid 1971, a proof of the renormalisability of gauge field theories was given. This means that with the use of sophisticated mathematical techniques, sensible approximate calculations were carried out. Accepting this proof, gauge theorists had to believe in the existence of the weak neutral current. The experimenters, however, were not yet able to show that these neutral currents existed. By adjusting their beliefs to fit the theoretical expectations, experimenters interpreted their results in a new fashion. This lead to the first item of empirical support for a class of quantum field theories, gauge theories, in mid 1973. Given the opportunities its

[1] In short, that is because by definition the only way to justify induction, is to derive from all *individual* cases of succesful induction that induction *always* works (if we would be able to justify induction without moving from individual cases to universal statements, it would be called *deduction*). To conclude that induction *in general* is a legitimate method of proof requires that exact same principle of induction that we are trying to justify. This falls down to *begging the question*, which is an invalid method of proof. For the complete description of the problems of induction, please read [40].

existence offered for future experimental and theoretical practise, in [29] it is assumed that particle physicists accepted the existence of the neutral current for the social-desirable outcome.[2] This example proves how experiments are not passive and objective observations, but mouldable by the accepted theories of their time. The other way around, the outcome of experiments generally affects the focus and choices of theorists. Hence, one should no longer claim that experiment independently tests theory, but admit that there exists an interrelation between experiment and theory.

A more formal argument that demonstrates the interrelation of theory and experiment is given in [14]. It is argued that it is extremely hard to do a good experiment. In fact, uncertainty about ability is an inevitable feature of doing experiment, which leads to the *Experimenters' Regress*. When there is an accepted theory, we can judge whether an experiment failed or succeeded: the experiment succeeded when the results match the theory, and the experiment failed when there is a discrepancy between the results and the theory. In the last case, the experimenter is accused of lack of expertise or failure of apparatus. However, when there is not yet one accepted theory, we cannot tell when the experiment is properly carried out, i.e., we have no theory to compare it to. Because of this mutual dependency, new and disputed areas must inevitably resort to subjective factors, such as competence of the experimenters themselves. This makes science part of the cultural world rather than standing outside it. In the next section we will see how this cultural world has an effect on scientific knowledge.

2.3 Social dimensions of scientific knowledge

The above mentioned influence of theory on experiment suggests that researchers are biased. There are other social dimensions that influence doxastic choices, such as perception, memory, reasoning or introspection ([19]). In the light of this thesis it is important to realise what social factors can influence the beliefs of scientists, because my aim is to be able to construct a context that increases the chances to repair false beliefs.

In [42], Zollman refers to Kuhn, ([24]) noting that if there would be an algorithm at hand to get the best out of experiment and find the true theory, then all conforming scientists would make the same decision at the same time and we would not have disagreements amongst scientists. However, there is no algorithm and there are often disagreements between scientists. Besides disagreements during scientific revolutions as described by Kuhn's paradigm shifts in [23], such disagreements also occur within one paradigm. In both cases, the disagreement can be due to the fact that science is conducted by humans, who are never independent of their judgments, experience, skills, etc.

According to [26], there is more attention for the social impact on science since 1980. Contextual empiricists argue that the cognitive process that determines knowledge is a social product. Agreeing with this position, I must take into account that

[2] All of the above details on the discovery of the weak neutral current are extracted from [29].

scientists are subject to psychological mechanisms that influence their work, e.g. greed to fraud, personal and national loyalties, devotion to political causes or moral judgements, gender and financial interests. As a result, scientsts may unconsciously and in some cases even consciously overlook crucial factors that greatly affect their labresults.

When scientists work together on projects (e.g. in the cases of multiple authorship or peer reviews), the social influence becomes even more apparant. In [23], Kuhn argues that we need social factors to settle disputes between competing theories or paradigms. Factors such as deliberation, (mis)communication, testimony and (dis)trust become essential aspects of knowledge. From a reductionistic perspective, we should use observation, memory and induction to judge testimony. From an antireductionistic view, one is justified in trusting someone's testimony without prior knowledge about the testifier's sincerity. Furthermore, we can distinguish the constitutive impact on epistemic outcomes, i.e., the meaning of justifiedness of beliefs can depend on local norms of an epistemic system. In [20] we read how some famous philosophers think one should deal with these aspects. Hume, for example, believes that only with adequate reasons based on personal observations one may rely on factual statements of others. And Locke too, has strong doubts about giving authority to the opinion of others. In [26] we read that Mill, who argues that knowledge is best achieved after critical interaction between scientists, and Peirce, who says that truth is beyond reach of any individual thus critical interaction is needed to approach truth, do support deliberation anyhow. I will try to find out what position to take in this debate in my logical analysis of networks of scientific communities in the subsequent chapters.

I argue that science is a social product. On the one hand this means that experimental results can simply be wrong, because the scientists conducting the experiments are no perfect robots but social and subjective beings. On the other hand, outcome of research is also influenced by the interaction between scientists. I did not come with empirical data to prove these claims; I solely argue that we cannot deny that scientific knowledge is affected by social dimensions. To what extend exactly this happens goes beyond the scope of this thesis.[3] The following case study will demonstrate some elementary effects on scientific knowledge.

2.4 The Einstein-De Haas experiment

I will now describe the event of the "discovery" of the Einstein-De Haas effect to show how inapt communication between scientists and social dimensions such as status can lead to undesirable outcomes.[4]

[3] I believe that it would go beyond the scope of this thesis to include empirical evidence, because I assume it will be impossible for anyone to claim that scientific knowledge, which is a cultural product, is *not* under influence of social factors. If one would argue that scientific knowledge is not a cultural product, then I suppose that he or she refers to a different kind of knowledge; not the one that is presented in papers and books, but a knowledge that then apparently exists independently of us. To be clear: in this thesis I speak about the scientific knowledge that is discovered, believed in and presented by human beings, i.e., in the cultural world.

[4] All of the details on the history of the Einstein-De Haas effect are extracted from [17, 16, **?**].

Firstly, let's describe the context. During the 1910s, Einstein and De Haas wanted to empirically test Ampère's hypothesis, who claimed in 1820 that magnetism is caused by circulation of electric charge. The fact that Einstein wanted to empirically test something deserves some attention, since Einstein is known for his disapproval of experiment. In fact, he could be very stubbornly convinced of a theory even if empirical data seemed to falsify the theory. Against all the odds, in 1914 Einstein started his only experimental work ever published. By that time, Einstein had already built up quite an impressive reputation, which minimized the distrust of other scientists to his claims.

Secondly, let's see what happened during and after the "discovery" of the Einstein-De Haas effect. Einstein and De Haas wanted to investigate the nature of magnetism and intented to show that the spin of a magnetic momentum is of the same nature as the spin of rotating bodies in classical mechanics. They predicted a so called *gyromagnetic ration* of 1.0. Experiments of Einstein and De Haas showed that $g = 1.02$ and $g = 1.45$. Next, Einstein and De Haas discarded the result of $g = 1.45$ (which mismatched Ampère's hypothesis) and published in the spring of 1915 that $g = 1.02$, claiming that experiment approximately confirms Ampère's theory. Their paper did include an elaborate description and discussion of the experimental setup and an analysis of possible errors and ways to overcome these. While others later repeated the experiment and got values around $g = 2$, Einstein insisted that $g = 1$. It wasn't until the 1920s that other scientists published that the Einstein and De Haas were wrong and that the correct value of 2.0 got accepted.

Thirdly, let's analyse what went wrong during and after the Einstein-De Haas experiment. A crucial mistake is made by Einstein and De Haas themselves. In their paper, they did not share their anomale result that $g = 1.45$. Furthermore, Einstein and De Haas were too priored at the start of the experiment because they were strongly committed to the theory. The desire to prove the theory was strong, because there were a lot of related problems that could be explained with a gyromagnetic ration of 1.0. This clearly affected their treatment of the data. Besides this mistake of Einstein and De Haas, their colleague-experimenters could also have been more critical. Because of Einstein's fame, the results of other experimenters got overshadowed by the publication of the Einstein-De Haas effect. Finally, here too we see the influence of the interrelation of theory and experiment. Earlier, Barnett (in 1909) and Maxwell (in 1861) did some experiments on the subject that conflicted with Ampère's theory. However, they lacked crucial theory on currents and electrons to properly interpet and design the experiment.

Note that the case of the Einstein-De Haas effect is not representative for science; such collective faults seem to occur only rarely. However, we should still try to prevent such faults. It seems that sharing only belief and keeping some evidence private, as Einstein and De Haas did, can lead to epistemic group failure. Likewise, we see that priors should not be too high because they might be based on false assumptions while preventing scientists from switching to another belief.

Chapter 3

Theoretical Framework

Now that the philosophical framework on science and its practitioners in general is set, I can start to focus on the effects of the network on the reliability of epistemic communities. In this chapter, I will discuss several relevant studies on information control problems among deliberating agents.

3.1 Irrational behavior of groups

Groups might seem to have an epistemic advantage over individuals, because they have access to more information, but they are at the same time very vulnerable to irrational collective behavior. In [34], the problems of deliberating groups are discussed. Ideally, a deliberating group would show the following principles: the best members pull the others to their level of expertise, the information of all group members is combined and group discussion creates extra insights. In practise we see something different: group members tend to become more confident of their judgments after they speak with one another ("amplification of cognitive errors"), groups usually get to the level of their average members and people with extreme views tend to have more confidence that they are right and as people gain confidence, they become more extreme in their beliefs ("group polarization"). Exposure to the views of others might lead people to silence themselve for two reasons: i) *informational pressure*, i.e., strong new informational signals contradict and outweigh private signals, and ii) *social influence*, i.e., people do not want to be different from the rest.

Well-studied phenomena of irrational behavior of groups include *informational cascades*, *pluralistic ignorance* and the *bystander-effect*, for example in [21]. People in a network can influence each other's behavior and decisions. An *informational cascade* occurs when it is optimal for the individuals of a group to follow the behavior of the crowd whilst ignoring their private evidence, because the information they get from the crowd outweighs their private information ([13]). We speak of a false cascade when this leads to a false group belief. Hence in false informational cascades the agents' behavior is individually rational, but irrational for the group. Such informational cascades can occur easily, but they can fortunately also easily be broken, for example when an individual with hard (true) information appears. When people go along with the crowd in order to maintain the appreciation of others, we speak of a *reputational cascade*.

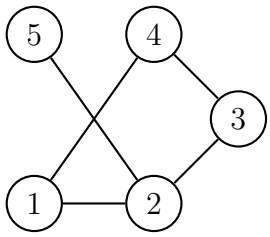

Figure 3.1: *A network graph with 5 nodes, labelled '1','2','3','4','5' and edges between the pairs (1,2),(1,4),(2,3),(2,5) and (3,4). The network is "connected", because there is a path going between every pair in the network.*

One way to model the behavior of people in an informational cascade is to use *Bayesian probabilities* and *network theory*. With Bayesian reasoning, we can determine the probabilities of events given the information that is observed or obtained by communication. For the probability of event A I write $Pr[A]$. For the probability of A given that B has occured I write $Pr[A|B]$. *Bayes' rule* states that

$$Pr[A|B] = \frac{Pr[A] \times Pr[B|A]}{Pr[B]}$$

We can use Bayes' rule for example to detect email spam.

A *network graph* (see Figure 3.1) consists of a set of objects, called *nodes*, with certain pairs of these objects connected by links called *edges*. For example, the World Wide Web is an enourmous information network with nodes being webpages and the edges are links leading from one page to another. For the purpose of this thesis, nodes will represent the agents and undirected edges will represent the communication between agents. The fact that the edges are undirected implies that communication is always *symmetric*, flowing two-ways. Furthermore, we say that two agents are *friends*, or neighbors, if they are connected by an edge. A *path* is a sequence of nodes such that each consecutive pair in the sequence is connected by an edge. In a *connected network*, every pair of agents is connected by a path.

A fundamental feature of a network setting is that we evaluate the actions of agents not in isolation, but with the expectation that the world will react to what any agent does.

Let an agent's choice between strategy A or B be based on the choices made by all of her friends. Consider any network and suppose everyone in the network has chosen B. Then let some initial adopters switch to A. If their direct friends copy their behavior, making *their* friends to adopt *their* behavior, a cascade has formed. When everyone in the network switches, we speak of a *complete cascade*. It can also happen that the cascade stops before everyone has switched. This depends on the structure of the network, specifically on the density of cluster. A *cluster of density x* is a set of nodes such that each node in the set has at least a fraction x of its network friends in the set. For example, the set of nodes 1,2,3,4 forms a cluster of density $\frac{2}{3}$ in the network in Figure 3.2. Now if the remaining network (those that did not yet

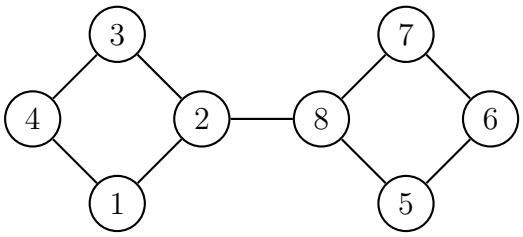

Figure 3.2: *A network graph with two clusters of density* $\frac{2}{3}$

switch) contains a cluster of density greater than $1 - q$, for q being the threshold, then the set of initial adopters will not cause a complete cascade. Whenever a set of initial adopters does not cause a complete cascade with threshold q, the remaining network must contain a cluster of density greater than $1 - q$ ([15, ch.19]).

It can also happen that the people in the network do not know who chose option A, for example when it is forbidden to talk about it. It can happen that everyone in the network wants to switch to A, but does not do it because they do not want to be the only one. We call this *pluralistic ignorance*. A special case of this is the *bystander effect*, expressing that the more individuals who are gathered in one place, the less the likelihood of people coming to the aid of a person in need. The observation of others' lack of action may lead one to believe that there is no reason to take action ([21, p.23]).

3.2 The effect of the network structure

An important paper on the effect on the process of social learning of network structure, i.e., the specific configuration of how evidence flows in a community, is about a technical investigation performed in 1998 by the two economists Bala and Goyal ([3, 4]). The authors consider an infinite society whose members face a decision problem: to choose an action at regular intervals without knowing the true payoffs from other actions. The agents use their experience along with the experience of their friends to upgrade their beliefs. Given these beliefs, each agent repeatedly chooses an action that maximises the expected utility. It is argued that humans cannot process complex calculations that include reasoning about unobserved agents (friends of friends) and therefore an analysis that relies on the agents' limited rationality (omiting higher-order reasoning abilities) is more realistic.

Bala and Goyal show that in a connected network agents' beliefs necessarily converge to a limit and that these limits are equal for all agents in a connected society. This implies that in the long run, all agents in a connected network have the same belief, which is called *social conformism*. Whether or not this action is optimal, depends on the distribution of prior beliefs, the structure of neighborhoods and the informativeness of all actions. Bala and Goyal develop conditions that ensure optimal choices. They consider agents arranged on a line where each agent can only communicate with those agents to the immediate left and right of them. If there is an infinite number of agents, convergence in this model is guaranteed so long as the agent's priors obey some mild assumptions. They also consider adding a special

group of individuals to this model, a 'royal family'. The members of the royal family are connected to every individual in the model. For this network structure, the probability of converging the the wrong result is no longer zero. Negative results obtained by the royal family infect the entire network and mislead every individual. It is claimed that the conclusions are consistent with empirical findings ([3, sec.5]).

Kevin Zollman analysed in further detail Bala and Goyal's counterintuitive result that in some contexts a weakly connected community is more reliable than a highly connected community in [41, 42, 43, 44]. Zollman works with models of finite groups instead of infinite groups, which is closer to real-science than Bala and Goyal's infinite model. As in Bala and Goyal's models, Zollman considers situations called *Bandit problems* where the agents are faced with a dilemma to gain information and meanwhile get the highest payoff. Suppose there are two medicines, medicine A and medicine B, and each agent believes that either A or B has the best healing power. The payoff of the old medicine A is known by every agent and the payoff of B, the new medicine, is unknown. The agents' beliefs determine their actions: all agents believing that A is superior will use medicine A on their patients and all agents believing in B will use medicine B on their patients. Agents want to cure their patients, so it would be irrational to test the inferior medicine.[5] Note that the incoming evidence depends on the actions of the agents. Moreover, learning demands communication: the believers of the old medicine A need the evidence of agents using the opposite medicine in order to compare the two payoffs and if neccessary switch to the new medicine. Zollman uses computer simulations to compare three different networks: the cycle, the wheel and the complete graph (see Figure 3.3) and different strenghts of prior beliefs.

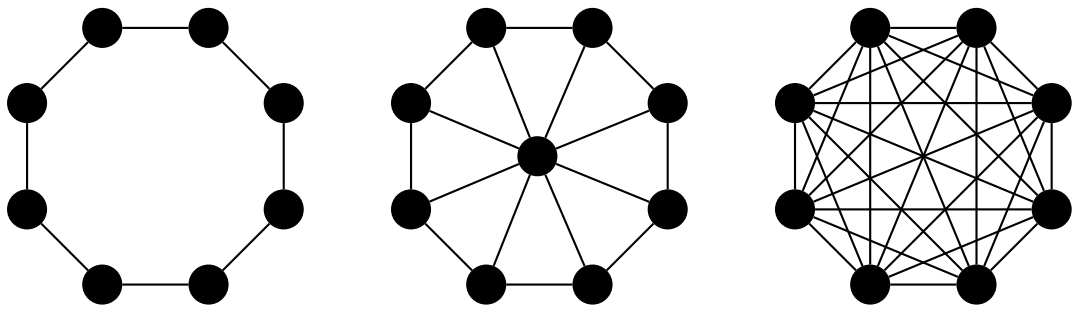

Figure 3.3: *An 8 person-circle, 9 person-wheel and 8 person-complete graph*

I will now sum up the most important conclusions from Zollman's work and give a short explanation of each of these. In [41] the conclusions are that:

i) in some contexts, scientific communities with less connections are more reliable than communities with more connections, and

[5]This is a simplified version of reallife science. It is not always irrational to test the inferior medicine, because scientists do realise that gaining information is also worth something. In any case, at some point (sometimes after n trials trying both medicines, sometimes right after the presentation of a new medicine) scientists are faced with the dilemma to get more information or choose for the superior medicine. The research for medicines for HIV, for example, was stopped before the planned amount of experiments with both medicines was conducted, because one medicine showed a successrate that was considerably much higher than the other, so it really was immoral to continue testing the inferior medicine on patients.

ii) there is a tradeoff between speed of adopting beliefs and reliability to track the truth. It depends on the epistemic goals of the community whether speed or reliability is more important.

These conclusions are explained by the fact that in less connected networks bad results and good results both spread slower, so variety is preserved longer. When variety is preserved longer, beliefs in the true theory are more likely to survive the emergence of a *false informational cascade* whereas in a highly connected community the good beliefs can disappear before they get the chance to repair the false beliefs.

So cognitive diversity, i.e., having all theories investigated by at least one agent, helps communities to choose the best action. There are two ways to achieve this, as argued in [42]:

i) by limiting the information that gets to the agents, and
ii) by implementing scientists with strong beliefs.

However, when a group holds both properties, its members will never switch belief. This is obviously a bad consequence, because if cognitive diversity is maintained indefinitely, then as a result agents fail to converge to the truth. We want *transient diversity*. Zollman uses a network graph and beta-distributions such that he can vary with the connections and priors α and β (representing the strengths of beliefs) to prove claim i) and ii).[6] These claims are illustrated by a study of the research on Peptic Ulper Disease (PUD). For a long time, people believed in the wrong theory to explain PUD (because they used the wrong method) and no one tried the other method. Zollman argues that this could have been prevented when the researchers would have taken either i) or ii) into account.

We can see the resemblance between Bandit problems and science, for two bandits (or, in the case of PUD, 'medicines') can be treated as two competing theories, such as in a scientific revolution as described by Kuhn. The reward for the doctors in [42] is to cure patients from PUD, and the reward for scientists in general is to develop a true theory. There are some problems that arise when we want to use logic to analyse Bandit problems, though. I will discuss these in section 5.1.1.

Zollman argues that division of labor improves the truth-tracking ability of the group. If that is achieved, then belief in good methods and theories persist longer, such that these can repair the bad results. Information about a theory or method can only be gathered by scientists actively pursuing it. Since the effort for developing an inferior theory is often regarded as a waste, we want to give scientists some interest in pursuing the inferior theory in order to divide the cognitive labor. In the next section we will see how this can be done (and that we are already doing this in western science).

[6] A beta-distribution uses Bayesian reasoning for complex probabilistic predictions. It is a function that represents an agent's belief over infinitely many hypotheses - values of α and β. Learning via beta distributions is relatively efficient, because agents directly learn after every update.

3.3 Cognitive division of labor

Because of the mismatch between individual rationality (i.e., persuing the superior theory) and collective rationality (i.e., cognitive division in labor in order to repair false beliefs in the group), we need to give scientists individual reasons to purchase collective rationality. In [22], Kitcher argues that a good scientist makes individual rational choices when she belongs to a community in which the chances of discovering the correct answer are maximised. Good scientists should agree in advance that it may sometimes be necessary for some to persue an inferior theory, and that it may fall to her to play this role. We cannot simply force scientists to try the inferior theory, but we must do it indirectly by promoting the investigation of a new method.

In [33], Strevens claims that our current reward system actually leads to cognitive division of labor. That is, because scientists are being rewarded for being the first to discover something. This reward system, reward being prestige (power, credibility, quotations), follows the *priority rule* (reward in the sense of salary is rewarded to all scientist that are employed at a university or other scientific institutions). Hence our reward system affects the behavior of scientists desirably: it stimulates the cognitive division of labor. Strevens claims that the priority rule has always and everywhere ruled in Western science.

3.4 The Independence Thesis

Often, science is depicted as done by isolated scientists, while scientists are always part of some larger community. My emphasis to look at the group as a whole instead of isolated individuals, is motivated by the claim in [27] that rationality of individuals and rationality of groups are independent properties of groups. Henceforth, we should consider the rationality of individuals as well as the rationality of groups when analysing social knowledge. Note that even though Bala and Goyal and Zollman's Bayesian models use the input of the network graph, and thereby the relations between the entire community, the calculations themselve are restricted to one individual.

Chapter 4

Logical Model

Now that I have built the philosophical framework and studied the most relevant work on the interaction within (scientific) communities, I can start with the logical analysis. There have been Bayesian analyses on the effect on the epistemic achievements of groups of the network structure, i.e., the specific configuration of how evidence flows in a community. For each individual agent, the authors in [3, 4, 41, 42] count the data of both the agent's own observation and the testimony of others, to calculate with the Bayesian Law and beta-distributions which theory the agent should regard as most plausible. Believing in this chosen theory apparently should give the highest payoff, henceforth the agent should behave as if that theory is indeed the true theory, by designing and interpreting her experiments in the light of the selected theory. The approaches in [3, 4, 41, 42] incorporate shared information on experimental results that agents receive from their friends in the network and omit reasoning about other agents' minds (e.g. "my friend b knows that all of her friends believe p, and since she has a lot of friends, I should regard her behavior as more informative than the behavior of my lonely friend c") and other higher-order reasoning powers of the agents (e.g., realising the network structure in general). Analysing the effects of a specific network configuration on the behavior of agents by using a system that includes higher-order reasoning, can shed a new light on the behavior of scientists.

Logic provides the tools and techniques to reason about the higher-order processes in the agents' minds. Since there are many different logics, each constructed for specific objectives, I will first have to choose the particular logic(s) I want to work with. There are a couple of tasks I want my logic to be able to do, such that I can analyse the truth-tracking power of scientific communities. Most importantly, I need a *Kripke model* and a language with *epistemic operators* K and B such that I can model different states of the world and agents' knowledge and belief about these states. In addition I want a *multi-agent logic*, such that besides modelling the agents' uncertainty about atomic facts, I can also gain insight on the mutual uncertainty about other agents' knowledge and beliefs. Furthermore, I want to see how agents justify their knowledge and beliefs, so I need *evidence-managing* tools. Since I will simulate a dynamic context, where agents update their beliefs, knowledge and evidence, I need a *dynamic logic* to model actions and a *temporal relation*. In the philosophical and theoretical framework I have learnt about some factors that can have an effect on the epistemic achievements of the group. Firstly, one

of the most striking results from [41, 42] is that there is a trade-off between the speed at which beliefs spread in a community and the truth-tracking ability that is caused by the network structure. Therefore, I want to include the *network structure* to describe who communicates with whom. Secondly, Zollman shows us in [41, 42] that the strenght of prior beliefs has an effect on the adopting behavior of the agents, so I need to be able to adjust the weights of the agents' priors, i.e., their *biasses*. Thirdly, from the Einstein-De Haas debacle I learnt that it also matters whát the agents communicate: so I wish to have flexible techniques for *sharing data*.

This is quite a list of desiderata, but fortunately there are some logics that are good candidates to handle this list. However, none of them are good enough to capture the entire list. For example, Justification Logic (JL) provides techniques to input evidence and justification, but only in a static situation. Standard Dynamic Epistemic Logic (DEL) uses dynamic models for updates, but is not refined enough to talk explicitly about evidence, justification and reliability. Classical DEL is often extended with tools from Belief Revision Theory (BR) for dealing with fallible evidence and "soft" information. The logic presented by the authors in [9] combines these three logics into one logic, the Logic of Dynamic Justified Belief (DJB). Unfortunately, this logic is only suitable for single-agent models. Therefore I will adjust DJB such that it can produce a multi-agent model. Besides this adjustment, in section 4.3 I will add some other necessary tools to the logic and throw out superfluous features. With the resulting system, I can adjust variables such as communication connections, distribution of priors and weight of priors. In section 4.3.5 I will tell how one can extend the logic into a more universal system.[7] The model will have different components, including the network structure as well as the epistemic structure and evidence of individual agents. In my presentation of this model I will highlight a selection of some specific features of the global model, as the total picture can become rather complex to draw.

I will first discuss the preliminaries, by briefly introducing DEL, BR and JL, such that I can thereafter present the relevant features of DJB in section 4.2. After that, I will describe the Multi-agent Dynamic Justification Logic (MDEL) in section 4.3, which is built up from ingredients of the former systems. Note that the situations I want to model will have all ingredients incorporated in one setting.

4.1 Preliminaries

In this section I will present preliminaries that are necessary to understand the Logic of Dynamic Justified Belief and the Multi-Agent Dynamic Evidence-based Logic, which are based on techniques from DEL, BR and JL. I will only briefly discuss the logics because the reader is expected to be familiar with propositional, first-order logic and formal definitons of truth, and because most technical details of the extended logics will be explained in the subsequent sections.

[7] In [30, 31], Renne combines DEL and JL in a multi-agent setting that allows for private communication. However, this model allows only for *deleting* evidence instead of *adding* evidence, which will be a crucial action of my analysis. Other logics that combine dynamic models with concepts of justification logic include [6, 25, 32, 38] . All of these logics could be explored in the future.

4.1.1 Dynamic Epistemic Logic

The framework of Dynamic Epistemic Logic (DEL) as presented in [12] describes how various changes such as observations by an agent or communication between the agents affect the epistemic and doxastic states of the agents. Classical DEL is not hospitable to belief revision, but in most recent literature tools for belief revision are added. For example, the author of [36] presents a dynamic logic for belief revision and the authors of [11] give a qualitative theory of dynamic interactive belief revision. Since I want a DEL that does include the possibility of upgrading beliefs, I will now introduce a *soft* version of DEL that uses tools from BR.

I use Kripke frames and models to define semantics for epistemic logics. A *Kripke frame* is a 2-tuple $F = (W, \sim)$ where W is a set of possible worlds and $\sim \subseteq W \times W$ is the *indistinghuisable relation* on W. A *Kripke model* is a 3-tuple $M = (W, R, [\![\cdot]\!])$ where $[\![\cdot]\!] : W \to \mathcal{P}(\mathscr{F})$ is a valuation map for $\mathscr{F}$, being the set of propositional formulas φ of the language. Given a set Φ of atomic sentences, a simple language $\mathscr{L}$ for DEL is defined by recursion:

$$\varphi := \bot | p | \neg\varphi | \varphi \wedge \varphi | \Box\varphi \text{ with } p \in \Phi$$

This language can be extended as we will see in the subsequent sections. In epistemic logic, $\Box\varphi$ is to be read as 'I know that φ', but this interpretation can be specified in further detail, as we will see in section 4.2.1. I use the following abbreviations:

$$\begin{aligned} \top &:= \neg\bot \\ \varphi \vee \psi &:= \neg(\neg\varphi \wedge \neg\psi) \\ \varphi \to \psi &:= \neg(\varphi \wedge \neg\psi) \end{aligned}$$

A *pointed model* is a pair (M, w) consisting of a model M and a designated world w in M called the "actual world" (or the "real world").

Definition 4.1.1. (Truth for DEL) The *satisfaction relation* $w \vDash \varphi$, short for $(M, w) \vDash \varphi$ when M is fixed, is defined as follows:

$$\begin{aligned} &w \vDash \bot && \text{never} \\ &w \vDash p && \text{iff } w \in [\![p]\!] \\ &w \vDash \neg\varphi && \text{iff } w \nvDash \varphi \\ &w \vDash \varphi \wedge \psi && \text{iff } w \vDash \varphi \text{ and } w \vDash \psi \\ &w \vDash \Box\varphi && \text{iff } v \vDash \varphi \text{ for every } vRw \end{aligned}$$

We can extend the valuation map $[\![\cdot]\!]$ to all sentences φ, by putting $[\![\varphi]\!] = \{w \in W | w \vDash \varphi\}$.

We say that 'φ is true at w in M' iff $M, w \vDash \varphi$. We say that 'φ is valid' iff φ is valid on the class of all frames.

Definition 4.1.2. (The logic K) The logic K is given by the following axiomatization:

$$\begin{aligned} \text{(Necessitation) If } &\vdash \varphi \text{ (“}\varphi \text{ is a } \textit{propositional tautology}\text{”), then } \vdash \Box\varphi \\ \text{(Modus Ponens) If } &\vdash \varphi \text{ and } \vdash \varphi \to \psi, \text{ then } \vdash \psi \\ \text{(K) If } &\vdash \Box(\varphi \to \psi), \text{ then } \vdash \Box\varphi \to \Box\psi \end{aligned}$$

Definition 4.1.3. (The logic S4) The logic S4 is obtained by adding the following axioms to K:
(4) $\vdash \Box\varphi \to \Box\Box\varphi$

Definition 4.1.4. (The logic S5) The logic S5 is obtained by adding the following axioms to K:
(5) $\vdash \neg\Box\neg\varphi \to \Box\neg\Box\neg\varphi$

The rules of S4 entail *positive introspection*: “if I know something, then I know that I know it”. The rules of S5 entail also *negative introspection*: “if I do not know something, then I know that I do not know it”.

When constructing a multi-agent Kripke model for a set of agents A, the operator $\Box$ needs an index $i \in A$ to specify who knows φ: $\Box_i\varphi$. The nice thing about using modal logic in epistemology, is that we can express sentences like “Alice knows that Bob knows that p”, i.e., $\Box_a(\Box_b p)$. We can also express that something is *common knowledge* for a set of agents G, written as $C\Box_G$. If φ is common knowledge to G, then every agent in G knows that φ and everyone knows that everyone knows φ, etc. As an example of a multi-agent epistemic model, consider Figure 4.1.

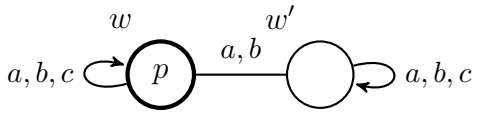

Figure 4.1: *A multi-agent epistemic model with three agents a, b and c. In the real world w, p is true. p is not true in w'. We can see in this model for example that agent a and b do not know whether p is true or not, but agent c does know that p is true (he can distinguish between w and w'). Furthermore, c knows that a and b do not know whether p. And a and b know that c knows whether p.*

If I want to model personal beliefs, I have to include another binary relation that specifies the plausibility order amongst the possible worlds, often written as $\leq_i$ and depicted by an arrow in the model. We define belief $B_i\varphi$ as truth in the most plausible worlds:

$$M, w \models B_i\varphi \text{ iff } M, w' \models \varphi \text{ for all } w' \in max_{\leq_i}\{w' \in W | w \sim_i w'\}$$

For example, consider Figure 4.2

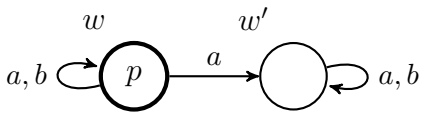

Figure 4.2: *A multi-agent epistemic model with two agents a and b. Here, agent b knows that p and agent a does not know whether p. a thinks that world w' is more plausible, hence she believes that $\neg p$. Moreover, a believes that b knows that $\neg p$.*

So far the epistemic logic is static. I do want to capture the truth conditions of statements concerning the change of knowledge and belief due to new information becoming available. A framework that can deal with this is the logic of public announcement PAL. Intuitively, a public announcement of φ removes all possible worlds where φ is false. Besides public announcements, we can also imagine *private announcements*: Alice tells Bob a secret, but not to Charlie. The framework of DEL provides a canonical way to model actions. The essential idea of action structures is that we describe actions as Kripke structures: a 2-tuple (E, R) where E is a set of events and R the equivalence relation on E. For every $\alpha \in E$ we have a formula $\mathsf{pre}(\alpha)$ which is called the precondition of α that defines when an event can happen (e.g., I can only see a unicorn if there is a unicorn). An action is a 3-tuple (E, R, α) where α should be seen as the 'actual action'. Combining the epistemic model and the event model, we get a *product update*. The product update is a partial function that maps pointed models to pointed models by an action. Please read [7, 8, 12, 39] for more details on models of product update.

4.1.2 Justification Logic

Even though DEL is an impressive and eleborate system, it cannot deal with evidence. A lot of formal and philosophical studies on the meaning of knowledge or belief are based on the criticised claim that 'knowledge' is equal to 'true justified belief'. Gettier described a few counterexamples in [18] that show that this claim is not always applicable, suggesting that there is a missing ingredient to the triple 'truth', 'justification' and 'belief'. [18] set fire to many different proposals of these missing ingredients, some of them focussing on the fact that the justification should be relevant or truthful. For example, the Defeasibility Theory defines 'knowledge' as 'true justified belief that is stable under belief revision with any new evidence'. As the authors in [9] point out, the interpretation of 'evidence' is not always clear from the context; do I have to consider *all* evidence, or only *true* information? Therefore, it is good to be very careful and explicit when we define evidence.

Justification Logic provides us with the tools for reasoning about justification and evidence ([1, 2]). JL introduces structured syntactic objects called *terms*. There are different kinds of evidence: directly observing t; testimonial evidence (given by friends), logical evidence (theorems from logic) and inferential evidence (derived by combining other pieces of evidence by Modus Ponens, managing or aggregating compound terms). JL allows us to form new formulas of the form $t :_i \varphi$: "t is agent i's justification that φ is true", and $t \gg_i \varphi$: 't is agent i's admissible evidence for φ". Justification Logic does not directly analyse what it means for t to justify φ beyond the format $t : \varphi$, but rather attempts to characterize this relation axiomatically.

The basic operation on justifications are *application* $\cdot$ and *sum* $+$. More elaborate logics introduce additional operations on justifications. The simplest justification logic J_0 is axiomatised by

(Classical Logic): All classical propositional axioms and the rule Modus Ponens
(Application): $s : (\varphi \to \psi) \to (t : \varphi \to (s \cdot t) : \psi$
(Sum): $s : \varphi \to (s + t) : \varphi$ and $s : \varphi \to (t + s) : \varphi$

4.2 The Logic of Dynamic Justified Belief

In [9], Baltag et al. introduce dynamic operations of evidence introduction, evidence-based inference, strong acceptance of new evidence and irrevocable acceptance of additional evidence. In this section I will discuss some elements of the Logic of Dynamic Justified Belief, DJB, which will be the basis of the Multi-agent Dynamic Evidence-based Logic presented in section 4.3. We will see how DEL, BR and JL are combined to construct the single-agent Logic of Dynamic Justified Belief, DJB as defined in [9, pp.2-3]. I refer the reader to Appendix A for all the details on DJB that I will not mention.

4.2.1 Syntax

Definition 4.2.1. (Language JB) Given a set Φ of atomic sentences, the language $\mathscr{L} := (\mathscr{T}, \mathscr{F})$ consists of the set $\mathscr{T}$ of *evidence terms* t and the set $\mathscr{F}$ of *propositional formulas* (sentences) φ defined by the following double recursion:

$$\varphi ::= \bot | p | \neg\varphi | \varphi \wedge \varphi | Et | t \gg \varphi | \Box\varphi | K\varphi | Y\varphi \text{ with } p \in \Phi$$

$$t ::= c_\varphi | t \cdot t | t + t$$

Subterms and subformulas are defined to construct preconditions. The operation $(\cdot)^Y$ is introduced in order to deal with the famous Moore sentence "$\varphi \wedge \neg B\varphi$". Please see Appendix A for the construction of these objects.

Explaining formulas of $\mathscr{L}$

Et says that *evidence t is available* to the agent (though not necessarily *accepted*). $t \gg \varphi$ says that *t is admissible evidence* for φ: if accepted, this evidence supports φ ("*t justifies φ*"). $\Box\varphi$ says that *the agent (implicitly) defeasibly knows φ* (rules of S4 so positive introspection). $K\varphi$ says that *the agent (implicitly) infallibly knows φ* (rules of S5, so negative introspection). And $Y\varphi$ says that *"yesterday"* (i.e.,before the last epistemic action) *φ was true.*

In [9], two different types of knowledge are defined: infallible knowledge K (absolutely unrevisable belief - even in the face of false evidence), corresponding to the principles of S5; and defeasible knowledge $\Box$ (unrevisable belief in the face of any new true information), corresponding to the principles of S4. This implies that $\Box$ does not have negative introspection, while K does. Belief is defined as $\neg\Box\neg\Box\varphi$ and is abbreviated as $B\varphi$. Note that K and B are universal operators, i.e., true independently of the possible worlds, as opposed to $\Box$ which is to be evaluated on

a specific world. The relation $\gg$ is also universal in a model M, that is, it is not defined on a specific world but any formula of the form $t \gg \varphi$ holds over the entire model.

Explaining evidence terms of $\mathscr{L}$

c_φ is an *evidential certificate*: a canonical piece of evidence in support of sentence φ. $t \cdot s$ combines two pieces of evidence t and s, using MP. $t + s$ aggregates (without performing logical inference) all evidence provided by t and s.

Definition 4.2.2. (Admissibility) *Admissibility* is the smallest binary relation $\gg \subseteq \mathscr{T} \times \mathscr{F}$ satisfying the following conditions:
(1) $c_\varphi \gg \varphi$;
(2) if $t \gg (\psi \Rightarrow \varphi)$ and $s \gg \psi$ then $(t \cdot s) \gg \varphi$; and
(3) if $t \gg \varphi$ or $s \gg \varphi$, then $(t + s) \gg \varphi$.

Definition 4.2.3. (Admissible terms) $\mathscr{T}^e := \{t \in \mathscr{T} | \exists \varphi \text{ such that } t \gg \varphi\}$ is the *set of admissible terms*.

Definition 4.2.4. (Propositional content) For every term $t \in \mathscr{T}$, the *propositional content* con_t *of* t is the conjunction of all the formulas for which t is admissible evidence: $\mathsf{con}_t := \bigwedge\{\theta | t \gg \theta\}$. For $t \notin \mathscr{T}^e$, this is the conjunction of an empty set of formulas.

Further notes

One of the objectives of [9] is to deal with the problems of *logical omniscience*. Agents are logically omniscient if they know or believe all of the logical consequences of their knowledge or beliefs. For instance, logical omniscient agents necessarily know all theorems of the logic in use. The classical interpretations of the modalities $K, \Box$ and B satisfy logical omniscience. In an ordinary sense, people do not possess such supernatural reasoning powers. By distinguishing between *implicit* and *explicit* knowledge (belief), the authors of [9] allow for non-logically omniscient agents. In JB, only *implicit* knowledge, $K\varphi$ or $\Box\varphi$, and *implicit* belief, $B\varphi$, satisfy logical omniscience. "Implicit knowledge may be thought of as "potential knowledge" of φ that the agent might in principle obtain, though perhaps she will never have this knowledge in actuality" ([9, p.8]). In other words, knowledge (belief) of φ that can be derived in the epistemic model. *Explicit* knowledge (belief) represents the agent's actual knowledge (belief), obtained when the agent realises her implicit knowledge (belief) and can verify, or reason about, the evidential certificate for φ, i.e., c_φ is in her evidence set:

$$K^e\varphi := K\varphi \wedge Ec_\varphi$$
$$\Box^e\varphi := \Box\varphi \wedge Ec_\varphi$$
$$B^e\varphi := B\varphi \wedge Ec_\varphi$$

To get a better grip on the relationship between formulas and terms, consider the following abbreviations:

- $A(t)$ is short for *(implicitly) accepting* t, i.e., when the agent believes all sentences φ for which $c_\varphi \in \mathsf{sub}(t)$) (define $\mathsf{sub}(t)$ as in Appendix A;

- $G(t)$ stands for *t is good (implicit) evidence*, i.e., the agent defeasibly knows all sentences φ for which $c_\varphi \in \mathsf{sub}(t)$;
- $I(t)$ for *t is infallible (implicit) evidence*, i.e., the agent infallibly knows all sentences φ for which $c_\varphi \in \mathsf{sub}(t)$; and
- $t : \varphi$ for *t is (implicit) evidence for believing that φ*, i.e., the agent accepts t and $t \gg \varphi$ ("t justifies φ and t is accepted as justification of φ").

Similarly as with $K^e, \Box^e$ and B^e, we can state that *t is explicit evidence for belief of φ*:

$$t :^e \varphi := t : \varphi \wedge Et$$

An important conceptual difference between evidence on the one side and knowledge and belief on the other side, is that evidence can be contradicting, while knowledge and beliefs cannot. For example, it is possible that $t \gg \varphi$ and $t' \gg \neg\varphi$, and $t \in E$ and $t' \in E$, for either t or t' can be unaccepted. It cannot, however, occur that both t and t' are accepted, for then the agent would believe φ and $\neg\varphi$, which leads to a logical contradiction.

I will not go into further detail on the subject of implicitness and explicitness, since it is not the focus of this thesis. In chapter 5 I will assume that all evidence is automatically accepted. Though in theory, as we have seen above, evidence does not need to be accepted. In that case, the consequences of t would not necessarily be believed even though $t \in E$.

4.2.2 Semantics

Definition 4.2.5. (Model for JB) A *model* $M = (W, [\![\cdot]\!], \sim, \geq, \rightsquigarrow, E)$ is a structure consisting of a nonempty set W of *possible worlds*; a *valuation map* $[\![\cdot]\!] : \Phi \to \mathcal{P}(W)$; binary relations $\sim$ ("epistemically indistinguishable from"), $\geq$ ("no more plausible than"), and $\rightsquigarrow$ ("the temporal predecessor of") on W; and an *evidence map* $E : W \to \mathcal{P}(\mathscr{T})$. Model M satisfies a number of conditions that can be found in Appendix A.

Definition 4.2.6. (Standard Model) A model M is *standard* if the strict plausibility relation $>$ is conversely well-founded and the immediate temporal predecessor relation $\rightsquigarrow$ is well-founded.

Definition 4.2.7. (Best World Assumption) A model M satisfies the *Best World Assumption* iff for every non-empty set $P \subseteq W$ such that $w \sim w'$ for all $w, w' \in P$, the set

$$\min_{\geq} P := \{w \in P | w' \geq w \text{ for all } w' \in P\}$$

is also non-empty. That is, there is always at least one "most plausible world".

Lemma 4.2.1. (Best Worlds Assumption) Every standard model satisfies the Best Worlds Assumption as defined in definition 4.2.7.

Proof. This follows from the converse well-foundedness of $>$ and the *Local Connectedness* condition in definition A.2.1. □

Truth for JB is defined in Appendix A.

4.2.3 Proof system

Theorem 4.2.2. (Proof system) Consider the following theorems that hold for JB:

i) For each $\varphi \in \mathscr{F}$, we have $\vdash \varphi$ iff there exists a logical term t such that $\vdash I(t) \wedge t \gg \varphi$ (Internalization)
ii) JB is sound and strongly complete with respect the the class of all models
iii) JB is sound and weakly complete with respect the the class of standard models
iv) JB is decidable

Please read [9, pp.7-11] for the complete theory of JB and proofs for i), ii), iii) and iv).

4.2.4 Evidence dynamics

Now let's add the actions to transform JB into a dynamic logic. The authors of [9] introduce four types of epistemic actions: $t+$, $t \otimes s$, $t!$ and $t \Uparrow$.

Definition 4.2.8. (Language DJB) $\mathscr{L}^{act} := (\mathscr{T}^{act}, \mathscr{F}^{act})$ is the extension of the static language for JB (see definition A.2.1) obtained by adding modal operators $[\alpha]$ for epistemic actions $\alpha \in \{t+, t \otimes s, t!, t \Uparrow\}$, for every $t, s \in \mathscr{T}$. The notions of subterm, subformula, admissibility and model are lifted to $\mathscr{L}^{act}$ in the obvious way.

The actions are to be interpreted as follows: $t+$ means that the evidence term t becomes available (not necessarily accepted), that is, added to the evidence set E. By performing $t \otimes s$, the agent forms a new term $t \cdot s$ representing the logical action of performing a Modus Ponens interference and hence adding $t \cdot s$ to E. $t!$ updates with some *hard* evidence t (coming from an absolutely infallible source), such that all worlds that do not fit the new evidence get eliminated. Finally, $t \Uparrow$ upgrades with some *soft* evidence t (coming from a strongly trusted, though not infallible, source), and as a consequence, the new evidence is accepted and all worlds that fit the new evidence become more plausible than the worlds that do not fit it. Note that these actions are only suitable for updating with terms; not for updating with formulas.

Please see Appendix A for the preconditions pre_α that capture the condition of possibility of action α, and the evidence set $\mathscr{T}(\alpha)$ that consists of all the evidence terms that become available due to α. Furthermore, in Appendix A the reader can find the truth definition for DJB.

4.2.5 Shortcomings

Comparing my list of desiderata and the characteristics of DJB, I need to adjust a couple of aspects to get a logic that fits the goal of my analysis. Firstly, I need to make the logic suitable for multi-agents, including techniques for private communication. Secondly, I want to include prior-evidence of agents, which is conceptually different from regular evidence, so I need to distinguish the priors from the normal evidence. Thirdly, I want to update not only with evidence terms, but also with formulas. In the next two sections I will describe how I can integrate these features into a Multi-agent Dynamic Evidence-based Logic.

4.3 Multi-agent Dynamic Evidence-based Logic

Recall the motivation behind constructing a Multi-agent Dynamic Evidence-based Logic: I want to compare specific network configurations and see how they affect the ability to repair false beliefs of the agents in the group. In specific, I want to test Zollman's hypothesis that transient diversity guarantees a high reliability, and that this is achieved by either limiting the communication between agents or by strenghtening the priors. I will first present the Multi-agent *Static* Evidence-based Logic, MSEL, that shows similarities with JB.

4.3.1 Syntax

Definition 4.3.1. (Language MSEL) Given a set Φ of atomic sentences, and a set of agents A, the language $\mathscr{L}^* := (\mathscr{T}^*, \mathscr{F}^*)$ consists of the set $\mathscr{T}^*$ of *observational evidence terms* t and the set $\mathscr{F}^*$ of *propositional formulas* (sentences) φ defined by the following double recursion:

$$\varphi ::= \bot | p | \neg\varphi | \varphi \wedge \varphi | E_i(t,m) | C_i(t,m) | N_i j | t \gg \varphi | \Box_i \varphi | K_i \varphi | Y\varphi$$

$$\text{with } p \in \Phi,\ i, j \in A \text{ and } m \in \mathbb{N}$$

$$t ::= o_\varphi \text{ with } \varphi \in \mathscr{L}^-$$

Notes on language

Consider the following informal readings of each language construct:

1. The formulas $\bot, p, \neg\varphi$ and $\varphi \wedge \varphi$ are classic formulas saying, respectively, 'falsum', 'proposition p holds', 'φ does not hold' and 'φ and φ hold'.
2. We can construct $\vee$ by using $\wedge$ and $\neg$ as usual in predicate logic, i.e., $\neg(\neg\varphi \wedge \neg\psi) \Leftrightarrow \varphi \vee \psi$
3. $E_i(t,m)$ says that 'evidence term t occurs m times in the evidence set of agent i'.
4. Likewise, $C_i(t,m)$ says that 'evidence term t occurs m times in the bias set of agent i'.
5. $N_i j$ says that 'j is a friend of i'.
6. $t \gg \varphi$ says that 't is admissible evidence for φ'. Note that $\gg$ is not labelled for agents. We already saw in section 4.1.2 that admissibility $\gg$ is universal for all worlds. Now that we have a multi-agent model, $\gg$ is also universal for all time and all agents.
7. $\Box_i\varphi$ and $K_i\varphi$ are lifted from the single agent formulas of DJB saying 'agent i defeasibly knows φ' and 'agent i infallibly knows φ'.
8. $Y\varphi$ says that 'yesterday (i.e., before the last epistemic action) φ was true'.
9. Finally, $\mathscr{L}^-$ is the Boolean propositional fragment of the language (the fragment built up by means of propositional letters, their negation, conjunction and disjunction) and o_φ is a piece of observational evidence for φ. Compared to JB, this is the replacement of the more general evidential certificate for φ: c_φ. Note that if $o_\varphi \in E_a$ then it is agent a who observed that φ, so we can see from the context who exactly observed that φ and do not need an index in the term construct itself. Recall that I agreed that observation can fail or be

misleading (see section 2.3 for the philosophical debate on the social dimensions of science), hence it might often occur that a piece of o_φ gets overriden by pieces of $o_{\neg\varphi}$.

Compared to JB, I omit the the term compounders + and ·, because I will not need them in my analysis and I want to keep things as simple as possible.

4.3.2 Semantics

Definition 4.3.2. (Model for MSEL) A *model* $M = (A, W, [\![\cdot]\!], \sim_i, \leq_i, \rightsquigarrow, N, E_i, C_i)$ is a structure consisting of a nonempty set of *agents* A, a nonempty set of *possible worlds* W, a *valuation map* $[\![\cdot]\!] : \Phi \to \mathcal{P}(W)$, binary relations $\sim_i$ (epistemically indistinguishability), $\leq_i$ (relative plausibility), and $\rightsquigarrow$ (immediate temporal precedence), a *friendship map* $N : A \to \mathcal{P}(A)$, an *evidence map* $E_i : W \to (\mathcal{P}(\mathscr{T}, m))$, and a *bias map* $C_i : W \to (\mathcal{P}(\mathscr{T}^*, m))$, satisfying the following conditions:

- $\sim$ is an equivalence relation (i.e., reflexive, symmetric and transitive) and $\geq$ is a preorder (i.e., reflexive and transitive).
- *Indefeasibility*: $w \leq v \Rightarrow w \sim v$
- *Local Connectedness*: $w \sim v \Rightarrow (w \leq v \vee v \leq w)$
- *Propositional Perfect Recall*: $(w \rightsquigarrow v \sim v') \Rightarrow \exists w'(w \sim w' \rightsquigarrow v')$ (i.e., knowledge of yesterday is still known today)
- *Evidential Perfect Recall*: $w \rightsquigarrow w' \Rightarrow (((t, m) \in E(w) \wedge (t, m') \in E(w')) \Rightarrow m' \geq m)$ (i.e., evidence of yesterday is still evidence today)
- *Uniqueness of Past*: $(w' \rightsquigarrow w \wedge w'' \rightsquigarrow w) \Rightarrow w' = w''$
- *Persistence of Facts*: $w \rightsquigarrow w' \Rightarrow (w \in [\![p]\!] \Leftrightarrow w' \in [\![p]\!])$ for $p \in \Phi$
- *Evidential Introspection*: $w \sim v \Rightarrow E(w) = E(v)$ (i.e., agents know what is in their evidence set)
- *Admissibility*: $o_\varphi \gg \varphi$
- *Bias Introspection*: $w \sim v \Rightarrow C_i(w) = C_i(v)$ (i.e., agents know what is in their bias set)
- *Persistence of Bias*: $w \rightsquigarrow w' \Rightarrow ((t, m) \in C_i(w) \Leftrightarrow (t, m) \in C_i(w')$
- *Consistency of Bias*: $t \gg \varphi \wedge t' \gg \neg\varphi \Rightarrow ((t, m) \in C_i(w) \Leftrightarrow (t', m') \notin C_i(w))$ (the consequences of bias evidence may not be contradicting)

Note the following general differences with JB:

1. The model contains a set of agents A and the binary relations $\sim_i$ and $\leq_i$ and the sets E_i and C_i are indexed with agent $i \in A$, to be able to refer to a specific agent.
2. We write $w \leq w'$ for "w' is at least as plausible as w" (whereas for JB we wrote $w \geq w'$).
3. $N : A \to \mathcal{P}(A)$ is the *friendship map* that tells who is friends with whom.
4. E_i maps a world $w \in W$ to a *multiset* $(\mathscr{T}, m)$, i.e., to an element of $\mathcal{P}(\mathscr{T}^*, m)$ such that for all m, $m(t) \in \mathbb{N}^+ = \mathbb{N} \setminus \{0\}$. [8]

[8] m is a partial map from $\mathscr{T}^*$ into the set of positive natural numbers. The image $m(t)$ is undefined if $t \notin \mathscr{T}^*$ and else m gives us the multiplicipty of $t \in \mathscr{T}^*$, i.e, how often t is included in the multiset $E_i(w)$. Note that for example the multiset $\{t, t, t, t', t'\}$ will also be denoted by me as $\{(t, 3), (t', 2)\}$ where I list each element of the set with its multiplicity. So I say $(t, 3)$ is an element of $\{(t, 3), (t', 2)\}$. The powerset of $(\mathcal{T}, m)$ is denoted as $\mathcal{P}(\mathcal{T}, m)$. The powerset is defined by taking all submuliplicities into account, so as an example the powerset of $\{(t, 1), t', 2)\}$ is given by $\{\emptyset, \{(t, 1)\}, \{(t', 1)\}, \{(t', 2)\}, \{(t, 1), (t', 1)\}, \{(t, 1), (t', 2)\}$.

5. $C_i(w)$ is a *bias map* that gives the biasses for every agent, also mapping worlds to the multiset, i.e., to an element of $(\mathcal{P}(\mathscr{T}^*, m))$. Even though E_i and C_i look similair, they are conceptually different. The bias set represents the agents' biases, and by presenting it as a multiset it specifically represents the agents' strength of bias. The bias set is fixed throughout time, while the evidence set gets updated with new evidence terms during the trial.

Definition 4.3.3. (Pointed Model) A *pointed model* is a pair (M, w) consisting of a model M and a designated world w in M called the "actual world".

Definition 4.3.4. (Standard Model) A model M is *standard* if strict plausibility relation $<$ and the immediate temporal predecessor relation $\rightsquigarrow$ are both well-founded.

Definition 4.3.5. (Truth for MSEL) The *satisfaction relation* $(M, w) \vDash \varphi$, writing $w \vDash \varphi$ when M is fixed, is defined as follows:

$$
\begin{array}{ll}
w \nvDash \bot & \\
w \vDash p & \text{iff } w \in [\![p]\!] \\
w \vDash \neg\varphi & \text{iff } w \nvDash \varphi \\
w \vDash \varphi \wedge \psi & \text{iff } w \vDash \varphi \text{ and } w \vDash \psi \\
w \vDash E_i(t, m) & \text{iff } (t, m) \in E_i(w) \\
w \vDash C_i(t, m) & \text{iff } (t, m) \in C_i(w) \\
w \vDash t \gg \varphi & \text{iff } t \gg \varphi \\
w \vDash \Box_i\varphi & \text{iff } v \vDash \varphi \text{ for every } v \geq_i w \\
w \vDash K_i\varphi & \text{iff } v \vDash \varphi \text{ for every } v \sim_i w \\
w \vDash Y\varphi & \text{iff } v \vDash \varphi \text{ for every } v \rightsquigarrow w
\end{array}
$$

Extend the valuation map $[\![\cdot]\!]$ to all sentences φ, for putting $[\![\varphi]\!] = \{w \in W | w \vDash \varphi\}$.

Lemma 4.3.1. (Belief) In a *standard model*, $M = (A, W, [\![\cdot]\!], \sim_i, \leq_i, \rightsquigarrow, N, E_i, C_i)$, "belief" is the same as "truth in the most plausible worlds":

$$(M, w) \models B_i\varphi \text{ iff } (M, w') \models \varphi \text{ for all } w' \in max\{w' \in W | w \sim_i w'\}$$

4.3.3 Network graph

In order to highlight the network structures, I will separately draw a network graph, as we know it from section 3.1, that shows us in a clear picture who is friends with whom, i.e., who communicates with whom. Recall that agents are represented by nodes and undirected edges represent communication.

Definition 4.3.6. (Set of friends) The group of friends of agent a, defined by $N(a)$, is the set of nodes that are connected to a via maximally one edge. Note that a is also a member of $N(a)$.

In definition 4.3.2 we saw that the epistemic model M contains the map $N : A \to \mathcal{P}(A)$, going from the set of all agents A to the powerset $\mathcal{P}(A)$, that determines whose evidence an agent takes into account when updating his belief. In definition 4.3.1 we saw that the sentence N_ij says that j is a friend of N.

One can use different network structures, hence different network mappings N and compare the effects they have on the epistemic model. It is required for any network to be *connected*, i.e., that there is a path between any pair of agents.

For the purpose of this thesis I let the network graph be stable throughout time, although it is possible that the network structure changes within one trial. Such an *evolution* of the network would appear when two non-friends become friends during the period of one trial. Also, it is only allowed for information to flow between friends. In principle it would be possible to let information flow between friends-of-friends of a, i.e., to the set $\{i|i \in N(j) \wedge j \in N(a)\}$, or more complex sets like "every agent that is friends with at least two of my friends": $\{i|i \in N(j) \wedge i \in N(k) \wedge j \neq k \neq a \wedge j, k \in N(a)\}$.

4.3.4 Evidence dynamics

So far, the new logic differs only in the details from JB. Adding the actions to make the model dynamic will make a greater difference. I will add two kinds of actions to the Multi-agent Static Evidence-based Logic: actions that do something with terms and actions that do something with formulas. Both actions are public: when agent a does α then everyone will learn that a did α. In case action α concerns an evidence term, the receiver(s) will update their evidence set; in case action α concerns a formula, the receiver(s) will update their epistemic state.

Definition 4.3.7. (Language MDEL) $\mathscr{L}^{*,act} := (\mathscr{T}^{*,act}, \mathscr{F}^{*,act})$ is the extension of the static language (see definition 4.3.2) for MDEL obtained by adding modal operators $[\alpha]$ for epistemic actions $\alpha \in \{t+_i, E_i(t,m)!_i, C_i(t,m)!_i, \bigwedge \mathsf{con}_t \Uparrow_i\}$, for every $t \in \mathscr{T}^*$, $\varphi \in \mathscr{F}^*$, $i \in A$, and $m \in \mathbb{N}$.

These action are to be interpreted as follows:[9]

- $t+_i$ means that (an extra instance of) evidence t becomes available to i (pubicly). For example, say $w \rightsquigarrow w'$ and $(t,m) \in E_a(w)$. After $t+_a$, $(t, m+1) \in E_a(w')$ becomes true.

- $E_i(t,m)!_i$ means that agent i publicly announces to his friends in $N(i)$ that she has m instances of evidence t in her evidence set.

- $C_i(t,m)!_i$ means that agent i pubicly announces to his friends in $N(i)$ that she has m instances of evidence t in her bias set.[10]

[9] I do not include *hard* updates of formulas supported by evidence (only hard updates on the content of agents' evidence sets $E_i(t,m)!_i$), coming from an infallible source, because I will assume that scientists only get *soft* evidence, i.e., observational evidence.

[10] This division between communicating E_i and C_i is quite unnatural. Even though evidence and bias are conceptually different objects, one expects the two to be communicated at the same time. However, it is not necessary to separately announce the two, because I want agents to upgrade their beliefs according to the knowledge they have about both the evidence and the bias (see precondition for $\bigwedge \mathsf{con}_t \Uparrow_i$ below). I take the bias to play the same role as the prior credences in probabilisitc frameworks as in [10]. This problem could be solved in more complex ways, but for reasons of ease I now assume that evidence and bias are separately communicated.

- $\bigwedge \mathsf{con}_t \Uparrow_i$ means that agent i upgrades her epistemic state with $\mathsf{con}_t := \{\varphi | t \gg \varphi\}$ (i.e., all formulas that are supported by t) and thereby accepting t and upgrading her plausibility order $\leq_i$ such that all worlds matching con_t become more plausible than all other worlds.

Note that adding evidence is a public action and announcing that an agent has some evidence is public too. Firstly, in this setting, it is unnessecary to have both actions represented in the model, because when the evidence is publicly added, then everyone allready knows what evidence the agent has. I do want to keep both actions, because it will be important when I would let both actions be private in further development of the logic. Secondly, not having private updates does not seem to match my objectives to model a situation where scientists privately conduct experiments and share their results with each other. This is not a problem, because I let the soft upgrades depend only on the knowledge of an agent about the evidence of her friends (see preconditions). Yet, it is unnatural that agents know about the evidence of everybody while only considering the evidence of friends, but for now it is the best solution to avoid the complexity that comes with making the actions private. Since the actions are public, the contents of everyone's bias and evidence sets are common knowledge. Given that the network structure is also common knowledge, agents are also able to derive how the other agents in the network upgrade their plausibility order. This results in a compact model where everybody knows each others evidence sets and beliefs. If private actions would be included, then the model will quickly grow for two reasons: uncertainty about evidence sets and uncertainty about doxastic states of other agents. In section 4.3.5 I will informally explain how to extend the language and the semantics such that the agents can perform private actions. For now, the evidence set of any agent is the same for every possible world within one timestep.

Lemma 4.3.2. (Admissible terms of $t \in \mathscr{T}^*$) $\mathscr{T}^* = \mathscr{T}^{*,e}$

Proof. For all $t \in \mathscr{T}^*$, it holds that t is of the form o_φ. Also, it always holds that $o_\varphi \gg \varphi$. Therefore, for all $t \in \mathscr{T}^*$ there exists a $\varphi \in \mathscr{F}^*$ such that $t \gg \varphi$. □

With t always being of the form o_φ and the only action that causes an evidential change being $+_i$, we can define the set of term $\mathscr{T}^*(\alpha)$ that becomes available due to α:

$$\mathscr{T}^*(\alpha) = t$$

Of course, if one would allow other types of evidence and actions, $\mathscr{T}^*(\alpha)$ would have to be generalised.

Preconditions actions

For every action α define the precondition pre_α:

$$\mathsf{pre}_{t+_i} := (t \gg \varphi \wedge t' \gg \neg\varphi) \wedge \bigwedge_{\varphi \in \mathsf{con}_{t'}} \neg B_i \varphi$$

$$\mathsf{pre}_{E_i(t,m)!_i} := E_i(t,m)$$

$$\mathsf{pre}_{C_i(t,m)!_i} := C_i(t,m)$$

A precondition is a formula that needs to be true for the action to be performed. The action $t+_i$ can happen when $(t \gg \varphi \wedge t' \gg \neg\varphi) \wedge \neg B_i \mathsf{con}_{t'}$ is true. Since I have only observational evidence and one theory p, in this setting the precondition for o_p+_i would be that $\neg B_i \neg p$. This is derived from the Restricted Outcome rule that I will introduce in definition 5.1.1, that states that an agent can only observe t when she already believes in the consequences of t or when she is indifferent about the consequences of t or t'. To announce $E_i(t,m)$, agent i needs to have indeed m instances of t in her evidence set. Let me emphasise that agents do not share the piece of evidence t, but the proposition that she has evidence t (i.e., she has done an experiment and observed t). The same applies to $C_i(t,m)!_I$.

Recall that I require that $m(t) \neq 0$, which implies that in order to communicate $E_i(t,m)$ or $C_i(t,m)$, there has to be at least one instance of t in the evidence (or bias) set of i.

Now let's define the preconditions for the soft upgrade $\Uparrow$. Let $\theta \in [0,1]$ be the *threshold* that determines what is required for agents to change their beliefs. Next, define the following abbreviations:

$$t_i^{m+n>0} := \bigvee_{m \in \mathbb{N}^+} E_i(t,m) \vee \bigvee_{n \in \mathbb{N}^+} C_i(t,n)$$

saying "there exists t-terms either as evidence or as bias", and

$$t_i^{m+n=0} := \bigwedge_{m \in \mathbb{N}^+} \neg E_i(t,m) \wedge \bigwedge_{n \in \mathbb{N}^+} \neg C_i(t,n)$$

for "there are no t-terms as evidence or bias".

Furthermore, define $t_i > t'_i$ as the formula

$$(t \gg \varphi \wedge t' \gg \neg\varphi) \wedge ((t_i^{m+n>0} \wedge t_i'^{m'+n'=0}) \vee \bigvee_{m+n>m'+n'} (t_i^{m+n>0} \wedge t_i'^{m'+n'>0}))$$

In words this expression captures that 'term t is related to φ and t' to $\neg\varphi$ and agent i has more evidence t than t''. This means that in any case the agent needs to have t-terms either as evidence or as bias. When the agent has terms t and no terms t', then she definetely has more t than t'. In case the agent has both t and t' in her evidence set, then the last part of the disjunction in the above formula expresses that there are more items of evidence t than items of evidence t'.
Now I can define the preconditions:

$$\mathsf{pre}_{\bigwedge \mathsf{con}_t} \Uparrow_i := \bigvee_{G \subseteq N(i)} \left(\frac{|G|}{|N|} > \theta \wedge \bigwedge_{j \in G} K_i(t_j > t'_j)\right)$$

This expression captures that agent i performs an upgrade of the consequences of t iff she infallibly knows that strictly more than θ of her friends (possibly including herself, possibly not[11]) have more evidence t than t'.[12]

[11]Note that in the settings of this thesis, the evidence of an agent i herself is considered exactly as important as the evidence of her friends. So $> \theta$ of her friends j having $t_i > t'_j$ while the agent herself has $t_i < t'_i$ still moves the agent to upgrade with $\bigwedge \mathsf{con}_t$.

[12]This causes one debatable situation: suppose $\theta = \frac{1}{2}$ and exactly $\frac{1}{2}$ of a's friends have more

Truth

For the truth clauses of the dynamic operators $[\alpha]$ in MDEL, I need to define an operation of what it means to add an element of multiplicity 1 to a multiset, i.e., how to define the union of two multisets where one is consisting of a singleton with one element and has multiplicity 1 (e.g. the sum of $\{(t,3),(t',2)\}$ and $\{(t,1)\}$ should give $\{(t,4),(t',2)\}$). To do this, I will define the extended multiset that has value 0 for each element not contained in the original set. Let T be the set (a regular set, *not* a multiset) of all terms that are defined by m, i.e., of all terms that are in the multiset under consideration. Let m^* be the extension of m, such that $m^*(t) = 0$ for every $t \notin T$ and $m^*(t) = m(t)$ for every $t \in T$. With the extensions given, I can define the multiset sum of (T, m^*) and (T', g^*) (a map that acts on the elements of $T \cup T'$) as follows:[13]

$$(m^* \oplus g^*)(t) = m^*(t) + g^*(t)$$

Let the resulting multiset be reduced to a regular multiset $(T \cup T', m \oplus g)$ where I delete the elements in which the multiplicity is 0.

Definition 4.3.8. (Truth for MDEL) Let w^α denote the ordered pair (w, α) to represent the "updated" world resulting from performing action α in world w. Then:[14]

$(M, w) \vDash [\alpha]\varphi$ iff $(M[\alpha], w^\alpha) \vDash \varphi$ with $M[\alpha] := (A^\alpha, W^\alpha, [\![\cdot]\!]^\alpha, \sim_i^\alpha, \leq_i^\alpha, \rightsquigarrow^\alpha, N^\alpha, E_i^\alpha, C_i^\alpha)$, and

$$\begin{aligned}
A^\alpha &:= A \\
W^\alpha &:= W \cup \{w^\alpha | w \in [\![\mathsf{pre}_\alpha]\!]\} \\
[\![p]\!]^\alpha &:= [\![p]\!] \cup \{w^\alpha \in W^\alpha | w \in [\![p]\!]\} \\
\sim_i^\alpha &:= \sim_i \cup \{(w^\alpha, v^\alpha) | w \sim_i v\} \\
\leq_i^\alpha &:= \leq_i \cup \{(w^\alpha, v^\alpha) | w \leq_i v\} \text{ for } \alpha \in \{t+_i, E_i(t,m)!_i, C_i(t,m)!_i\} \\
\leq_i^\alpha &:= \leq_i \cup \{(w^\alpha, v^\alpha) | (w \notin [\![\mathsf{con}_t]\!] \wedge v \in [\![\mathsf{con}_t]\!]) \vee (w \notin [\![\mathsf{con}_t]\!] \wedge w \leq_i v) \vee \\
&\qquad (v \in [\![\mathsf{con}_t]\!] \wedge w \leq_i v)\} \text{ for } \alpha = \bigwedge \mathsf{con}_t \Uparrow_i \\
\rightsquigarrow^\alpha &:= \rightsquigarrow \cup \{(w, w^\alpha) | w \in [\![\mathsf{pre}_\alpha]\!]\} \\
N^\alpha &:= N \\
E_i^\alpha(w) &:= E_i(w) \text{ for } w \in W \\
E_i^\alpha(w^\alpha) &:= (T \cup \{t\}, m \oplus g) \text{ for } \alpha = t +_i \text{ and } E_i(w) = (T, m) \\
E_i^\alpha(w^\alpha) &:= E_i(w) \text{ for } \alpha \neq t+_i \\
C_i^\alpha(w) &:= C_i(w) \text{ for all } w \in W^\alpha
\end{aligned}$$

t than t' and exactly $\frac{1}{2}$ of a's friends have more t' than t, and a knows all of this. The current precondition says that in this case, the agent does not upgrade, so she stays with her old belief. One could also propose a contraction action $\bigwedge \mathsf{con}_t \Downarrow_i$, making the con_t-worlds and $\mathsf{con}_{t'}$-worlds equiplausible. The authors in [25] propose a contraction operator that can be used. Another unnatural situation arises in the case where $N(c) = \{a, b, c\}$, $t_a > t'_a$ and both b and c have an equal amount of evidence t and for t'. It would be natural for c to believe in the consequences of t, but according to this definition she does not upgrade his beliefs and stays with his old plausibility order. This is something for further research to explore.

[13] This definition is found at `http://planetmath.org/operationsonmultisets`.

[14] Note that since it always holds that $t \in \mathscr{T}^e$, I do not need to give a seperate definition for the update of $t \notin \mathscr{T}^e$.

Let me informally explain how I obtain the updated model $M[\alpha]$.

- The agents are fixed throughout time, so $A^\alpha := A$.
- I keep the old worlds $w \in W$ and add new worlds that are the result of the action α, which are updated versions of those worlds that satisfy the precondition of α.
- The atomic facts that hold at w are carried on to the new worlds w^α.
- Since there are no hard updates, the indistinguishability relation $\sim_i^\alpha$ simply copies all pairs from the old model.
- When α is not the soft upgrade $\bigwedge \mathsf{con}_t \Uparrow_i$, then the plausibility order is also copied to the new worlds. When $\alpha = \bigwedge \mathsf{con}_t \Uparrow_i$, then I have to correctly update the new worlds to match α: I let $w^\alpha \leq_i v^\alpha$ for all pairs such that either i) con_t do not hold at w but do at v; or ii) con_t do not hold at w and in the original worlds $w \leq_i v$, or iii) con_t hold at v and in the original worlds $w \leq_i v$.
- The temporal relation $\rightsquigarrow^\alpha$ has to include the pairs between the original worlds w and their updated versions w^α.
- The network connections are stable, so $N^\alpha := N$.
- Evidence in the old worlds stays the same. Evidence in the new worlds w^α have to be updated when $\alpha = t+_i$, such that the new evidence set (obtained by adding the new evidence $(\{t\}, g)$) contains this extra term t. This is done with the help of the sum operation $\oplus$ on extended multisets m^* as explained above. If $\alpha \neq t+_i$, then the evidence sets stay intact.
- Finally, the bias set stays equal throughout any update.

4.3.5 Extension

There are many possibilities for extending this system into a more generally applicable system. To give an idea, I will name a few. The most urgent extension concerns the public actions $+_i$, $E_i(t,m)!_i$ and $C_i(t,m)!_i$. To simulate more naturally what happens in everyday science (scientists privately conduct experiments and later either publicly or privately communicate that they found a particular result), I must make these actions private. To let agent a privately announce $E_a(t,m)$, I have to distinguish the set of 'insiders', $N(a)$, and the set of 'outsiders' $A \setminus N(a)$. We can learn from DEL how to use action models that implement uncertainty of agents about the actions of others (for example, see [12]). There are different 'degrees' of privacy: the action can be completely private: then the outsiders might think nothing has happened, and the action can be semi-private: then the outsiders know that something has happened but it is unclear what (i.e., they know that agent a conducted an experiment, but they do not know the results. Or they know that a communicated something with her friends, but they do not know what exactly). The semantics need to be extended with uncertainty about actions that happened for different agents and I need full event models with uncertainty over actions. With these adjustments, I can let $+_i t$ be necessarily private and add $E_i(t,m)!_i^{N(i)}$ and $C_i(t,m)!_i^{N(i)}$ for private announcements of $E_i(t,m)$ and $C_i(t,m)$ from i to her friends in the set $N(i)$.

As an example, suppose that agent a privately updated with $+_a t$. Then in the updated model, I take a copy of all worlds and relations of the original timestep three times. One part keeps the same relations and evidence sets, one part adds t to

E_a and one part adds t' to E_a (or, if I would consider more outcomes of experiment, worlds with those terms added to E_a should also be added to the updated model). These three parts of the model will have arrows between them for all the outsiders who are uncertain between that part of the old copied model and the new parts of the model in which I added the new evidence t or t'. For example, $c \notin N(a)$ does not know whether $+_a t$, $+_a t'$ or that a observed nothing.

If I would let the biasses be private and keep the concept of prior beliefs, then the extension would have to adjust the system in one of the following ways: i) let the agents communicate their biasses in order for the agents to upgrade their beliefs; ii) make a special upgrading rule that applies to the initial state where agents did not yet hear about their friends' biasses; iii) else, let the prior belief be neutral and such that after one round of conducting experiments and accordingly communicating the results, the agents will upgrade their beliefs for the first time. I leave this choice open for further research. In the current settings with public updates the bias does not need to be communicated to get prior beliefs, because everybody knows about each others' biasses and network structure beforehand. Further research should specify the exact adjustments to the semantics of the extended system.

Furthermore, one can add compound terms $t \cdot s$ and $t + s$, as in JB. We can also add (negative) weights to the Y's in a formula, such that for instance old evidence is less important than more recent evidence. To handle these additions, we should design more complex tools to store and manage the evidence. Additionally, one could include other updates, such as a contraction operator that I mentioned earlier, or hard updates of infallible evidence.

Chapter 5

Logical Analysis

In the previous chapter I have created a logical system that will help us to analyse the effects of several network configurations on the reliability of scientific communities. Before I will actually model and compare different setups, I will first discuss some assumptions I have to make for the analysis to be manageable. In section 5.3 I will model some simple situations, to see how updating and adopting behavior works. Finally, I will model the more complex networks and conclude which factor has what effect on the reliability of the community.

5.1 Assumptions and simplifications

Since it is impossible to model every single aspect of reallife science, I will make simplifications and assumptions. Despite the philosophical discussion on the impossibility of knowing when we have reached the absolute truth, in the logical analysis I will speak of the 'true world'. I can do this, because I will be analysing from a meta-perspective, from which I can denote one of the worlds as the 'true world', also referred to as the 'actual world' or 'real world'. I assume that the agents involved all want to find out which is the true theory, what in Zollman's analyses stands for "getting the highest payoff".[15] Apart from this philosophical point, I make some other practical assumptions and simplicifactions that I will now discuss.

5.1.1 Zollman's Bandit-studies

One crucial point for Zollman's Bandit problems is that learning, or revising beliefs, demands the input concerning the observation of others. In the case mentioned in section 3.2, it is important that the agents in the network who believe in medicine A exchange their experimental outcomes in order for the agents believing in medicine B to adopt the belief in A. Recall that in Zollman's example ([42]) agents only focus on the medicine they believe is best, since it is generally taken to be irrational to test the inferior medicine. In Bandit problems, agents who believe in p observe a certain payoff $P(p)$. This payoff does not say whether the payoff for the other theory, $P(q)$, is higher or lower then $P(p)$. Only information about both payoffs gives the data which is needed to compare p or q and conclude which is the best

[15] If in reality there is not a true theory, the agents want to approach the true theory; they want to find the theory that most closely describes the truth. Then the goal remains to find out which theory gives the highest payoff.

medicine. When everyone believes in the same medicine, the group is stable and no one will ever learn about the other medicine, henceforth no one will ever test the other medicine again.

In bandit situations, the agent has to decide between two opposing theories p and q. Believing one will necessarily entail disbelieving the other. In classical logic, this implies that $B_i q$ is equivalent to $B\neg p$. Therefore, I will use $B_i p$ and $B_i \neg p$ to denote the belief in theory p or belief in the opposite theory q. Furthermore, using theory p in logic can only result in confirming that p or disconfirming that p; which implies that $\neg p$. The loss here is that we do not get payoff values which have to be compared in order to decide the value of the theory, but merely a "yes p (or $\neg p$)" or "no p (or $\neg p$)" that directly gives information about the opposing theories: "yes p" implies "no $\neg p$" and "no p" implies "yes $\neg p$". In this case, agents do not need the input of the experimental results of their friends to learn about the other theory. A rational strategy could be to simply individually conduct the experiment a million times and conclude whether p or $\neg p$. This independence is however not the type of situations I am studying in this thesis. My aim is to capture at least some aspects of the problems of communication in science. If we do not let learning be dependent on communication, then a community will end up in a state where no one will ever learn about the other medicine because everyone believes in the same medicine. This is an interesting state, because communities can end in this state after a few rounds of sharing and updating. I would like to see which factors increase the chances of getting to such a state where everyone believes in the true theory, compared to everyone believing the false theory. I call such a state *finished learning*.

Note that keeping two seperate propositions p and q will not make the outcomes be independent from each other (and thus forcing the agents to communicate), since to express the independence I have to add the bi-implication $B_i p$ iff $B_i \neg q$. I will take this independence into account, assuming the agents follow the rule:

Definition 5.1.1. (Restricted Outcome Rule) If an agent believes p, then she can conduct an experiment for p, which means observing either p ("confirming that p") or $\top$ ("observing nothing"); if an agent believes $\neg p$, then she can conduct an experiment for $\neg p$ and observe either $\neg p$ or $\top$. If an agent is indifferent about p or $\neg p$, then she conducts both experiments and observes either p or $\neg p$, i.e., she does not choose between testing p or $\neg p$.

Note that by testing p and not obtaining any confirming evidence (observing $\top$), the agent does not obtain confirming evidence for the opposing theory. To obtain confirming evidence for $\neg p$ the agent has to conduct an experiment for $\neg p$. The idea behind this is that the agents use the theory they believe in (because that's the most rational thing to do) and that the agents need the input about the experimental results of others in order for them to change their mind. With this rule, the group can reach a state of finished learning (because the group will never in the future reach a state where they can in the future learn about the other theory, since no one will be able to observe something about it). Restricting the agents to observe only either a confirmation of their belief or nothing, comes most close to the feature of Zollman's Bandit problems. The last sentence in definition 5.1.1 which deals with

the state of indifference does not match with Zollman's bandit cases, where agents always test either medicine 1 or medicine 2 and never both, but I think it is the most natural way to deal with agents who are forming their beliefs on the basis of an equal amount of evidence for p and for $\neg p$.

Compared to Zollman's analysis in [41, 42], this reconstruction cannot capture all elements at play. Even with the Restricted Outcome Rule, logic cannot fully simulate Zollman's Bandit studies one-on-one, because in principle the outcomes of Zollman's method are subjective probabilistic values and the outcomes of the qualitative logical setting are 0 and 1.[16] Another difference is found in the fact that in Zollman's use of a payoff function, the agents thereby also directly communicate their belief: an agent can communicate a payoff value $P(p) = x$ if and only if she believed in p and therefore tested p. By using the Restricted Outcome Rule in my setting, the agents only communicate their belief if the outcome is p or $\neg p$; if they observe $\top$ this can be the result of believing both p or $\neg p$.

Keeping the differences with Zollman's setting in mind, the added value we get from logic, i.e., insight on the agent's higher-order reasoning, is still good enough to proceed with my analysis. We can still analyse the effect of specific network configurations on reliability of communities. We cannot put the results directly next to Zollman's, due to the limitations mentioned above, but we can say sensible things about effects of different network configurations on the epistemic behaviour of scientific groups.

5.1.2 Distribution of priors and failures

By using logic instead of computer simulations (as done in [41, 42]) I have to make another "sacrifice". Zollman lets his agents believe in a medicine; then they conduct the experiment that tests that medicine; then the agents share their results with their neighbours; then they maybe change their beliefs, etc. He lets the computer run this protocol for 10.000 iterations. Then, he lets the computer run this trial many times. Describing the outcomes of these simulations, Zollman does not explicitly say how he distributes the failures of the experiments, i.e., how often does an experiment give the 'wrong result'? Suppose medicine 2 is the best. Then how often do we get a payoff for medicine 2 that is lower than the avarage payoff for medicine 1? It looks like Zollman lets the computer randomly distribute the failures. Zollman does mention how he distributes the prior believes: he "assigns the agents random beliefs uniformly drawn from the interior of the probability space" [41, p.579]. By running so many different experiments and trials, it does not really matter how these are distributed. It is the average outcome that is presented in his papers, so extreme individual trials will be overriden by the mean value. In my models, I will model each trial manually, looking closely at what happens at each step: giving the agents prior evidence - implementing prior beliefs - choosing a method and apply - getting results - updating evidence set -communicating evidence - updating epistemic state - choosing a method - etc. Since I will not repeat this 10.000 trials, it does matter how the priors and failures are distributed. In fact, Bala and Goyal ([3]) clearly

[16] One could turn to a probabilistic logic or multi-valued logic, however I leave that analysis for future work. Probabilistic logics have been used to model social scenarios in [5].

state that the distribution of priors has an effect on the success of the group. The main question is then: how should I decide how to distribute the priors and failures of the experiments?

The important thing to note here, is that the most important condition is that I have to keep the distributions of the priors and failures equal. When I compare two network structures, I have to keep these distributions the same to conclude something about the effect of the network structure. Since I have to deal with the Restricted Outcome Rule, it may sometimes occur that an agent at timestep x observed o_φ in one trial, but believes $\neg\varphi$ at timestep x in the other trial, such that she cannot observe o_φ. In this case, let the agent observe $\top$. I can change these distributions and see how this affects the comparison between different network structures. Let me emphasize that my results will not be statistically motivated, but I focus on getting more insight into the details; into the process of agents' adopting behavior.

5.1.3 Other assumptions

I highlight a couple of other simplifications and assumptions for my analyses. As for the network, I let the number of agents be fixed. In further research, one could also compare different sizes of groups.[17] I also assume that everyone in the group knows the network structure, i.e., who communicates with whom. Furthermore, let evidence be automatically accepted. Of course, this can easily be loosened, since the authors in [9] already gave the techniques to do this.[18] Further, I keep the threshold θ for updating fixed at $\frac{1}{2}$, such that an agent will adopt a belief p when more than half of her friends have more evidence t for this belief than evidence t' for the opposite belief $\neg p$. I also keep the strenght of biasses per agent the same, as well as the reliability of the testimony of all friends. These three factors can be varied in further research. For now, however, I have kept the system as simple as possible. In section 5.4 I will use this to compare the few variations that I do allow for, dealing with network structure and the universal weight of biasses.

5.2 Definitions

The next two sections will be dedicated to computing different circumstances. To compare the outcomes in an organised way, please consider the following definitions:

Definition 5.2.1. (Round) When I refer to a *round* in the analysis, I refer to one sequence of 'conducting an experiment', according to the Restricted Outcome Rule (action $t{+}_i$), then 'communicating the results' (action $Et!_i$) and finally 'upgrading belief' (action $\bigwedge \mathsf{con}_t \Uparrow_i$).

Definition 5.2.2. (Trial) When I model different circumstances, in the initial state the biasses are distributed. The next thing I do is upgrade the prior beliefs such that they correctly match the biasses. Next, I repeatedly compute rounds. The total of

[17] The ongoing investigations of the team directed by S. Perovic in Belgrade does focus exactly on this team number aspect.

[18] This could actually be quite interesting, since we saw for example in section 2.2 that sometimes experimental results are only accepted when there is a theory that can explain it, as in the discovery of the weak neutral current.

all of the prior-work and these rounds is called one *trial*. By adjusting the variables I will compare different trials.

Definition 5.2.3. (Stable) A community is *stable* if performing one round of updating and communicating will have no effect on the doxastic states of any agent in the community.

A community can be stable for one round, but may change again after another round. When the community will be stable no matter how many rounds I compute, the community has finished learning.[19]

Definition 5.2.4. (Finished learning) A community has *finished learning* if the community is stable for any future round.

Definition 5.2.5. (Successful learning) A community has *successfully finished learning* if all agents $a \in A$ believe the true theory.

Definition 5.2.6. (Semi-successful learning) A community has *semi-successfully finished learning* if at least $\frac{1}{2}$ of the set A believes the true theory *and* the community has finished learning.

Definition 5.2.7. (Failed learning) A community has *failed learning* if less than $\frac{1}{2}$ of the set A believes the true theory and the community is finshed learning.

5.3 Basic trials

Now let me slowly increase the complexity of the models, starting with some simple networks that are intented to illustrate how MDEL works and how to judge networks on their behaviour. In the following sections, let $t = o_p$ and $t' = o_{\neg p}$, hence $t \gg p$ and $t' \gg \neg p$. In the drawings of my Kripke models I leave all reflexive and transitive arrows for $\sim_i$ and $\leq_i$ implicit. I will sometimes write $E_i = \{t, t\}$ and sometimes $E_i = \{(t, 2)\}$.[20] Note that since I assume agents have the same biasses in every possible world, I will often use C_i instead of $C_i(w)$. Finally, let p be the true theory.

Basic trial I: Two agents without bias

Let's start with a network of two friends, a and b.

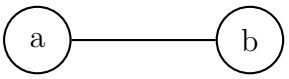

Figure 5.1: *Network graph for two friends a and b*

[19]The authors in [25] also define a community that will never have finished learning, because the agents will endlessly keep on switching back and forth from one belief to another, as being *in a flux*. In my analysis communities will never be in a flux in that sense. The most important cause for this difference is that in [25] more credits are given to the belief of a friend than to those of the agent herself, whereas in boths settings the evidences are valuad equally.

[20]Note that I leave out the case when an agent has no evidence for a term, as the multiplicity map is partial and leaves out the case for $m = 0$, although the notion of multiset can be extended to include this if needed.

Note that this is a *complete graph*, since every pair of nodes is connected by one edge. This gives us the map of N such that $N(a) = \{a, b\} = N(b)$. Suppose that a and b's bias sets C_a and C_b are empty. The initial epistemic model looks as follows:

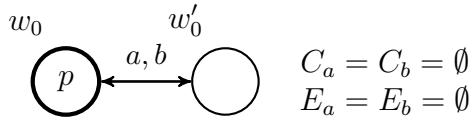

Figure 5.2: *Initial epistemic model for Basic trial I*

In Figure 5.2 we see that it is common knowledge that nobody conducted any experiment untill now, because the multisets E_i are the same at every possible world. Hence agent a knows that b has no evidence and b knows that a knows this, etc. Since all updates $+_i t$, $E_i(t, m)!_i$ and $C_i(t, m)!_i$ are public, this will always be the same at every world in one timestep.

I will now stepwise update the model. Suppose that the first actions are $+_a t$ and $+_b t'$. That is, agent a observes that p and agent p observes that $\neg p$. These outcomes are legimite according to the Restricted Outcome Rule, since both agents were still indifferent about whether p or $\neg p$. Further, note that the order in which one updates the two actions does not matter, so I will update the model with both actions in one timestep. I will only update multiple actions at the same time when this does not change the results. This is always the case when agents simultanously do the same kind of action: updating their evidence set, communicating their evidence and bias, or upgrading their beliefs. In the updated model $M[+_a t, +_b t']$ (Figure 5.3) it holds that $E_a = \{t\}$ and $E_b = \{t'\}$. I will only write the bias states for the first time moment, because it will not change within one trial.

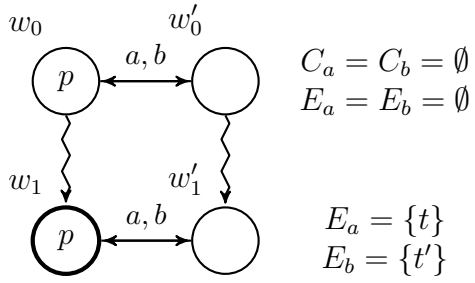

Figure 5.3: *Basic trial I after* $+_a t, +_b t'$

Next, the agents will communicate their observations: $E_a(t, 1)!_a$ and $E_b(t', 1)!_b$. In the updated model the evidence sets nor do doxastic states of the agents have changed (see Figure 5.4).

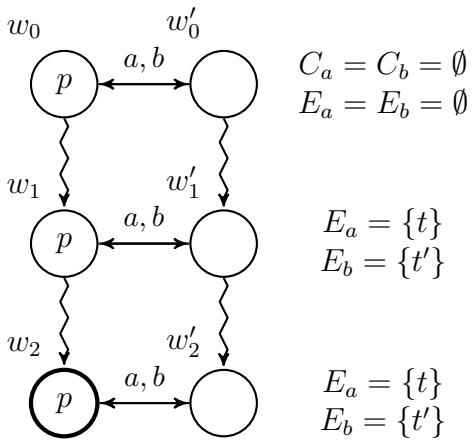

Figure 5.4: *Basic trial I after* $E_a(t, 1)!_a, E_b(t', 1)!_b$

Let's see if the agents will upgrade their plausibility order. Since the fraction of friends supporting t is exactly $\frac{1}{2}$ for both agent a and b, they will both stay with their initial belief state: being indifferent about whether p or $\neg p$. Henceforth, no upgrade action happens. Then, after another round of experiments with full visibility to all agents, we have public updates $+_a t$ and $+_b t$ that add terms to the evidence sets of a and b (and inform everybody about this event). See Figure 5.5 for the corresponding epistemic model.

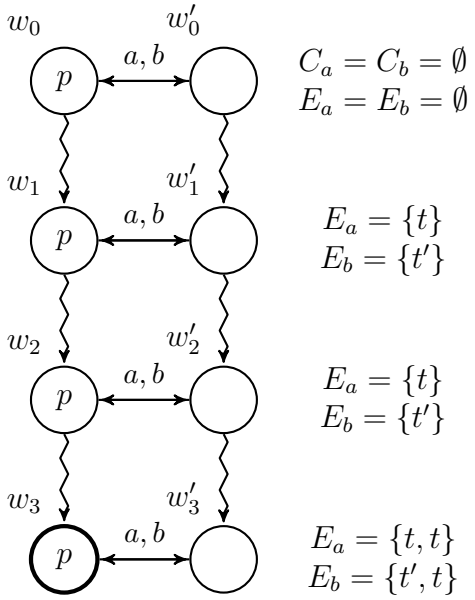

Figure 5.5: *Basic trial I after* $+_a t, +_b t$

After updating their evidence sets, the agents will publicly communicate $E_a(t, 2)!_a$, $E_b(t, 1)!_b$ and $E_b(t', 1)!_b$. Since nothing changes in the model, I do not draw this timestep now. Note that the fraction of agents having more support for p than $\neg p$ is equal to $\frac{1}{2}$, hence no soft upgrade will be made. In the next round, both a and b observe p. This causes the public updates $+_a t$ and $+_b t$, as depicted in Figure 5.6.

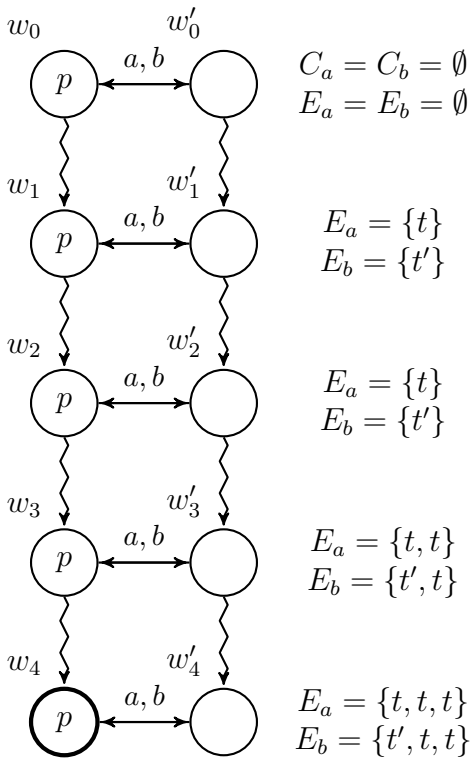

Figure 5.6: *Basic trial I after another round of* $+_a t, +_b t$

Consequently, agent a and agent b will communicate $E_a(t, 3)!_a$, $E_b(t, 2)!_b$ and $E_b(t, 1)!_b$. Now, both agents know that every agent in their set of friends has more instances of t than t' in her evidence set and bias set. Because of that they will upgrade their plausibility set such that for all worlds $w' \in [\![p]\!]$ it holds that $w <_A w'$ for $A = \{a, b\}$ (see Figure 5.7). Since the agents have common knowledge about the network structure and common knowledge about each others' evidence sets, they can both deduce the plausibility order of their friend.

Now, the community has *finished learning*: both agents believe that p, so by the Restricted Outcome Rule they will only get new evidence p or $\top$, and no one will ever observe $\neg p$ in the future. Furthermore, since p is the true theory, the community has achieved a status of *successful finished learning*.

Let's quickly analyse *Basic trial I*. Firstly, since all agents in the network share the same information about each others evidence, the content of the evidence sets are *common knowledge* to all agents in the network. Secondly, imagine that the agents observed more often that $\neg p$ (which is a false result, hence a 'failed experiment'). One can guess that this will change the adopting process. This means that the distribution of failures probably has a great effect on the outcome of the trial.

In the next examples I will only highlight some aspects of the trial. When the network grows and the network needs more time to stabilise, the epistemic model can get very big. In the actual analysis in section 5.4 I will present the outcomes (whether or not the community has successfully finished learning) in a table and

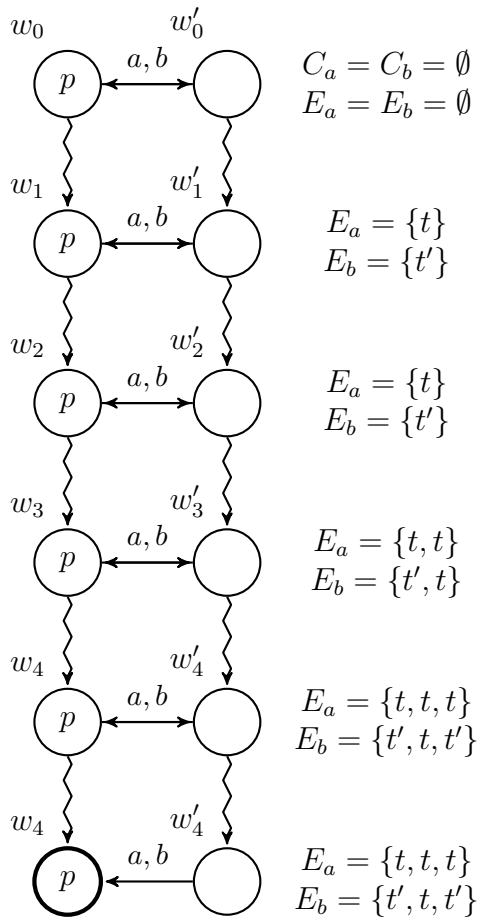

Figure 5.7: *Basic trial I after* $\bigwedge \mathsf{con}_t \Uparrow_a, \bigwedge \mathsf{con}_t \Uparrow_b$

afterwards discuss the most striking behaviors.[21]

Basic trial II: Two agents with bias

Let's see what happens when I add strong biasses to the model and then compare the process and the outcome to the previous example. Since I will be comparing the effect of the bias, I have to be careful to keep the other variables, i.e., the network structure and the distribution of failures, exactly the same.

For this trial, I give a a strong bias for p, so $C_a = \{t, t, t\}$ and b a strong bias for the opposite theory, so $C_b = \{t', t', t'\}$. We see that the strenght of the biasses has a value of 3 and that this value is universal for all agents in the network. Recall that the bias sets are public, so the agents know from each other what bias they have. Before I let the agents conduct experiments, I first have to set he agents' prior beliefs. Since it is not the case that more than half of a's friends have more evidence for t or for t', a will not upgrade her beliefs and stay indifferent about whether p or $\neg p$. The same holds for agent b. See Figure 5.8 for the initial model.

[21] If the reader is interested, he or she can ask the author for the calculations.

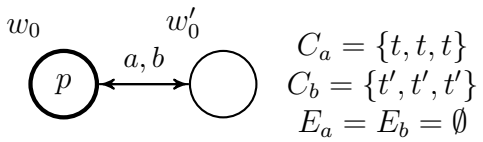

Figure 5.8: *Initial epistemic model for Basic trial I*

As in *Basic trial I*, the first round of updates will be $+_a t$ and $+_b t'$, the second round of updates will be $+_a t$ and $+_b t$ and the third round of updates will be $+_a t$ and $+_b t$. After three rounds, b's evidence set $E_b = \{t', t, t\}$ is not strong enough to overrule her bias set $C_b = \{t', t', t'\}$. As a consequence of b's strong bias set, we have to 'wait' for at least five updates of $+_b t$ in order for $t_b > t_b'$ ("having more evidence t than t', see precondition for $\mathsf{con}_t \Uparrow_i$). Assuming a realistic distribution of failures, after a certain number of rounds both a and b reach a state of $t > t'$. After that round, more than $\frac{1}{2}$ of $N(a)$ and more than $\frac{1}{2}$ of $N(b)$ will have $t_i > t_i'$. As both agents will know this, they will upgrade their beliefs such that $B_a p$ and $B_b p$. As a result, the community will have *successfully finished learning*.

Comparing these results to *Basic trial I*, we see that it took this community longer to reach the state of finished learning: in *Basic trial I* it took 3 rounds, while in *Basic trial II* this took 6 rounds.

We can learn from the first two examples that it is more informative to let agents have an uneven number of friends, because in those cases the fraction of friends having more evidence of one kind than of the other will be more often $> \frac{1}{2}$. Using uneven numbers of friends will force the agents to upgrade more often.

Basic trial III: Four agents in a circle

Now I will increase the number of agents to make the model slightly more complex. Consider a network of four agents a, b, c and d, that are structured as a *circle*:

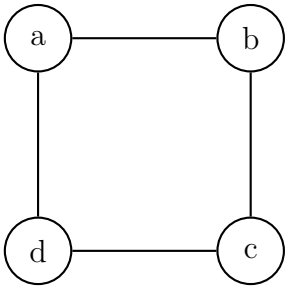

Figure 5.9: *Network graph for four friends a, b, c and d in a circle*

Let the biasses be such that $C_a = C_b = \{t\}$ and $C_c = C_d = \{t'\}$. In Figure 5.10 we see the initial state and the upgrades according to the biasses. Recall that since the bias sets are public and thus common knowledge, as well as the network structure, all agents know what the other agents in the network believe. If I would let the

bias sets be private, then the epistemic model would be a lot bigger, modelling the agents' mutual uncertainty about each other's beliefs.

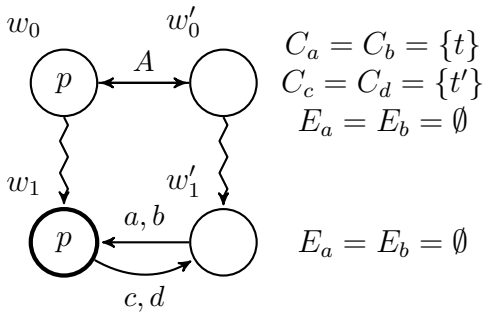

Figure 5.10: *Basic trial III after* $\mathsf{con}_t \Uparrow_a, \mathsf{con}_t \Uparrow_b, \mathsf{con}_{t'} \Uparrow_c, \mathsf{con}_{t'} \Uparrow_d$

This network has now reached an interesting state: as two friends a and b both believe in p and two friends c and d both believe in $\neg p$, the agents will never change their beliefs in the future. Two friends form a team that is too strong for a third friend with opposite belief to change the beliefs of the two friends. See section 5.4.2 for the explanation of this phenomenon. Since the network will be stable throughout any update, the network has *finished learning*. Exactly half of the network believes in the correct theory p, so the network has *semi-successfully finished learning*.

5.4 Comparing actual models

In this section I will compare some specific network configurations. As mentioned before, I will not draw each model nor describe what happens at every step. I will present a table with the outcomes of different network configurations and informally describe what conclusions we can draw from the analysis.

For the analysis I use networks consisting of five agents. With five agents the number of friends in a complete graph is uneven, which gives more decisive power (see section 5.3). With less than five agents either the number of friends in a complete graph is even (in the case of four agents) or we cannot distinguish a circle from a complete graph (for less than four agents). A network with 5 agents is still relatively small such that the calculations are relatively simple. I will compare two different structures: the circle and the complete graph,[22] see Figure 5.11. I will also vary with a bias of weight 1 and bias of weight 2, and see how this influences the difference between the circle and the complete graph.

[22]This is a fair comparison, because in both cases the communication is equally divided amongst the agents. Zollman ([42]) and Bala and Goyal ([3]) also analysed 'the wheel'. However, as argued by Zollman himself, the difference in the results between the wheel network and the complete network can also be caused by the fact that the wheel network has unequal connections: the hub is connected to everyone while the others are not.

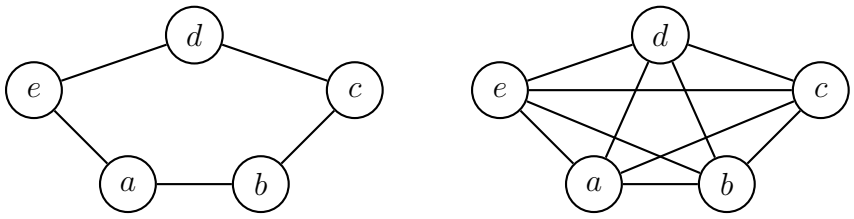

Figure 5.11: *A 5 person-circle and a 5 person-complete graph*

5.4.1 The results

Since I already suspect that the distribution of biasses and failures will affect the behaviour of the agents, I have modelled different distributions. In this section, I will present the results of some specific distributions. Let me emphasise that the results below highlight some extraordinary behaviours, that show the differences between the different configurations. These results should not be regarded as statistical data, such as the results of Zollman in [42]. These results are the motivation behind the informally described effects in the next section. As I have argued in chapter 4, the good thing about using modal logic compared to statistical tools, is exactly that I can focus on the details of behaviour adoption, instead of studying the results of 10.000 trials.

For that reason, I will present the results of the following cases in Table 5.1 below (let 'i observes φ or $\top$' be short for 'i observes φ if $\neg B_i \neg \varphi$, else i observes $\top$):

A: **Biasses**: $C_a = C_b = C_c = C_d = \{t\}$, $C_e = \{t'\}$
Experimental results: 0

B: **Biasses**: $C_a = C_b = C_c = \{t\}$, $C_d = C_e = \{t'\}$
Experimental results: 0

C: **Biasses**: $C_a = C_b = \{t\}$, $C_c = C_d = C_e = \{t'\}$
Experimental results: 0

D: **Biasses**: $C_a = C_c = C_e = \{t\}$, $C_b = C_d = \{t'\}$
Experimental results: Round 1: all agents $i \in A$ observe $\neg p$ or $\top$.
Round 2: a, e observe p or $\top$ and b, c, d observe $\neg p$ or $\top$.

E: **Biasses**: $C_b = C_d = \{t\}$, $C_a = C_c = C_e\{t'\}$
Experimental results: Round 1: all agents $i \in A$ observe p or $\top$.
Round 2: a, e observe $\neg p$ or $\top$ and b, c, d observe p or $\top$.

The following distributions are derived from the distributions above by multiplying the biasses by 2:

A*: **Biasses**: $C_a = C_b = C_c = C_d = \{t, t\}$, $C_e = \{t', t'\}$
Experimental results: 0

B*: **Biasses**: $C_a = C_b = C_c = \{t, t\}$, $C_d = C_e = \{t', t'\}$
Experimental results: 0

C*: **Biasses**: $C_a = C_b\{t, t\}$, $C_c = C_d = C_e = \{t', t'\}$
Experimental results: 0

D*: **Biasses**: $C_a = C_c = C_e = \{t, t\}$, $C_b = C_d = \{t', t'\}$
Experimental results: Round 1: all agents $i \in A$ observe $\neg p$ or $\top$.

Round 2: a, e observe p or $\top$ and b, c, d observe $\neg p$ or $\top$. Round 3: a, e observe p or $\top$ and b, c, d observe $\neg p$ or $\top$.

E*: **Biasses**: $C_b = C_d = \{t, t\}$, $C_a = C_c = C_e\{t', t'\}$
Experimental results: Round 1: all agents $i \in A$ observe p or $\top$.
Round 2: a, e observe $\neg p$ or $\top$ and b, c, d observe p or $\top$. Round 3: all agents $i \in A$ observe p or $\top$.

I will abbreviate the outcomes (see section 5.2 for the meanings of these outcomes):

$$\begin{aligned} \text{SFL} &:= \text{Successful Finished Learning} \\ \text{SSFL} &:= \text{Semi-Successful Finished Learning} \\ \text{FFL} &:= \text{Failed Finished Learning} \end{aligned}$$

In Table 5.1, the following explanations are given to specify roughly what the main cause for the outcome is:[23]

(1) := In the finished state, the network is split up into two groups of friends that each believes in the opposite theory.
(2) := The distribution of the biasses is such that more than half of the agents have $t_i < t'_i$ or $t_i > t'_i$.
(3) := The distribution of failures is extraordinary.

Consider Table 5.1 that presents the results of trials A, B, C, D, E, A*, B*, C*, D* and E* for both a circle and a complete graph:

	A	B	C	D	E
circle	SFL,(2),0	SSFL,(1),0	FFL,(1,2),0	FFL,(1,3),2	SSFL,(1),2
complete	SFL,(2),0	SFL,(2),0	FFL,(2),0	SFL,(2),0	FFL,(2),0

	A*	B*	C*	D*	E*
circle	SFL,(2),0	SSFL,(1),0	FFL,(1,2),0	FFL,(1,3),3	SSFL,(1),3
complete	SFL,(2),0	SFL,(2),0	FFL,(2),0	SFL,(2),0	FFL,(2),0

Table 5.1: *The results of the analysis. Let X, Y, Z be such that $X :=$ the outcome, $Y :=$the rough explanation and $Z :=$ the number of rounds it takes to reach a state of finished learning*

5.4.2 Effects described

There are a couple of remarkable behaviours that I will discuss in this section. The most striking result is that all trials of the complete graph do not need any round of conducting experiments and upgrading to get to a state of finished learning. Because of the connections in the network, everybody is friends with everybody, so all beliefs are upgraded according to the evidence and bias set of the entire network. Since the

[23]Since p is the true theory, it is "extraordinary" when observations of $\neg p$ (i.e., $o_{\neg p}$) occur more often than observations of p (i.e., o_p).

network has an uneven number, and I give all agents prior evidence that supports either p or $\neg p$, the agents directly choose between believing p or $\neg p$. Whether the entire community believes p or $\neg p$ depends on the distribution of the biasses. So in a complete network with an uneven number of agents, the distribution of the bias is crucial for whether the network will be successful or not.

In general, we see that the distribution of biasses and the distribution of priors has a great effect on the outcome of the trial. There are two cases where the circle network could overrule the distribution of biasses, namely in D and E. In D, more than half of the agents in the network got a bias in t, so every agent in the complete graph directly upgraded to believing p, achieving successful finished learning. However, in the circle an extraordinary sequence of experimental results moved the network to a state of failed finished learning, where three agents believed in $\neg p$ and only two in the true theory p. The same thing happened in E, only the other way around: biasses were distributed such that more than half of the agents in the network got a bias in t', drawing the complete graph to a state of failed finished learning. In E, the circle could repair the bad distribution of biasses. Note that the distribution of failures in E is not extraordinary, while the distribution of failures in D is (because it involves a lot of observations of the false theory). So even though there are cases where the circle network causes the network to achieve successful learning as well as failed learning, the former is more realistic.

Another factor that greatly affects the adopting behaviour of agents in a circle, is the presence of two friends that have the same belief. We see the same result in [25]. For example in situation B, we have a, b and c believing in p and d and e believing in $\neg p$. Since the threshold θ is $\frac{1}{2}$, a p-believer needs two friends to have more t' than t to affect her doxastic state such that she believes $\neg p$ (which is a consequence of t'). Note that b is in the middle of two p-believers, so she is least likely to switch to the group of $\neg p$-believers. Let's consider a and c. Note that a and c both have only one $\neg p$-believer as a friend, $N_a e$ and $N_c d$. These friends cannot beat the beliefs of the two p-believing-friends of a and b. For this reason, agents a and c will never change their beliefs. The same holds for the duo d and e, who will never switch to believing that p. So in a circle network, whenever friends hold the same belief they will never change this belief. Moreover, if there exists two such groups of friends in a circle network with opposing beliefs, then the group will never reach a state of successfully finished learning.

Finally, we see that increasing the weights of the biasses from 1 to 2 does not have a big impact on the behaviour of the agents. Neither increasing the weight even more will have a big impact on the models. Only in settings D and E the adopting behaviour was slightly slowed down after the weights of the biasses were increased, but the final outcome stayed the same.

Chapter 6

Conclusion

In this thesis, I studied the effect of different network configurations on the reliability of scientific communities. In the philosophical framework, I have argued that science is a social product. On the one hand this means that experimental results can simply be wrong, because the scientists conducting the experiments are no perfect robots but social and subjective beings. On the other hand, outcome of research is also affected by the interaction between scientists. In the theoretical framework, I introduced Zollman's hypothesis that transient diversity guarantees a high reliability, and that this is achieved by either limiting the communication between agents or by strenghtening the priors.

By introducing a new Multi-agent Dynamic Evidence-based Logic, I aimed at being able to say under what circumstances a network is more likely to recover a false group-belief. Even though the model is built on many simplifications and no general conclusions can be drawn, there are some interesting features that can be investigated in more detail in further research. Firstly, the distribution of priors and failures greatly affects the outcome of the network. Especially, in a complete network with an uneven number of agents, the distribution of the bias is crucial for whether the network will be successful or not. Secondly, in some circumstances the circle network is better at tracking the truth, and in some circumstances the complete graph is. However, the circumstance where the circle is more realistic is more realistic. Thirdly, whenever friends in a circle network hold the same belief, they will never change this belief. This implies that in my setting, cognitive diversity does not per se improve the truth-tracking abilities of a community. Finally, in my setting, increasing the weights of the biasses does not have a strong impact on the behaviour of the agents.

Summarizing the above results, I conclude that whether a scientific community will track the truth or not, depends more on the distribution of the biasses and failures than on the network structure or weights of the bias. Previous studies by Zollman ([41, 42]) and Bala and Goyal ([3]) suggest that the effect of the latter factors is more explicit. The reason for this difference is most likely that I have used a different system than Zollman and Bala and Goyal to analyse the behaviour of epistemic agents in a network. In section 5.1.1 I explained that logic is not suitable to analyse Bandit problems in the way Zollman did, so I had to make some adjustments to the context. In future research, one could investigate the option to use probabilistic

logic or multi-value logic to deal with the Bandit problem.

The outcomes of the analysis in chapter 5 rely not only on the distributions of the biasses and failures, but also greatly on the set-up of the semantics of MDEL. In future research, one can adjust the semantics (and the language), aiming for the most realistic set-up. For example, I can adjust the value of the threshold θ or the preconditions for the actions. I can also add private actions to the language. The case-study of the Einstein-De Haas effect suggest that the content of communication is crucial. In my system, agents communicate that they have a piece of evidence t, namely $E_i(t, m)!_i$. In future research, I could investigate what happens to the reliability of communities when the agents communicate t or Bp. Furthermore, I can generalize the model in such a way that it allows for more variation. For example, I could give relative weights to different kinds of evidence (e.g. testimony from a respected professor versus testimony from an inexperienced student).

The system I have presented can be quite easily transformed into a complete proof system, adding the relevant theorems and principles. A completeness proof could be constructed by first giving the completeness for the static logic MSEL and then prove completeness of dynamic models by reduction of all dynamic sentences into static sentences, as shown in [37].

Appendix A

The Logic of Dynamic Justified Belief: Details

A.1 Syntax

Definition A.1.1. (Language JB) Given a set Φ of atomic sentences, the language $\mathscr{L} := (\mathscr{T}, \mathscr{F})$ consists of the set $\mathscr{T}$ of *evidence terms* t and the set $\mathscr{F}$ of *propositional formulas* (sentences) φ defined by the following double recursion:

$$\varphi ::= \bot | p | \neg\varphi | \varphi \wedge \varphi | Et | t \gg \varphi | \Box\varphi | K\varphi | Y\varphi \text{ with } p \in \Phi$$

$$t ::= c_\varphi | t \cdot t | t + t$$

The set $\mathsf{sub}(t)$ of *subterms* of a term t is defined by induction on the construction of t as follows:

$$\begin{aligned} \mathsf{sub}(c_\varphi) &= \{c_\varphi\} \\ \mathsf{sub}(s \cdot u) &= \{s \cdot u\} \cup \mathsf{sub}(s) \cup \mathsf{sub}(u) \\ \mathsf{sub}(s + u) &= \{s + u\} \cup \mathsf{sub}(s) \cup \mathsf{sub}(u) \end{aligned}$$

The set $\mathsf{sub}(\varphi)$ of *subformulas* of a formula φ is defined by induction on the construction of φ as follows:

$$\begin{aligned} \mathsf{sub}(\bot) &= \{\bot\} \\ \mathsf{sub}(p) &= \{p\} \\ \mathsf{sub}(\neg\theta) &= \{\neg\theta\} \cup \mathsf{sub}(\theta) \\ \mathsf{sub}(\theta \wedge \theta') &= \{\theta \wedge \theta'\} \cup \mathsf{sub}(\theta) \cup \mathsf{sub}(\theta') \\ \mathsf{sub}(Et) &= \{Et\} \\ \mathsf{sub}(t \gg \theta) &= \{t \gg \theta\} \\ \mathsf{sub}(\Box\theta) &= \{\Box\theta\} \cup \mathsf{sub}(\theta) \\ \mathsf{sub}(K\theta) &= \{K\theta\} \cup \mathsf{sub}(\theta) \\ \mathsf{sub}(Y\theta) &= \{Y\theta\} \cup \mathsf{sub}(\theta) \end{aligned}$$

The operation $(\cdot)^Y : \mathscr{T} \cup \mathscr{F} \to \mathscr{T} \cup \mathscr{F}$ is defined by setting for terms:

$$\begin{aligned} (c_\varphi)^Y &:= c_{(\varphi^Y)} \\ (t \cdot s)^Y &:= t^Y \cdot s^Y \\ (t + s)^Y &:= t^Y + s^Y \end{aligned}$$

and for formulas:

$$\begin{aligned}
\bot^Y &:= \bot \\
p^Y &:= p \\
(\neg\varphi)^Y &:= \neg\varphi^Y \\
(\varphi \wedge \psi)^Y &:= \varphi^Y \wedge \psi^Y \\
(Et)^Y &:= Et^Y \\
(t \gg \varphi)^Y &:= t^Y \gg \varphi^Y \\
(\Box\varphi)^Y &:= Y\Box\varphi \\
(K\varphi)^Y &:= YK\varphi \\
(Y\varphi)^Y &:= YY\varphi
\end{aligned}$$

A.2 Semantics

Definition A.2.1. (Model for JB) A *model* $M = (W, [\![\cdot]\!], \sim, \geq, \rightsquigarrow, E)$ is a structure consisting of a nonempty set W of *possible worlds*; a *valuation map* $[\![\cdot]\!] : \Phi \to \mathcal{P}(W)$; binary relations $\sim$ ("epistemically indistinguishable from"), $\geq$ ("no more plausible than"), and $\rightsquigarrow$ ("is the temporal predecessor of") on W; and an *evidence map* $E : W \to \mathcal{P}(\mathscr{T})$, satisfying the following conditions:

- $\sim$ is an equivalence relation (i.e., reflexive, symmetric and transitive) and $\geq$ is a preorder (i.e., reflexive and transitive).
- *Indefeasibility*: $w \geq v \Rightarrow w \sim v$
- *Local Connectedness*: $w \sim v \Rightarrow (w \geq v \vee v \geq w)$
- *Propositional Perfect Recall*: $(w \rightsquigarrow v \sim v') \Rightarrow \exists w'(w \sim w' \rightsquigarrow v')$ (i.e., knowledge of yesterday is still known today")
- *Evidential Perfect Recall*: $w \rightsquigarrow w' \Rightarrow \{t^Y | t \in E(w)\} \subseteq E(w')$ (i.e., evidence of yesterday is still evidence today)
- *Uniqueness of Past*: $(w' \rightsquigarrow w \wedge w'' \rightsquigarrow w) \Rightarrow w' = w''$
- *Persistence of Facts*: $w \rightsquigarrow w' \Rightarrow (w \in [\![p]\!] \Leftrightarrow w' \in [\![p]\!])$ for $p \in \Phi$
- *(Implicit) Evidential Introspection*: $w \sim v \Rightarrow E(w) = E(v)$ (i.e., agents know what is in their evidence set)
- *Subterm Closure*: If $t \cdot t' \in E(w)$ or $t + t' \in E(w)$, then $t \in E(w)$ and $t' \in E(w)$ (i.e., a compound evidence is available to the agent only if its component pieces of evidence are available)
- *Certification of Evidence*: If $t \in E(w)$ and $t \gg \varphi$ then $c_\varphi \in E(w)$ (i.e., every actual evidence in support of a sentence φ can be converted into a canonical piece of evidence c_φ that certifies it, implying that all explicit knowledge can be certified)

Definition A.2.2. (Truth for JB) The *satisfaction relation* $w \vDash \varphi$, short for $(M, w) \vDash$

φ when M is fixed, is defined as follows:

$$\begin{array}{ll} w \not\models \bot & \\ w \vDash p & \text{iff } w \in [\![p]\!] \\ w \vDash \neg\varphi & \text{iff } w \nVdash \varphi \\ w \vDash \varphi \wedge \psi & \text{iff } w \vDash \varphi \text{ and } w \vDash \psi \\ w \vDash Et & \text{iff } t \in E(w) \\ w \vDash t \gg \varphi & \text{iff } t \gg \varphi \\ w \vDash \Box\varphi & \text{iff } v \vDash \varphi \text{ for every } v \leq w \\ w \vDash K\varphi & \text{iff } v \vDash \varphi \text{ for every } v \sim w \\ w \vDash Y\varphi & \text{iff } v \vDash \varphi \text{ for every } v \rightsquigarrow w \end{array}$$

We can extend the valuation map $[\![\cdot]\!]$ to all sentences φ, for putting $[\![\varphi]\!] = \{w \in W | w \vDash \varphi\}$.

Definition A.2.3. (Belief) We define belief: $B\varphi := \neg\Box\neg\Box\varphi$ as truth in the most plausible worlds:

$$M, w \models B\varphi \text{ iff } M, w' \models \varphi \text{ for all } w' \in min\{w' \in W | w \sim w'\}$$

A.3 Evidence dynamics

Definition A.3.1. (Language DJB) $\mathscr{L}^{act} := (\mathscr{T}^{act}, \mathscr{F}^{act})$ is the extension of the static language for JB (see definition A.2.1) obtained by adding modal operators α for epistemic actions $\alpha \in \{t+, t \otimes s, t!, t \Uparrow\}$, for every $t, s \in \mathscr{T}$. The notions of subterm, subformula, admissibility and model are lifted to $\mathscr{L}^{act}$ in the obvious way.

Definition A.3.2. (Preconditions and evidence set for DJB) For every action α, define a sentence pre_α, called the *precondition* of α, and a set of terms $\mathscr{T}(\alpha)$ called the *evidence set* of α:

$$\begin{array}{ll} \mathsf{pre}_{t+} = \mathsf{pre}_{t\Uparrow} & := \top \\ \mathsf{pre}_{t!} & := \mathsf{con}_t = \bigwedge\{\theta | t \gg \theta\} \\ \mathsf{pre}_{t\otimes s} & := Et \wedge Es \\ \mathscr{T}(t+) = \mathscr{T}(t!) = \mathscr{T}(t \Uparrow) & := \mathsf{sub}(t) \cup \{c_\theta | s \gg \theta \text{ for some } s \in \mathsf{sub}(t)\} \\ \mathscr{T}(t \otimes s) & := \{t \cdot s\} \cup \{c_\theta | t \cdot s \gg \theta\} \end{array}$$

Definition A.3.3. (Truth for DJB) Let w^α denote the ordered pair (w, α) to represent the "updated" world resulting from performing action α in world w. Then:

$(M, w) \vDash [\alpha]\varphi$ iff $(M[\alpha], w^\alpha) \vDash$ with $M[\alpha] := (W^\alpha, [\![\cdot]\!]^\alpha, \sim^\alpha, \geq^\alpha, \rightsquigarrow^\alpha, E^\alpha)$, and

$$
\begin{aligned}
W^\alpha &:= W \cup \{w^\alpha | w \in [\![\mathsf{pre}_\alpha]\!]\} \\
E^\alpha(w) &:= E(w) \text{ for } w \in W \\
E^\alpha(w^\alpha) &:= \{u^Y | u \in \mathscr{T}(\alpha) \cup E(w)\} \\
[\![p]\!]^\alpha &:= [\![p]\!] \cup \{w^\alpha \in W^\alpha | w \in [\![p]\!]\} \\
\sim^\alpha &:= \sim \cup \{(w^\alpha, v^\alpha) | w \sim v\} \\
\rightsquigarrow^\alpha &:= \rightsquigarrow \cup \{(w, w^\alpha) | w \in [\![\mathsf{pre}_\alpha]\!]\} \\
\geq^\alpha &:= \geq \cup \{(w^\alpha, v^\alpha) | w \geq v\} \text{ for } \alpha \in \{t+, t \otimes s, t!\} \\
\geq^{t\Uparrow} &:= \geq \cup \{(w^{t\Uparrow}, v^{t\Uparrow}) | (w \notin [\![\mathsf{con}_t]\!] \wedge v \in [\![\mathsf{con}_t]\!] \wedge w \geq v)\} \text{ for } t \in \mathscr{T}^e \\
\geq^{t\Uparrow} &:= \geq \cup \{(w^{t\Uparrow}, v^{t\Uparrow}) | w \geq v\} \text{ for } t \notin \mathscr{T}^e
\end{aligned}
$$

Appendix B

Bibliography

[1] Sergei N. Artemov. The logic of justification. *Review of Symbolic Logic*, 1(4):477–513, 2008.

[2] Sergei N. Artemov and Melvin Fitting. Justification logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/archives/fall2012/entries/logic-justification/`, fall 2012 edition.

[3] Venkatesh Bala and Sanjeev Goyal. Learning from neighbours. *Review of Economic Studies*, 65:565–621, 1998.

[4] Venkatesh Bala and Sanjeev Goyal. A noncooperative model of network formation. *Econometrica*, 68(5):1181–1229, 2000.

[5] Alexandru Baltag, Zoé Christoff, Jens U. Hansen, and Sonja Smets. Logical models of informational cascades. In Johan van Benthem and Fenrong Liu, editors, *Logic across the University: Foundations and Applications - Proceedings of the Tsinghua Logic Conference*, volume 47, pages 405–432. College Publications: London, 2013.

[6] Alexandru Baltag, Zoé Christoff, Rasmus K. Rendsvig, and Sonja Smets. Dynamic epistemic logics of diffusion and prediction in social networks. Draftpaper, April 2015.

[7] Alexandru Baltag and Larry S. Moss. Logic for epistemic programs. *Synthese*, 139(2):165–224, 2004.

[8] Alexandru Baltag, Larry S. Moss, and Slawomir Solecki. The logic of common knowledge, public announcement, and private suspicions. *Technical report, Indiana University*, 1999.

[9] Alexandru Baltag, Bryan Renne, and Sonja Smets. The logic of justified belief change, soft evidence and defeasible knowledge. In L. Ong and R. de Queiroz, editors, *Proceedings of the 19th Workshop on Logic, Language, Information and Computation*, volume 7456 of *Lecture Notes in Computer Science*, pages 168–190. Springer-Verlag Berlin Heidelberg, 2012.

[10] Alexandru Baltag and Sonja Smets. Probabilistic dynamic belief revision. In Johan van Benthem, Shier Ju, , and Frank Veltman, editors, *Proceedings of LORI'07*. College Publications: London, 2007.

[11] Alexandru Baltag and Sonja Smets. A qualitative theory of dynamic interactive belief revision. In Giacomo Bonanno, Wiebe van der Hoek, and Michael Wooldridge, editors, *Logic and Games 3*, pages 13–60. Amsterdam University Press, 2008.

[12] Alexandru Baltag, Hans P. van Ditmarsch, and Larry S. Moss. Epistemic logic and information update. *Elsevier*, 8:361–455, 2008.

[13] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. Learning from the behavior of others: Conformity, fads, and informational cascades. *Journal of Economic Perspectives*, 12(3):151–170, 1998.

[14] Harry M. Collins. The meaning of experiment: Replication and reasonableness. In Lisa Appignanesi and Hilary Lawson, editors, *Dismantling Truth: Science in Post-Modern Times*. London: Weidenfeld, 1988.

[15] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.

[16] V. Ya Frenkel. On the history of the einstein-de haas effect. *Sov. Phys. Usp*, 22(7):580–587, 1979.

[17] Peter L. Galison. *How Experiments End*. Chicago: University of Chicago Press, London, 1987.

[18] Edmund L. Gettier. Is justified true belief knowledge? *Analysis*, 23:121–123, 1963.

[19] Alvin I. Goldman. *A Guide to Social Epistemology*, chapter 1, pages 11–37. Oxford University Press, 2011.

[20] Alvin I. Goldman and Thomas Blanchard. Social epistemology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/archives/sum2015/entries/epistemology-social/`, summer 2015 edition.

[21] Pelle G. Hansen and Vincent F. Hendricks. *Infostorms: How to take Information Punches and Save Democracy*. New York: Copernicus Books, 2013.

[22] Philip Kitcher. The division of cognitive labor. *Journal of Philosophy*, 87:5–22, 1990.

[23] Thomas S. Kuhn. The structure of scientific revolutions. *Science*, 136(3518):760–764, 1962.

[24] Thomas S. Kuhn. Collective belief and scientific change. In *The Essential Tension*, pages 320–339. Chicago: University of Chicago Press, 1977.

[25] Fenrong Liu, Jeremy Seligman, and Patrick Girard. Logical dynamics of belief change in the community. *Synthese*, 191:2403–2431, 2014.

[26] Helen Longino. The social dimensions of scientific knowledge. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/archives/spr2015/entries/scientific-knowledge-social/`, spring 2015 edition.

[27] Conor Mayo-Wilson, Kevin J.S. Zollman, and David Danks. The independence thesis: When individual and social epistemology diverge. *Philosophy of Science*, 78(4):653–677, 2011.

[28] Bradley Monton and Chad Mohler. Constructive empiricism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/archives/spr2014/entries/constructive-empiricism/`, spring 2014 edition.

[29] Andy Pickering. Against putting the phenomena first: the discovery of the weak neutral current. *Studies in History and Philosophy of Science Part A*, 15(2):85–117, 1984.

[30] Bryan Renne. Evidence elimination in multi-agent justification logic. In *TARK*, 2009.

[31] Bryan Renne. Multi-agent justification logic: Communication and evidence elimination. *Synthese*, 185(S1):43–82, 2012.

[32] Jeremy Seligman, Fenrong Liu, and Patrick Girard. Facebook and the epistemic logic of friendship. In *TARK*, 2013.

[33] Michael Strevens. The role of the priority rule in science. *Journal of Philosophy*, 100:55–79, 2003.

[34] Cass R. Sunstein. *Deliberating Groups versus Prediction Markets (or Hayek's Challenge to Haberbas)*, chapter 14, pages 314 – 337. Oxford University Press, 2011.

[35] David R. Topper. *Quirky Sides of Scientists: True Tales of Ingenuity and Error from Physics and Astronomy*, chapter 1. Tenacity and Stubbornness: Einstein on Theory and Experiment. Springer Science, 2007.

[36] Johan van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 14(2):1–26, 2004.

[37] Johan van Benthem, Jelle Gerbrandy, and Barteld Kooi. Dynamic update with probabilities. *Studia Logica*, 93:67–96, 2009.

[38] Johan van Benthem and Eric Pacuit. Dynamic logics of evidence-based beliefs. *Studia Logica*, 0:1–31, 2011.

[39] Hans P. van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. Dynamic epistemic logic. *Synthese*, 337, 2007.

[40] John Vickers. The problem of induction. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/search/searcher.py?query=problem+of+induction`, spring 2014 edition.

[41] Kevin J.S. Zollman. The communication structure of epistemic communities. *Philosophy of Science*, 74(5):574–587, 2007.

[42] Kevin J.S. Zollman. The epistemic benefit of transient diversity. *Erkenntnis*, 72(1):17–35, 2010.

[43] Kevin J.S. Zollman. Network epistemology: Communication in epistemic communities. *Philosophy Compass*, 8(1):15–27, 2013.

[44] Kevin J.S. Zollman. Learning to collaborate. Draftpaper, July 2014.