Social Interaction
A Formal Exploration

Social Interaction
A Formal Exploration

Proefschrift
ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 22 mei 2015 om 14.15 uur
door

Dominik Klein

geboren op 3 april 1983 te Bonn-Bad Godesberg,
Duitsland

Promotores:	Prof. dr. S. Hartmann
	Prof. dr. J.M. Sprenger
Copromotor:	Dr. E.J. Pacuit
Overige Leden:	Dr. H.C.K. Heilmann
	Dr. R.A. Muskens
	Prof. dr. J.W. Romeijn
	Prof. dr. S.J.L. Smets

Abstract

My work is built around the application of formal methods in philosophy. The idea that formal tools, logic in particular, could help advance the philosophic endeavor is not new. It dates back to antiquity. However, it is only recently that this idea has gained new momentum. By now, the use of formal methods in philosophy is not restricted to logic anymore, but includes a large variety of different methods and techniques. In this thesis, I concentrate on three of these: logical and statistical methods and computer simulations. In doing so, I follow two main goals. The first is to study these frameworks abstractly in order to get a better understanding of their qualities and drawbacks. The second goal then is to apply these tools to particular issues in philosophy. Here, I mainly focus on applications in epistemology and political philosophy.

This thesis consists of five chapters, each of which constitutes an autonomous scientific paper. The first chapter contains a logical framework for modeling the beliefs in interactive game situations and the dynamics thereof. In the second, I compare different logical frameworks with respect to two criteria relevant for choosing between different modeling tools. The third chapter contains a statistical model on group decision making. I present a model for aggregating individual judgments that is sensitive to differences in competence between the different agents. The fourth chapter contains a mathematical model on voting behavior and opinion dynamics, based on the idea that voters interests are driven by an underlying agenda of topics. Finally, in the fifth and last chapter, I present a simulation model on the emergence and maintenance of trust in larger societies.

Contents

Chapter 1

Introduction

The use of formal tools is a recent and quickly spreading phenomenon in philosophy and some of its neighboring fields. Formal methods appear in the names of "mathematical philosophy" and "formal epistemology", two recent subfields of philosophy, but also in the wide spread application of game theory or computer simulations in philosophy, sociology or political science. But *what* is it that people hope to gain from using formal tools? And *how* are these methods applied exactly? As it turns out, there is no unique answer to either of these questions. Just to the contrary, there is a wide variety of formal approaches differing, for instance, in the tools they employ, the situation they address or the goals they pursue. In this thesis, I will present five applications of formal tools that illustrate the wide range of formal methods in present day philosophy. But before going into detail, let's have a slightly more systematic look at some aspects in which formal approaches can differ. In the following, we introduce four such aspects: the formal methods used, the target system of the model, the way in which the model relates to that target system and, finally, the precise goals pursued by a formalization.

The first aspect is the particular framework used for a formal model. Current literature produced a plethora of different modeling frameworks: Quantitative methods, qualitative methods, logic, probabilistic approaches, mathematical models, game theory, rational choice theory, Bayesianism, network theory and computer simulations to name but a few. Notably, these different frameworks can be related in various possible ways. They could be incompatible with each other, they could be combinable with each other or some framework could be a specification or subfield of another. Qualitative and quantitative models, for instance, are incompatible, describing mutually exclusive ways of addressing some target system. Game theory, on the other hand, is a general framework

for representing interactive strategic situations that is compatible with different formal frameworks, using probabilistic or logical methods. In the extreme case, the exact relationship between two frameworks depends on the standpoint of the observer. Seen from a taxonomy of scientific fields, logic would be classified as a subfield of mathematics, while sociologically these are two different fields. Logicians work in different institutes than mathematicians, label themselves differently, apply different methods and, sometimes even have a different way of thinking about abstract systems.

The second aspect we introduce is the target system of a formal model. Each formalization or formal model relates to some target system. These target systems could, in principle, be about anything, a philosophic argument, a formal or informal theory, a concrete social situation, a particular concept, a piece of data, a social practice, or even another formal model. Moreover, a formalization may relate to various such target systems at once, or even remain intentionally opaque about the intended target system. For instance, a formalization of some philosophic theory about the nature of knowledge may be treated as a representation of that theory and, at the same time, as a formal model for the practice of knowledge. In both roles, the formal framework can be praised or criticized for a more or less of accuracy, it can be compared to other models, subjected to criticism, tested in various ways and so on. And of course, what is faithful to the underlying theory need not be very accurate in tracking the phenomenon of knowledge and vice versa, thus the different possible target systems might trigger contradictory evaluations of the same formal model.

Thirdly, different formalizations do not only differ in the nature of their target systems, but also in *how* they relate to them. Some formalizations aim to represent some phenomenon in its entirety, others only aim at particular aspects that, in reality, never appear in isolation. On a related note, some formalizations aim at being mere abstractions, in principal adequate representations of a target system, that merely abstract away from some complicating factors. Others are idealizations, intentionally omitting or misrepresenting relevant aspects of their target system. In the extreme case, some might want to argue, certain formal studies may be mere parabolas, instructive stories that are meant to suggest some moral about an informal system.

And, as a fourth and last aspect, the function served by a formal model can vary. For instance, one function of a formal model can be to *clarify* certain properties of a target system. Having a representation in an adequate, well defined formal framework can help to discern certain properties of the target

system, such as, for instance, the logical structure of an argument. In a second step, formal frameworks may then be applied for a *verifying* or controlling purpose, checking whether some informally given argument is in fact conclusive or whether some mechanism produces the results it is believed to produce. And, as a third and last aspect, formal representation can help to *explore* the properties of some target systems, for instance by using computers to replicate some dynamical processes under controlled input parameters or by applying a highly developed mathematical apparatus.

In this thesis, we will present five applications of formal tools, differing in the four aspects we have just presented: their formal frameworks, their target systems, their relations to the respective target system and their modeling goals. These applications will be put forward in the next five chapters. Taken together, they give a snapshot on the wide range of formal models in contemporary philosophy. Each application covers a different area of interest, thus the individual chapters can be accessed independently of each other. In principal, this thesis is intended to be self contained, although some passages, especially the proof sections, may assume some familiarity with the underlying formal frameworks of epistemic and doxastic logic (chapter 2 and 3), probability theory (chapter 4), basic linear algebra (chapters 4 and 5), computer simulations (6) and game theory (chapters 2 and 6). In the concluding chapter 7, finally, we will offer some general remarks about the use of formal tools in philosophy and their potential roles and purposes. For a start, we present an outline of the different chapters.

In the first chapter 2, we assume a logical perspective on epistemic game theory. Traditionally, game theory aims to identify the strategic structure of social situations, which agent has which strategies, how these strategies relate to each other and so forth. However, knowledge of the underlying game structure might not always be sufficient for predicting the behavior of different agents, or advising some player on how *ideally* to move next. The optimal behavior in many situations, be they cooperative or competitive, depends on what the other player thinks, wants or does. Therefore, so the starting idea of epistemic game theory, it is not enough to learn about the game structure, but any successful strategic analysis needs to take the opponents' moves and strategies into account. However, there is a certain complication hidden in this picture. We could expect a rational opponent to adopt a similar strategy, making her moves depend on what she expects us to do. Thus, in order to anticipate her behavior, we need to go *second order* and incorporate her potential beliefs about us in the analysis, that is, we need to form beliefs about her beliefs. Anticipating the opponent to

do likewise again leads us to third, fourth and higher levels of knowledge. In the limit, these chains of reasoning about each others' beliefs easily lead to *infinite* levels of mutual beliefs ascriptions, that can be represented with probabilistic [73] or logical [52] methods. Taking this perspective of epistemic game theory as a starting point, we want to add a further perspective to the analysis of games. Classically, epistemic game theory assumes the players' beliefs to be externally given and static. But this, of course, is an idealization. The players' beliefs have emerged through some dynamic epistemic progress, and they might well continue to change *during* the game. Of course, the players' beliefs will change as the game goes on. Each move made by one of the player is a new piece of information influencing the beliefs of all other players. But also events *external* to the actual game can impact the players beliefs and expectations. For instance, some accidental side comment, dropped voluntarily or involuntarily by one of the players, may change the way in which other players perceive the game. Consequentially, so the starting assumption of our model, we need a *logic of change*, a formal framework to incorporate external informational events into epistemic game theory. In chapter 2 we develop exactly that. Starting from a logical model of higher order information, we develop a mechanism for incorporating new informational events into the players' epistemic states. This mechanism is based on product updates, a tool from epistemic logic developed for incorporating external informational events into Kripke Models. On a more conceptual level, our model thus broadens the target system of epistemic game theory by incorporating the agents' belief *dynamics* into the game model.

In our second chapter 3, we explore various formal frameworks for multi agent information, referring again to first and higher order levels of information. Social situations such as arranging a dinner with friends will crucially depend on the information available to the different agents. To successfully arrange a joint dinner, all guests, of course, need to know about the time and place of the convention. But this *first order* information will, in general, not suffice to make people actually show up at a dinner. The different invitees may want to make sure that they are not the only ones coming. That is, they want to have the *higher order* information that also the other guests know about the proposed date and time and are also planning to come. Thus, a formal representation of the relevant informational attitudes for social interaction needs to incorporate first and higher order informational settings. In chapter 3, we explore different formal frameworks for representing the informational states in such interactive situations, all based on epistemic and doxastic logic.

Notably, different social events will call for different formal frameworks. For instance, in the above case of friends coordinating for dinner, more information is always better, thus a suitable formal framework can concentrate on acquiring new information. Other situations might also be sensitive towards restricting the access to information. As an example, consider situations of secure communication. In communicating with our banks, one of the central concerns is that some third party, trying to eavesdrop *cannot* learn about the content of our messages. That is, we want to restrict the access to the available information. A corresponding formal framework thus needs to keep track of which information the agents can and cannot acquire. Generally speaking, the decision of which framework is best suited for a given situation will depend upon some characteristics of the target system, but also on the taste and the goals of the modeler.

In chapter 3, we identify two criteria relevant in choosing formal modeling tools, and use these criteria to compare different logical frameworks. The first criterion, expressive power, states that some framework is *fine enough* for a given target system, allowing to represent all features relevant for subsequent analysis. The second criterion, realizability, expresses that a framework is not *too fine*, for instance by making too subtle distinctions or allowing for lengthy, overly complex constructions that do not mirror any observable properties. Obviously, these two criteria drag in opposite directions. In choosing some formal framework for a particular situation, the ideal balance between these two depends, for instance, on the exact properties of the target system or the goals pursued by the modeler. Some situations require a finer formal language to be represented adequately than others. Equally, some modeling purposes are more sensitive towards differences in complexity than others. Our main aim in this chapter is to clarify the relation between these two properties, expressive power and realizability, and to relate them to some prominent representational frameworks.

The next chapter 4, deals with the problem of judgment aggregation. When confronted with major decisions, political authorities, economic boards or NGOs will often refer to external expertise about, for instance, future population growth, the expected sales of a product or the likelihood of some major ecological disaster. Usually, this external expertise is provided by a board of experts, appointed especially for that purpose. However, the different experts will, almost inevitably, disagree about the value in question, thus the decision maker faces the problem of how to combine the individual assessments. One of her

central questions is: Should she take *differences* in competence between the various experts into account. In this chapter, we propose a mathematical model for judgment aggregation, sensitive to differences in competence between experts. As it turns out, the practical problem of identifying expertise is extremely hard. In particular, a vast body of empirical research shows that properties such as status, reputation or self assessment are, at best, unrelated to the actual quality of some individual's judgments. With other words, the *actual* distribution of competence present in some given expert panel might be extremely difficult to determine. Therefore, we are not so much concerned with identifying the *ideal* weights for any particular expert panel, but in devising a model that *stably* outperforms non-differential methods, even under the presence of various complicating factors.

In judging a particular expert panel, the decision maker will face some uncertainty about the degrees of expertise, but also about the existence of biases or correlations between the different members. Taking this uncertainty about the target system into account, we primarily aim to identify a large parameter range where our framework fares better than non-differential rules. Ideally, this range covers all of the decision maker's uncertainty about the target system, such that she can rely on our method performing sufficiently well, even without having access to the precise characteristics of the target system.

The following chapter 5 focuses on voting behavior in political elections. We present a mathematical model of individual voters and how they choose between candidates in a major political election. While the idea of describing voters as boundedly rational actors aiming to maximize their own returns is as old as Schumpeter's [141], it is yet debated *how* to adequately represent the motivations of voters. Notably, there are two competing accounts in making sense of voters as rational actors, instrumental and expressive voting. The first of these, instrumental voting, sees voters as interested in the *outcome* of an election, deriving their utility from future political decisions. On the other hand, expressive voting holds that voters already obtain their gratification from voting from their preferred party or siding with some camp they support. Both of these approaches are well supported by theoretical and empirical arguments, yet both sides have their respective weak spots, certain prominent phenomena they cannot explain. There are several arguments, however, that actual voting behavior can best be explained by a superposition of instrumental and expressive voting, that is, voters caring about the attributes and values of some party as well as the resulting outcomes of an election. The relative weights attached

to both modes, expressive and instrumental, vary with, for instance, the exact policy at stake or how close the election is expected to be. Thus, in order to understand complex voting situations, it is best to first analyze these from both perspectives individually, in order to later combine the insights obtained. In chapter 5, we assume an expressive perspective on voting in elections where all voters and candidates are interested in a particular agenda of topics.

In the first part of this chapter, we use our framework to compare different voting systems such as plurality vote or approval voting. In particular, we are interested in how good the different voting systems are in fostering a high electoral turnout. That is, we are interested in the number or probability of abstentions under the different voting systems. In the second part of the chapter, we then add a dynamic module to our framework. The beliefs and preferences of voters, just as the beliefs in games studied in chapter 2, are not fixed once and for all. Just on the contrary, these preferences develop gradually, reacting to various pieces of information the voter obtains. In particular, recent political news, but also electoral campaigns or private discussions can change the voters' beliefs about the best course of action. We outline a particular formal tool, focus changing matrices, that integrates these various elements of opinion change into our discussion of voting behavior. In this part, we also show *how* a related logical model relates to our mathematical model, thus giving an example for the interaction of different formal frameworks.

Finally, in chapter 6 we present a computer simulation on the dynamics of generalized trust. Generalized trust, the belief that strangers we have never met before will act cooperatively in complex, interactive situations, is a major determinant for the success of modern societies. A high level of trust is relevant for the performance of political institutions and the economic capacities of states, but also for individual benefits such as health or the quality of life. Consequentially, recent years have seen an increased research interest in the architecture and determinants of trust. In chapter 6, we present an agent based computer simulation on the dynamics of trust, based on factors and mechanisms identified in previous empirical and theoretical research. The aims of our model are twofold. First, the simulation should help to obtain new insights in *how* different mechanisms relevant for the emergence of trust interact with each other. Here, our model will build on various factors and mechanisms identified in the theoretical and empirical literature on trust, but also on insights from experimental psychology and game theory. Second, our simulation should help to evaluate various claims from the theoretical literature. Is some proclaimed

mechanism strong enough to explain differences in trust? Does a certain parameter, say the mobility within a society, have the impact it is believed to have? These questions can be settled by implementing said mechanisms in a computer simulation in order to replicate the underlying dynamic system. In particular, our simulational model will relate to two different target systems. First, our simulations are targeted at the actual phenomenon of generalized trust in larger societies. In order to validate our simulation, we need to show that it tracks particular aspects of generalized trust sufficiently well. In a second step, we then shift target systems and use our simulations to test various theoretical predictions from the literature.

In chapter 7, finally, we conclude. We do so by offering some general remarks about the use of formal tools in philosophy. This chapter is primarily meant to provide some framing for the work presented in chapters 2 to 6 and to help the reader put this work into context. In our discussion, we return to the four aspects mentioned above: the modeling techniques available, the different possible target systems, the various ways in which a model could relate to such target systems and, fourth, the goals and motivations pursued by a formal model. We will further address the possible relationships between the various formal frameworks and discuss a variety of ways in which formal tools can relate to dynamic patterns. Finally, we will also mention some practical factors guiding the choice between different formal frameworks.

Most of the points we mention in this conclusion will likely be old news to people working with formal tools. They are, however, usually little addressed in the respective formal communities, making it unnecessarily difficult for interested outsiders to assess the exact scope, goals or motivation behind some formal models. We hope that the discussion we offer remedies this fact and thereby helps to assess some of the work presented in this thesis.

Chapter 2

Changing Types: Information Dynamics for Qualitative Type Spaces

2.1 Introduction

The central thesis of the epistemic program in game theory is that the basic mathematical model of a game situation should include an explicit parameter describing the players' *informational attitudes*.[1] See [28] for the relevant references and a discussion of the key results, and [131] for an introduction to this literature. Games are played in specific *informational contexts*, in which players have specific knowledge and beliefs about each other.[2] Many different formal models have been used to represent such informational contexts of a game (see [23, 156, 157], and references therein, for a discussion). In this chapter, we are not only interested in structures that describe the informational context of a game, but how these structures can *change* in response to the players' observations, communicatory acts or other dynamic operations of information change (cf. [154]).

We focus our attention on the players' *hard information* about the game

This chapter is based on joint work with E. Pacuit. It is an extended version of [91].

[1] This is, of course, something of a truism regarding games of *incomplete* or *imperfect* information. But the thesis is intended to apply to *all* game situations. See [29, Section 5] for a precise description about the crucial differences between an epistemic model of a game and a *Bayesian game*.

[2] This is nicely explained by Adam Brandenburger and Amanda Friedenberg ([30, pg. 801]): "In any particular structure, certain beliefs, beliefs about beliefs, ..., will be present and others won't be. So, there is an important implicit assumption behind the choice of a structure. This is that it is "transparent" to the players that the beliefs in the [type] structure – and only those beliefs – are possible.... The idea is that there is a 'context' to the strategic situation (e.g., history, conventions, etc.) and this 'context' causes the players to rule out certain beliefs."

(which we refer to as *knowledge* following standard terminology in the game theory and epistemic logic literature) and its dynamics. Broadly speaking, there are two different types of models that have been used to describe the players' knowledge (and beliefs) in a game situation. Both types of models include a nonempty set S of *states of nature* (elements of S are intended to represent possible outcomes of a game situation).[3] The first type of models are the so-called *Aumann-* or *Kripke-structures* [9, 53]. These structures describe the players' knowledge in terms of an *epistemic indistinguishability* relation over a (finite) set of states W. The second type of models are the knowledge structures of [52, 54], which are non-probabilistic variants of *Harsanyi type spaces* [73].[4] The key concept here is a **type** which describes the players' infinite hierarchy of knowledge (i.e., what the players know about the ground facts, what the players know about each others' knowledge of the ground facts, what players know about what the others know about each others' knowledge of the ground facts, and so on). The precise relationship between these two types of models was clarified in [52, 54].

Our goal in this chapter is to show how to adapt recent work modeling information change on Kripke structures as a product update with an *event model* [160] to the more general setting where the players' knowledge is represented using knowledge structures. To the best of our knowledge, this is the first attempt to develop a theory of information change for knowledge structures in the style of recent work on dynamic epistemic logic. Our main result (Theorem 2.25) characterizes precisely when a type in a fixed knowledge structure can be transformed into another type in that structure using the product update operation.

There are two main motivations for this technical study. The first is to explore generalizations of the product update operation. This is done in Section 2.3.1 where we also generalize a result of [158] characterizing when a Kripke structure can be transformed into another Kripke structure by a product update. The second motivation for this work is to initiate a study of information dynamics for epistemic models of games. Each player of a game can obtain new information about the game. Before the game, for instance, some player might gradually learn about the informational states of her opponents, their mutual relationships and what they think about the game. Also, that player might

[3] Often, it is assumed that the elements of S can be described by some logical language (for example, propositional logic), but this is not crucial for us in this chapter.

[4] See [145] for a modern introduction to type spaces as models of beliefs and [120] for a discussion of Harsanyi's classic paper.

acquire some new *factual* information, for instance about some face-down cards on the table. All these informational events may, of course, be relevant for her choosing a strategy. The player can only reasonably decide on which strategy to play after having incorporated all available information in her beliefs. Similarly, our player might also acquire new information *during* the play of an extended game. For instance, some opponent might accidentally drop her hand or a gust of wind may allow a subset of the players to see certain cards. Of course, also such external events may lead the player to revise her strategic considerations.[5] We agree that the type of events we have in mind here, gusts of wind and the like, are irrelevant to a game-theoretic analysis. But these events do change the *context*[6] of a game by revealing or hiding important information to all or some of the players and, more generally, changing their beliefs. This chapter is a first step towards a more general project that uses the dynamic epistemic logic framework to represent changes in the informational context of a game.

The remainder of this chapter is organized as follows. Section 2.2 provides the necessary background on (dynamic) epistemic logic and knowledge structures. Note that this Section was written for a reader already familiar with the key concepts and definitions. Consult [154] and [52] for motivations and a broader discussion of the literature. Our main result is in Section 2.3.2 with the technical preliminaries to be found in Section 2.3.1. We conclude in Section 2.4 with a discussion of topics for future research.

2.2 Background

2.2.1 A Primer on Dynamic Epistemic Logic

We assume the reader is familiar with the basics of (dynamic) epistemic logic, and so we only give the key definitions here (see the textbooks [53, 154] for an introduction to the subsequent definitions). Let I be the finite set of players and At a (finite or infinite) set of atomic propositions.[7]

Definition 2.1 (Epistemic Language)**.** The **epistemic language**, denoted

[5] This reasoning squares nicely with the many moments interpretation of extensive games, see [45]. The many moments interpretation holds that a player chooses anew at each of her action nodes in an extended game. In picking a strategy for an entire game, the player thus needs to predict which choices she will make at her future decision nodes, based on the information she envisages herself to have at the respective node. This expectation can, of course, turn out to be wrong, in which case she might want to move differently than anticipated.

[6] Here, we take the "context" of a game to be *all* events that influence the players' beliefs in the game situation.

[7] Atomic propositions are intended to represent properties of states of nature.

$\mathcal{L}_{EL}$, is the smallest set of formulas generated by the following grammar:

$$\varphi \ ::= \ p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi$$

where $p \in \mathsf{At}$ and $i \in I$. Define $L_i\phi$ as the dual of K_i (i.e., $L_i\phi := \neg K_i \neg\phi$) and the other boolean connectives (e.g., $\vee, \rightarrow$) as usual. $\triangleleft$

The intended interpretation of $K_i\phi$ is "agent i knows that ϕ (is true)". The standard semantics for $\mathcal{L}_{EL}$ are Kripke structures.

Definition 2.2 (Kripke Structure)**.** A **Kripke structure** (for a set of atomic propositions At) is a tuple $\langle W, \{R_i\}_{i\in I}, V\rangle$ where W is a set of states, $R_i \subseteq W \times W$ is an equivalence relation[8], and $V : \mathsf{At} \rightarrow \wp(W)$ is a valuation function. To simplify notation, we may write $w \in \mathcal{M}$ when $w \in W$. $\triangleleft$

Formulas of $\mathcal{L}_{EL}$ are interpreted at states in a Kripke model in the standard way, we only remind the reader of the definition for the knowledge modality:

$$\mathcal{M}, w \models K_i\phi \text{ iff } \text{ for all } v \in W \text{ if } wR_iv \text{ then } \mathcal{M}, v \models \varphi$$

The central idea of *dynamic* epistemic logic is to describe events that change a situation and the (uncertain) perceptions of these events by the agents as a so-called event model.

Definition 2.3 (Event Model)**.** An **event model** is a tuple $\langle E, \{Q_i\}_{i\in I}, \mathsf{pre}\rangle$ where E is a set of basic events, $Q_i \subseteq E \times E$ is an equivalence relation[9] and $\mathsf{pre} : E \rightarrow \mathcal{L}_{EL}$ assigns to each primitive event a formula that serves as a **precondition** for that event. We write $e \in \mathcal{E}$ if e is an event in $\mathcal{E}$. $\triangleleft$

The primitive events represent the basic observations available to the agents in a dynamic situation. Similar to Kripke structures, uncertainty about which events are taking place is represented by relations Q_i. Given our assumptions that each Q_i is an equivalence relation, the intended interpretation of eQ_if is that agent i cannot distinguish between events e and f. The key operation of *product update* describes how to incorporate into a Kripke structure $\mathcal{M}$ (describing an epistemic situation) the epistemic event described by an event model $\mathcal{E}$.

Definition 2.4 (Product Update)**.** The **product update** of a Kripke model $\mathcal{M} = \langle W, \{R_i\}_{i\in I}, V\rangle$ and an event model $\mathcal{E} = \langle E, \{Q_i\}_{i\in I}, \mathsf{pre}\rangle$ is a Kripke model $\mathcal{M} \oplus \mathcal{E} = \langle W', \{R'_i\}_{i\in I}, V'\rangle$ defined as follows:

[8] In this work, we restrict attention structures where the epistemic relations are equivalence relations. These are known in the literature as S5-structures or Aumann structures.

[9] To keep things manageable for this initial study, we restrict attention to event models with equivalence relations. For much of what follows, this assumption is not crucial.

- $W' = \{(w, e) \in W \times E \mid \mathcal{M}, w \models \mathsf{pre}(e)\}$
- $(w, e)R_i'(w', e')$ iff wR_iw' and eQ_ie'
- $(w, e) \in V'(p)$ iff $w \in V(p)$ ◁

This operation (together with variants appropriate for modeling belief and preference change) has been extensively studied in the literature. We do not provide an overview of this literature here: see [154, 160] for an extensive analysis. Rather, the focus is on how to understand this theory of information dynamics in the context of models of knowledge (and beliefs) typically found in the game theory literature. We need one additional notion from the general theory of modal logic.

Definition 2.5 (Bisimulation)**.** Suppose that $\mathcal{M}_1 = \langle W_1, R_1, V_1\rangle$ and $\mathcal{M}_2 = \langle W_2, R_2, V_2\rangle$ are Kripke structures. A nonempty relation $Z \subseteq W_1 \times W_2$ is a **bisimulation** provided for all $w_1 \in W_1$ and $w_2 \in W_2$, if w_1Zw_2 then:

(atomic harmony) For all $p \in \mathsf{At}$, $w_1 \in V_1(p)$ iff $w_2 \in V_2(p)$.

(zig) If $w_1R_1v_1$ then there is a $v_2 \in W_2$ such that $w_2R_2v_2$ and v_1Zv_2.

(zag) If $w_2R_2v_2$ then there is a $v_1 \in W_1$ such that $w_1R_1v_1$ and v_1Zv_2.

We write $\mathcal{M}_1, w_1 \leftrightarroweq \mathcal{M}_2, w_2$ if there is a bisimulation relating w_1 with w_2. We write $\mathcal{M}_1 \leftrightarroweq \mathcal{M}_2$ if there is a bitotal bisimulation between $\mathcal{M}_1$ and $\mathcal{M}_2$, that is a bisimulation Z such that for every $v \in \mathcal{M}_1$ there is some $W \in \mathcal{M}_2$ with vZw and vice versa. The relation Z is called a **simulation from $\mathcal{M}_1$ to $\mathcal{M}_2$**, denoted $\mathcal{M}_1, w_1 \rightrightarrows \mathcal{M}_2, w_2$, if Z satisfies the **atomic harmony** and **zig** properties. Z is called **total** provided for each $w_1 \in W_1$ there is a $w_2 \in W_2$ such that w_1Zw_2. Finally, Z is called **functional** if it is total and a function from W_1 to W_2 (i.e. for every $w_1 \in W_1$ and $w_2, \tilde{w}_2 \in W_2$ it is the case that w_1Zw_2 and $w_1Z\tilde{w}_2$ implies $w_2 = \tilde{w}_2$). ◁

2.2.2 Knowledge Structures

Knowledge structures were introduced in [52] as an alternative semantics for the basic epistemic language $\mathcal{L}_{EL}$.[10] They are non-probabilistic versions of Harsanyi type spaces which are the predominant model of knowledge and beliefs in the literature on the epistemic foundations of game theory ([29] offers some explanation about why this is the case).

[10] See [52] for an extended discussion of knowledge structures aimed at game theorists. Fagin [51] and Fagin and Vardi [55] show how variants of knowledge structures can provide an elegant semantics for many modal logics.

The key concept is a κ-**world** (also called a **type** in the game theory literature) describing the players' infinite hierarchy of knowledge (belief) of a given state of affairs.

Definition 2.6 (κ-world)**.** Let S be a (finite or infinite) nonempty set (whose elements are called states). A κ-**world** is a vector of functions $\mathbf{f} = \langle f_0, f_1, f_2 \ldots \rangle$ of length κ (a possibly infinite ordinal) defined inductively as follows:

- A **1-world** is a vector $\langle f_0 \rangle$ where f_0 is a state of nature (i.e., $f_0 \in S$).[11]
- For $\kappa > 1$ of the form $\kappa = \lambda + 1$ (i.e. κ is a successor ordinal) a κ-*world* is a vector $\langle f_0 \ldots f_\lambda \rangle$ such that $\langle f_i \mid i < \lambda \rangle$ is a λ-world and f_λ is a function from the set of agents I to the power set of the set of λ-worlds over S (i.e., $f_\lambda : I \to \wp(\mathcal{F}_\lambda(S))$, where $\mathcal{F}_\lambda(S)$ denotes the set of all λ-worlds over S) that satisfies the following conditions. Let $\mathbf{f}_{<\beta}$ denote the initial segment of $\mathbf{f}$ of length β.

  Extendability If $0 < \alpha < \lambda$, then $\mathbf{g} \in f_\alpha(i)$ iff there is some $\mathbf{h} \in f_\lambda(i)$ such that $\mathbf{g} = \mathbf{h}_{<\alpha}$ (i.e., higher-order worlds are extensions of lower-order worlds and every lower-order world has at least one higher-order extension).

  In addition, since we intend κ-worlds to represent the **knowledge** of the players, we impose two additional conditions:

  Correctness For each agent $i \in I$, $\mathbf{f}_{<\lambda} \in f_\lambda(i)$ (i.e., every agent must consider the actual state of the world possible).

  Introspection For all $i \in I$, if $\langle g_0, g_1, \ldots \rangle \in f_\kappa(i)$, then $g_\lambda(i) = f_\lambda(i)$, for all λ with $0 < \lambda < \kappa$ (i.e., players cannot consider states possible that differ in their description from their own lower-order beliefs). $\triangleleft$
- Finally, for κ a limit ordinal a κ-*world* is a vector of functions $\langle f_i \mid i < \kappa \rangle$ such that for every $\lambda < \kappa$ the vector $\langle f_i \mid i < \lambda \rangle$ is a λ-world.

We denote the set of all κ-worlds over S by $\mathcal{F}_\kappa(S)$.

The intended interpretation is that $f_\kappa(i) \subseteq \mathcal{F}_\kappa(S)$ is the set of all κ-worlds player i considers possible. Then, κ-worlds $\mathbf{f}$ are descriptions of the state of affairs and the players' higher-order knowledge (up to level κ). Thus, we can

[11] For the comparison with epistemic logic, it is useful to think of the set of states S as the set of propositional valuations on a set At of atomic propositions. In this case f_0 would be a propositional valuation function.

interpret the basic epistemic language at κ-worlds. For simplicity, we assume there is an atomic proposition E for every subset of the set of states S (i.e., $\mathsf{At} = \wp(S)$). This language is interpreted as follows:

$$\begin{array}{ll} \mathbf{f} \models E & \Leftrightarrow f_0 \in E \\ \mathbf{f} \models \neg\varphi & \Leftrightarrow \mathbf{f} \not\models \varphi \\ \mathbf{f} \models \varphi \wedge \psi & \Leftrightarrow \mathbf{f} \models \varphi \text{ and } \mathbf{f} \models \psi \\ \mathbf{f} \models K_i\varphi & \Leftrightarrow \text{ for each } \mathbf{g} \in f_l(i) : \mathbf{g} \models \phi \\ & \quad \text{where } l \text{ is the quantifier depth}^{12} \text{ of } \varphi. \end{array}$$

There is an alternative way of defining truth of the knowledge modality by defining an accessibility relation on $\mathcal{F}_\kappa(S)$, which transforms $\mathcal{F}_\kappa(S)$ into a Kripke model. We can then use the standard definition of a modal operator. For a κ-world $\mathbf{f} = \langle f_0, f_1, \ldots\rangle$, let $\mathbf{f}^i = \langle f_1(i), f_2(i), \ldots\rangle$ (note that the state of nature is not part of $\mathbf{f}^i$) and define a relation $\sim_i$ on the $\mathcal{F}_\kappa(S)$ as follows: $\mathbf{f} \sim_i \mathbf{g}$ iff $\mathbf{f}^i = \mathbf{g}^i$ (equality is defined component-wise). If $\mathbf{f} \sim_i \mathbf{g}$ then we say $\mathbf{f}$ and $\mathbf{g}$ are equivalent according to agent i. It is easy to see that these relations are equivalence relations. They turn $\mathcal{F}_\kappa(S)$ into a Kripke structure (with $\mathsf{At} = \wp(S)$ and the valuation function V defined by $w \in V(E)$ iff $w \in E$). Fagin *et al.* show ([54, Theorem 2.4]) that the interpretation of the epistemic language given above coincides with the interpretation of the epistemic language obtained by interpreting $\langle \mathcal{F}_\kappa(S), \{\sim_i\}_{i\in I}, V\rangle$ as a Kripke structure. So, there are two equivalent ways to interpret the basic epistemic language on the set $\mathcal{F}_\kappa(S)$ of κ-worlds. In the remainder of the chapter, we will use whichever definition is most convenient.

We are interested in general maps between Kripke structures and knowledge structures. To this end, we fix a set of atomic propositions At and assume that the state space S is the set of propositional valuations of At, i.e., $S = \wp(\mathsf{At})$. To simplify our exposition, we identify $p \in \mathsf{At}$ with $\{e \in S \mid p \in e\} \subseteq S$, i.e. the set of valuations containing p.

The key observation is that every Kripke structure can be naturally associated with a substructure of $\langle \mathcal{F}_\omega(S), \{\sim_i\}_{i\in I}, V\rangle$. The mapping is defined as follows:[13]

Definition 2.7 (Embedding from Kripke structures to knowledge structures)**.** Let $\mathcal{M} = \langle W, \{R_i\}_{i\in N}, V\rangle$ be a Kripke structure. We associate with each state $w \in W$ in $\mathcal{M}$ an ω-world $\mathbf{f}_{\mathcal{M},w} = \langle f_0^w, f_1^w, f_2^w, \ldots\rangle$ where the f_α^w are defined by synchronous induction on all worlds $w \in W$:

[12] Quantifier depth is defined as usual by induction on the structure of $\phi \in \mathcal{L}_{EL}$: Formally, $qd(p) = 0$, $qd(\neg\phi) = qd(\phi)$, $qd(\phi \wedge \psi) = \max(qd(\phi), qd(\psi))$, and $qd(K_i\phi) = 1 + qd(\phi)$.

[13] The mapping is a functional simulation but in general not a bisimulation onto its image. Nonetheless, it is a natural mapping in the sense that when applied to connected components $\mathcal{K}$ of $\langle \mathcal{F}_\omega(S), \{\sim_i\}_{i\in I}, V\rangle$ it is simply the embedding of $\mathcal{K}$ into $\langle \mathcal{F}_\omega(S), \{\sim_i\}_{i\in I}, V\rangle$

- $f_0^w = \{p \mid w \in V(p)\}$.
- To define the function f_{k+1}^w assume inductively that f_0^x, f_1^x, f_2^x, ..., f_k^x have been defined for all worlds $x \in W$ (k a natural number). Then, $f_{k+1}^w(i) = \{\langle f_0^x, f_1^x, \ldots f_k^x\rangle \mid wR_ix\}$.

Define the map $r : W \to \mathcal{F}_\omega(\wp(\mathsf{At}))$ as $r(w) = \mathbf{f}_{\mathcal{M},w}$. ◁

For every ordinal λ we can continue the construction to get a vector $\langle f_i^x \mid i < \lambda\rangle$. Thus this map naturally generalizes to maps $r_\lambda : W \to \mathcal{F}_\lambda(\wp(\mathsf{At}))$ for every ordinal λ. To simplify notation, assume for the rest of our analysis that $S = \wp(\mathsf{At})$ and that S is finite. The map r_κ gives a precise way to connect the class of all Kripke structures to a single structure $\mathcal{M}^\kappa = \langle \mathcal{F}_\kappa(S), \{\sim_i\}_{i\in I}, V\rangle$ for any κ. The following observation is immediate from the relevant definitions.

Observation 2.8. *Let $\mathcal{M} = \langle W, \{R_i\}_{i\in I}, V\rangle$ be a Kripke structure and $\mathcal{M}^\kappa$ be the structure $\langle \mathcal{F}_\kappa(S), \{\sim_i\}_{i\in I}, V\rangle$.*

i) The relation $wZ\mathbf{f}$ iff $r_\kappa(w) = \mathbf{f}$ is a functional simulation from $\mathcal{M}$ into $\mathcal{M}^\kappa$, but, in general, is not a bisimulation.

ii) There is an ordinal λ, depending on $\mathcal{M}$ such that Z is a bisimulation if $\kappa \geq \lambda$.[14]

iii In particular, if $\mathcal{M}$ is finite, then there is a bisimulation between $\mathcal{M}$ and $r(\mathcal{M}) = \langle r[W], \{\sim_i\}, V\rangle$. Moreover, $r(\mathcal{M})$ is the minimal bisimulation contraction of $\mathcal{M}$, i.e. the Kripke model of minimal cardinality that allows for a total bisimulation to $\mathcal{M}$.

Proof. *i*) The functionality of Z is obvious, since r_κ is a function. Atomic harmony holds by definition of f_0^w. To see that zig holds let $v_0, v_1 \in M$ with $v_0R_iv_1$ and $w \in \mathcal{M}^\kappa$ with v_0Zw. Since Z is functional we have $w = \mathbf{f}_{\mathcal{M},v_0}$ An induction shows that $f_i^{v_0} = f_i^{v_1}$ for every $i \leq \kappa$, thus $\mathbf{f}_{\mathcal{M},v_0}(i) = \mathbf{f}_{\mathcal{M},v_1}(i)$. Thus by definition of $\sim_i$ we have $\mathbf{f}_{\mathcal{M},v_0} \sim_i \mathbf{f}_{\mathcal{M},v_1}$. By definition of Z we also have $v_1Z\mathbf{f}_{\mathcal{M},v_1}$, thus zig holds. Example 3.10 of [52] shows that Z is in general not a bisimulation.
ii) Choose λ' such that for all $v, w \in \mathcal{M}$ holds: If there is some μ such that $r_\mu(v) \neq r_\mu(w)$, then $r_{\lambda'}(v) \neq r_{\lambda'}(w)$ and let $\lambda := \lambda' + \omega$. We have to show that zag holds: Let vZw with Z defined as above and let $w \sim_i w'$. We have to show

[14] In fact, for $\mathcal{M} = \mathcal{F}_\kappa(S)$ we have $\lambda = \kappa$. In particular, there are functional simulations between $\mathcal{M} = \mathcal{F}_\kappa(S)$ and $\mathcal{M} = \mathcal{F}_\lambda(S)$ for all $\kappa, \lambda > \omega$. Though $\mathcal{F}_\kappa(S)$ and $\mathcal{F}_\lambda(S)$ are not bisimilar for $\kappa \neq \lambda$.

that there is some $v' \in \mathcal{M}$ with $r_\lambda(v') = w'$. Indeed, since $w \sim_i w'$ we have for all $\mu < \lambda$ that $w' \restriction \mu \in w_\mu(i)$. By the construction of r_λ this implies that for every $\mu < \lambda$ there is some $v' \in \mathcal{M}$ such that $w' \restriction \mu = r_\mu(v')$. By the choice of λ' and the extendability condition, we have that $\exists \mu \in [\lambda'; \lambda) : r_\mu(v') \in w_\mu(i)$ implies $\forall \mu \in [\lambda'; \lambda) : r_\mu(v') \in w_\mu(i)$. In particular we have by the limit condition that $r_\lambda(v') = w'$ as desired. See chapter 3 of [52] for more details.
iii) Obvious from *ii*) and the definition of r_ω. □

2.3 Information Dynamics on Knowledge Structures

Our aim is to examine natural transitions between types in a knowledge structure. These transitions are intended to represent some type of reasoning process or information update about the state of nature of the beliefs of other players. For this initial study, we focus on the operation of product update (restricted to equivalence relations as in Definition 2.4).

2.3.1 Technical Preliminaries: Generalized Product Update

Our first contribution is to define a sequence of products $\times_{N_n}$ between *Kripke structures*. The idea to apply product update between Kripke structures (rather than Kripke structures and event models) was initially proposed by Jan van Eijck and colleagues [161]. We follow the same basic idea, although our approach differs in a technical, but crucial, way.

In order to generalize the product update operation so that it applies between two Kripke structures, we must replace the precondition function with something appropriate for merging two Kripke structures. Our approach is to explicitly mark which of the formulas we are interested in, and treat these formulas as atomic propositions.[15] Fix a set I of players and At of atomic propositions (for simplicity assume both are finite).

Definition 2.9 (Language extension)**.** 1. Let $\mathcal{T} \subseteq \mathcal{L}_{EL}$ with $\mathsf{At} \subseteq \mathcal{T}$. For every $\varphi \in \mathcal{T}$ we introduce a new constant $\check{\varphi}$ called the name of φ. Let $\check{\mathcal{T}} := \{\check{\varphi} | \varphi \in \mathcal{T}\}$. The **language extension** with $\mathcal{T}$, denoted by $\mathcal{L}_{EL}^{\mathcal{T}}$, is

[15] In general, this type of language extension can be used to model agents with limited memory. For instance, this is needed for an analysis of situations such as the sum and product riddle involving the dialogues: *A: I don't know* φ. *B: I* knew *you didn't know before you said that* (cf. [144] for an analysis of this puzzle in Public Announcement Logic).

the epistemic language with $\check{\mathcal{T}}$ as atomic propositions. By a slight abuse of notation we write p instead of $\check{p}$ for $p \in \mathsf{At} \subseteq \mathcal{T}$. We denote the valuation function over the language $\mathcal{L}_{EL}^{\mathcal{T}}$ by $V_{\mathcal{T}}$. As usual, we omit the subscript when it is clear from the context.

2. Let $\mathcal{M} = \langle W, \{R_i\}_{i\in I}, V\rangle$ be a Kripke model with atomic propositions At and let $\mathcal{T} \subseteq \mathcal{L}_{EL}$ with $\mathsf{At} \subseteq \mathcal{T}$. Then $\mathcal{M}$ can naturally be interpreted as a Kripke model over $\mathcal{L}_{EL}^{\mathcal{T}}$ by defining $V_{\mathcal{T}}$ as: $w \in V_{\mathcal{T}}(\check{\varphi})$ iff $\mathcal{M}, w \vDash \varphi$. We denote $\mathcal{M}$ viewed over $\mathcal{L}_{EL}^{\mathcal{T}}$ by $\mathcal{M}^{\mathcal{T}}$. ◁

In $\oplus$-updates every state v in the event model comes with a (generally complex) formula φ that is the precondition for v to occur. That is (w, v) is only defined if $\mathcal{M}, w \vDash \mathsf{pre}(v)$. This is exactly the idea of the $\times_{\mathcal{T}}$ update defined below: pairs of states are in the new model only if they agree on the formulas in $\mathcal{T}$.

Definition 2.10 (Product update)**.** i) Let $\mathcal{T} \subseteq \mathcal{L}_{EL}$ with $\mathsf{At} \subseteq \mathcal{T}$. Let $\mathcal{M} = \langle W, \{R_i\}_{i\in I}, V\rangle$ and $\mathcal{M}' = \langle W', \{R_i'\}_{i\in I}, V'\rangle$ be two Kripke models over $\mathcal{L}_{EL}^{\mathcal{T}}$. The *product model* $\mathcal{M}\times\mathcal{M}' = \langle W'', \{R_i''\}_{i\in I}, V''\rangle$ over $\mathcal{L}_{EL}^{\mathcal{T}}$ is defined as follows:

- $W'' = \{(w, w') \mid w \in W, w' \in W'$ and for all $\check{\varphi} \in \check{T} : w \in V_{\check{\mathcal{T}}}^{\mathcal{M}}(\check{\varphi})$ iff $w' \in V_{\check{\mathcal{T}}}^{\mathcal{M}'}(\check{\varphi})$;
- $(w, w')R_i''(v, v')$ iff wR_iv and $w'R_i'v'$; and
- $(w, w') \in V_{\check{\mathcal{T}}}''(\check{\varphi})$ iff $w \in V_{\check{\mathcal{T}}}^{\mathcal{M}}(\check{\varphi})$ (and thus also $w' \in V_{\check{\mathcal{T}}}'(\check{\varphi})$).

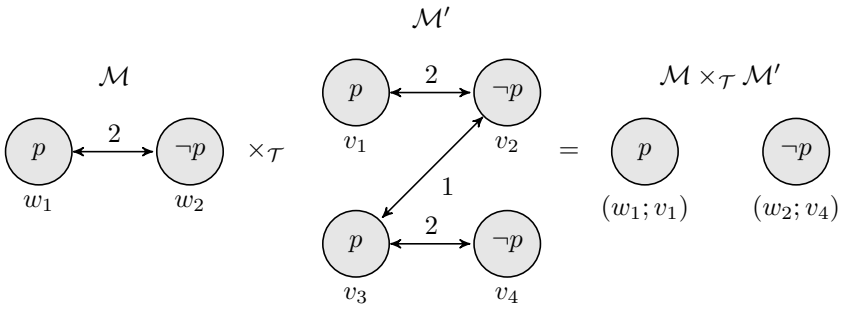
ii) The **generalized product update** of $\mathcal{M}$ and $\mathcal{M}'$ over $\mathcal{T}$, denoted by $\mathcal{M}\times_{\mathcal{T}}\mathcal{M}'$ is the model $\mathcal{M}\times\mathcal{M}'$ as defined above interpreted as a model over $\mathcal{L}_{EL}$. (That is: removing all atoms $\check{\varphi}$ with $\varphi \in \mathcal{T} \setminus \mathsf{At}$ and identifying $\check{p}$ with p for all $p \in \mathsf{At}$.) ◁

We write $\mathcal{M}\times_{\mathcal{T}}\mathcal{M}'$ where $\mathcal{M}$ and $\mathcal{M}'$ are Kripke models over $\mathcal{L}_{EL}$, meaning that we interpret $\mathcal{M}$ and $\mathcal{M}'$ as being models over $\check{\mathcal{T}}$ and do the $\times_T$-update as defined above. The procedure that we follow to compute this product runs as follows:

1. Pick a set $\mathcal{T}$ of statements to keep track of,
2. Build the Product in $\mathcal{L}_{EL}^{\mathcal{T}}$, and
3. Remove the additional information, i.e., restrict the valuation function from $\check{\mathcal{T}}$ to At.

The following example demonstrates this procedure.

Example 2.11: Let $\mathcal{T} = \{p, K_1p, K_2p, K_1\neg p, K_2\neg p\}$. Then the product of the two models is calculated as follows.



Note that the reflexive and transitive arrows are not drawn in the above picture for simplicity. The set $\mathcal{T}$ is rich enough to uniquely describe all knowledge assignments of level at most one. Thus, the product reflects a merging of models taking into account the agents' first-order information. The fragments of $\mathcal{T}$ true at the individual worlds are:

$$\mathcal{M}, w_1 \vDash \{p, K_1p\} \qquad \mathcal{M}, w_2 \vDash \{K_1\neg p\} \qquad \mathcal{M}', v_1 \vDash \{p, K_1p\}$$
$$\mathcal{M}', v_2 \vDash \emptyset \qquad \mathcal{M}', v_3 \vDash p \qquad \mathcal{M}', v_4 \vDash \{K_1\neg p\}$$

The only pairs satisfying the same fragment of $\mathcal{T}$ are (w_1, v_1) and (w_2, v_4). Observe that in the model $\mathcal{M} \times_{\mathcal{T}} \mathcal{M}'$ we have:

$$\mathcal{M} \times_{\mathcal{T}} \mathcal{M}', (w_1; v_1) \vDash \{p, K_1p, K_2p\}$$

which is different from the fragment of $\mathcal{T}$ satisfied by $\mathcal{M}, w_1$ and $\mathcal{M}', w_2$.

In general, taking a generalized product update consists of two steps: The first is picking a set of statements $\mathcal{T} \supseteq \mathsf{At}$ that one wants to keep track of and extending the language to $\mathcal{L}_{EL}^{\mathcal{T}}$. The second is to do generalized product update $\times_{\mathcal{T}}$, that is the normal product $\times$ over $\mathcal{L}_{EL}^{\mathcal{T}}$ followed by omitting all the information about the valuation of $\check{\mathcal{T}} \setminus \mathsf{At}$, i.e., making the newly created model an $\mathcal{L}_{EL}$ model again. The above example shows that the $\times_{\mathcal{T}}$ product does not preserve higher order information.

Remark 2.12: There are epistemic models $\mathcal{K}, w$ and $\mathcal{L}, v$ over $\mathcal{L}_{EL}$ a fragment $\mathcal{T}$ of $\mathcal{L}_{EL}$ and some $\varphi \in \mathcal{T} \setminus \mathsf{At}$ such that $(v, w) \in \mathcal{K} \times \mathcal{L}$ (the product over $\mathcal{L}_{EL}^{\mathcal{T}}$) and $\mathcal{K} \times \mathcal{L}, (v, w) \vDash \check{\varphi}$, but $\mathcal{K} \times \mathcal{L}, (v, w) \nvDash \varphi$. (Where, in the last formula, φ is evaluated as a formula of $\mathcal{L}_{EL}$.)

There is a close connection between generalized product update and the $\oplus$-update. In both cases, the result is not the complete cartesian product between the two state spaces, but a subset that is characterized by a certain set of formulas. The precise connection between the two concepts is clarified by the following lemma.

Lemma 2.13. *For every event model $\mathcal{E}$ there is some fragment $\mathcal{T} \subseteq \mathcal{L}_{EL}$ and a Kripke model $\mathcal{M}'$ (for the language $\mathcal{L}_{EL}^{\mathcal{T}}$) such that $\oplus\mathcal{E}$ is the same as $\times_{\mathcal{T}}\mathcal{M}'$ (i.e., for all Kripke models $\mathcal{M}$, $\mathcal{M} \oplus \mathcal{E}$ is isomorphic to $\mathcal{M} \times_{\mathcal{T}} \mathcal{M}'$).*

Proof. Let $\mathcal{E} = \langle E, \{Q_i\}_{i\in I}, \mathsf{pre}\rangle$ be an event model. Let $\mathcal{T}$ be the set $\{\mathsf{pre}(e) \mid e \in E\} \cup \mathsf{At}$. Construct the model $\mathcal{M}' = \langle W', \{R_i'\}, V'\rangle$ as follows: Let W' be the set of pairs (e, L_e) where $e \in E$ and $L_e \subseteq \mathcal{T}$ is a maximally consistent subset of $\mathcal{T}$ containing $\mathsf{pre}(e)$. The relations R_i' are defined as $(e, L_e)R_i(e', L_e')$ iff eQ_ie', and the valuation V' is defined by L_e (i.e., $(e, L_e) \in V(\check{\varphi})$ provided $\check{\varphi} \in L_e$). It is easy to check that this $\mathcal{M}'$ has the desired properties. □

Corollary 2.14. *If there is an upper bound for the quantifier depths of the preconditions in the event model $\mathcal{E}$ (i.e., the set $\{qd(\mathsf{pre}(e)) \mid e \in \mathcal{E}\}$ has an upper bound) then the set $\mathcal{T}$ in the above lemma can be chosen finite. This holds in particular if $\mathcal{E}$ is finite.*

Proof. Let n be an upper bound for the quantifier depths of $\{\mathsf{pre}(e) \mid e \in E\}$. Recall that $\mathcal{F}_n(\wp(\mathsf{At}))$ is finite, and so there are characteristic formulas ϕ_t for every $t \in \mathcal{F}_n(\wp(\mathsf{At}))$ (that is, $\mathcal{F}_n(\wp(\mathsf{At})), s \vDash \phi_t \Leftrightarrow s = t$). Let $\mathcal{T} := \{\phi_e \mid e \in \mathcal{F}_n(\wp(\mathsf{At}))\} \cup \mathsf{At}$ and construct a model $\mathcal{M}'$ as follows:

$$W' := \{(e,t) | e \in E, t \in \mathcal{F}_n(\wp(\mathsf{At})) \text{ and } \mathcal{F}_n(\wp(\mathsf{At})), t \models \mathsf{pre}(e)\},$$

let $(e,t)R_i'(e',s)$ if eQ_ie', and define V' as:

$$(e,t) \in V'(\check{\varphi}) \text{ iff } \mathcal{F}_n(\wp(\mathsf{At}), t \vDash \varphi$$

□

The sets $S = \{\phi_t \mid t \in \mathcal{F}_n(\wp(\mathsf{At}))\}$ chosen above are special in that these sets reflect all possible knowledge assignments up to depth n. We denote the resulting set of formulas by N_n (i.e., $N_n = \{\phi_t \mid t \in \mathcal{F}_n(\wp(\mathsf{At}))\} \cup \mathsf{At}$).

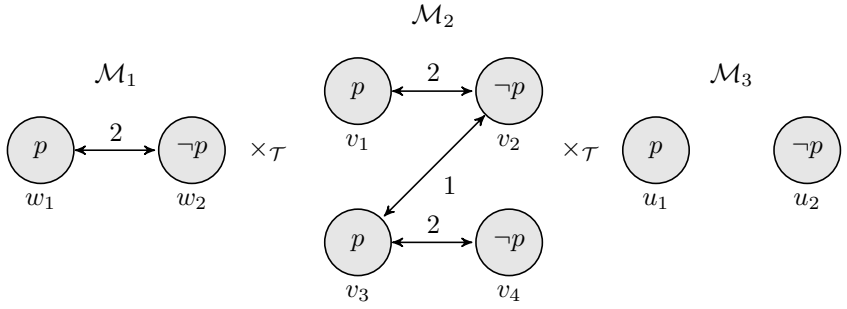
Remark 2.15:

i) In the above proof, we can turn $\mathcal{M}'$ into an event model $\mathcal{E}'$ by letting $\mathsf{pre}(e,t) = \varphi_t$. In this case we have $\mathcal{M} \times_{N_n} \mathcal{M}' = \mathcal{M} \oplus \mathcal{E}'$ for all $\mathcal{M}$. In particular $\mathcal{E}'$ is a special event model that only has preconditions from N_n. This follows a general pattern: The initial strength of arbitrary event models is that they allow for a very intuitive description of events in a multi-agent setting. However, from a technical point of view arbitrary event models can be difficult to handle. Therefore it sometimes proves useful to translate arbitrary event models into a certain subclass of event models which are easier to work with. For instance, [159] defined a class of canonical event models that are useful for studying when two event models are *equivalent*.

ii) The translation of an event model into a Kripke model blurs the distinction between static descriptions of situations and descriptions of events.

There is an interesting peculiarity of the $\times_{\mathcal{T}}$-products. Obviously, $\times_{\mathcal{T}}$ is commutative, but the following example shows that it is not associative.[16]

Example 2.16: This example is similar to Example 2.11. Suppose that $\mathcal{T} = \{p, K_1p, K_2p, K_1\neg p, K_2\neg p\}$. Consider the following $\mathcal{L}_{EL}$-models which we interpret as $\mathcal{L}_{EL}^{\mathcal{T}}$-models.



We now show that $(\mathcal{M}_1 \times_{\mathcal{T}} \mathcal{M}_2) \times_{\mathcal{T}} \mathcal{M}_3 \neq \mathcal{M}_1 \times_{\mathcal{T}} (\mathcal{M}_2 \times_{\mathcal{T}} \mathcal{M}_3)$. As we already noted in the previous example (Example 2.11), $\mathcal{M}_1 \times_{\mathcal{T}} \mathcal{M}_2 = \mathcal{M}_3$. In particular, $(\mathcal{M}_1 \times_{\mathcal{T}} \mathcal{M}_2) \times_{\mathcal{T}} \mathcal{M}_3 = \mathcal{M}_3 \times_{\mathcal{T}} \mathcal{M}_3 = \mathcal{M}_3$ where the last equivalence holds since u_1 and u_2 satisfy different formulas from $\mathcal{T}$.

[16] In general, it is clear that the process of consecutive learning is not commutative. One's actions in some event B can depend on having learned A before. In our formalization, the non-associativity captures this intuition: $(A \times_S B) \times_S C$ is to be read as being in situation A and learning B, then C, whereas $A \times_S (B \times_S C) = A \times_S (C \times_S B)$ corresponds to learning B and C at a time. A similar phenomenon has been noticed in the belief merging literature (cf. [115, Section 5.1]).

On the other hand, note that the following formulas from $\mathcal{T}$ are true at states in $\mathcal{M}_3$:

$$\mathcal{M}_3, u_1 \vDash \{p, K_1 p, K_2 p\} \qquad \mathcal{M}_3, u_2 \vDash \{K_1 \neg p, K_2 \neg p\}$$

However, there are no states in $\mathcal{M}_2$ satisfying precisely these formulas, so $\mathcal{M}_2 \times_{\mathcal{T}} \mathcal{M}_3 = \emptyset$ and consequently $\mathcal{M}_1 \times_{\mathcal{T}} (\mathcal{M}_2 \times_{\mathcal{T}} \mathcal{M}_3) = \emptyset$. Thus, we have $(\mathcal{M}_1 \times_{\mathcal{T}} \mathcal{M}_2) \times_{\mathcal{T}} \mathcal{M}_3 \neq \mathcal{M}_1 \times_{\mathcal{T}} (\mathcal{M}_2 \times_{\mathcal{T}} \mathcal{M}_3)$. [17]

The interpretation of this statement is that first learning $\mathcal{E}$ and then learning $\mathcal{E}'$ is different to learning $\mathcal{E}$ and $\mathcal{E}'$ *at the same time*. To be more precise, we have $(\mathcal{E} \times_{\mathcal{T}} \mathcal{F}) \times_{\mathcal{T}} \mathcal{G} \neq \mathcal{E} \times_{\mathcal{T}} (\mathcal{F} \times_{\mathcal{T}} \mathcal{G}) \neq \mathcal{E} \times_{\mathcal{T}} \mathcal{F} \times_{\mathcal{T}} \mathcal{G}$[18] This non-associativity shows that our framework is rich enough to distinguish between consecutive learning and receiving all information at once.

These observations should be contrasted with the theory developed in [161]. The authors of [161] are concerned with updates where all preconditions are boolean combinations of the ground variables (describing non-epistemic facts about the state of the world). Learning facts about the world is associative (cf. [161, Theorem 1]), whereas learning facts about the players' previous knowledge is not!

Van Eijck *et al.* [161] study the monoid generated by $\times_{\mathsf{At}}$ products. Our primary goal in this chapter is to understand how the $\oplus$-update works in type spaces. To that end, we first generalize a result from [158].

Theorem 2.17. *Let $\mathcal{M}_1$ be a Kripke structure such that for any $v, w \in \mathcal{M}$ there is an epistemic formula φ distinguishing v and w (i.e. $\mathcal{M}, w \vDash \varphi$ and $\mathcal{M}, v \vDash \neg\varphi$). Let $\mathcal{M}_2$ be an arbitrary Kripke structure. Then there is a set of formulas $\mathcal{T}$ and $\mathcal{L}_{EL}^{\mathcal{T}}$-Kripke structure $\mathcal{M}'$ such that $\mathcal{M}_1 \times_{\mathcal{T}} \mathcal{M}' \leftrightarrows \mathcal{M}_2$ if and only if there is a total simulation from $\mathcal{M}_2$ to $\mathcal{M}_1$. Furthermore, if the model $\mathcal{M}_1$ is finite the set $\mathcal{T}$ can be chosen finite.*

Proof. The direction from left to right is easy: Let $\mathcal{M}'$ and $\mathcal{T}$ be such that $\mathcal{M}_1 \times_{\mathcal{T}} \mathcal{M}' = \mathcal{M}_2$. It is easy to see that the map $\mathcal{M}_1 \times_{\mathcal{T}} \mathcal{M}' \to \mathcal{M}_1$ sending every pair (w, w') to w is a functional, hence total, simulation.

For the direction from right to left: Let Z be a total simulation from $\mathcal{M}_2$ to $\mathcal{M}_1$. First we define a Kripke model $\mathcal{M}^\circ = \langle W^{\mathcal{M}^\circ}, \{R_i^{\mathcal{M}^\circ}\}_{i \in \mathsf{Agt}}, V^{\mathcal{M}^\circ} \rangle$:

- $W^{\mathcal{M}^\circ} = \{(t_1, t_2) | t_i \in \mathcal{M}_i, i = 1, 2 \text{ and } t_1 Z t_2\}$

[17] There are examples where both $(\mathcal{M}_1 \times_{\mathcal{T}} \mathcal{M}_2) \times_{\mathcal{T}} \mathcal{M}_3$ and $\mathcal{M}_1 \times_{\mathcal{T}} (\mathcal{M}_2 \times_{\mathcal{T}} \mathcal{M}_3)$ are non-empty; however, they are more complicated while making the same point.

[18] Here $\mathcal{E} \times_{\mathcal{T}} \mathcal{F} \times_{\mathcal{T}} \mathcal{G}$ is the obvious generalization of $\times_T$ where all tuples (e, f) in the definition are replaced by triples (e, f, g).

- $(t_1, t_2) R_i^{\mathcal{M}^\circ} (s_1, s_2)$ iff $t_1 R_i^{\mathcal{M}_1} s_1$ and $t_2 R_i^{\mathcal{M}_2} s_2$
- $(t_1, t_2) \in V^{\mathcal{M}^\circ}(p)$ iff $t_2 \in V^{\mathcal{M}_2}(p)$ (and thus also $t_1 \in V^{\mathcal{M}_1}(p)$)

First we show that the model $\mathcal{M}^\circ$ is bisimilar to $\mathcal{M}_2$. We show that the projection map π_2 mapping every $(t_1, t_2) \in \mathcal{M}^\circ$ to $t_2 \in \mathcal{M}_2$ is a bitotal bisimulation (recall Definition 2.5). The atom condition is clear. For forth assume that $(t_1, t_2)\pi_2 t_2$ and that $(t_1, t_2) R_i^{\mathcal{M}^\circ} (s_1, s_2)$. By the definition of $R_i^{\mathcal{M}^\circ}$ we have $t_2 R_i^{\mathcal{M}_2} s_2$ and by definition of π_2 we have $(s_1, s_2)\pi_2 s_2$, thus forth is fulfilled.

Similarly, for back assume that $(t_1, t_2)\pi_2 t_2$ and that $t_2 R_i^{\mathcal{M}_2} s_2$. Since Z is a total simulation and $t_1 Z t_2$ holds by the construction of $\mathcal{M}^\circ$, there is some $s_1 \in \mathcal{M}_1$ with $s_1 Z s_2$ and $t_1 R_i^{\mathcal{M}_1} s_1$. But this means that $(s_1, s_2) \in \mathcal{M}^\circ$ and that $(t_1, t_2) R_i^{\mathcal{M}^\circ} (s_1, s_2)$, thus proving the back condition.

Since $\mathcal{M}_2 \leftrightarrows \mathcal{M}^\circ$, it suffices to show that there is some $\mathcal{M}'$ with $\mathcal{M}_1 \times_{\mathcal{T}} \mathcal{M}' = \mathcal{M}^\circ$.

Note, that the projection $\pi_1 : \mathcal{M}^\circ \to \mathcal{M}_1$ sending each pair (t_1, t_2) to t_1 is a functional left simulation. The atom condition is clear, and the rest can be shown with arguments similar to the ones given above.

Now, pick a set $\mathcal{T}^* \subseteq \mathcal{L}_{EL}$ that contains a distinguishing formula for any $v, w \in \mathcal{M}_1$ and let $\mathcal{T} := \mathcal{T}^* \cup \mathsf{At}$. Turn $\mathcal{M}^\circ$ into an $\mathcal{L}_{EL}^{\mathcal{T}}$-model $\mathcal{M}'$ by defining: $(t_1, t_2) \in V^{\mathcal{T}}(\check{\varphi})$ iff $\mathcal{M}_1, t_1 \vDash \varphi$. Since $\mathcal{T}^*$ is separating, $s_1 \in \mathcal{M}_1$ and $(t_1, t_2) \in \mathcal{M}^\circ$ satisfy the same $\check{\mathcal{T}}$-formulas iff $s_1 = t_1$. Therefore $\mathcal{M}_1 \times_{\mathcal{T}} \mathcal{M}' = \mathcal{M}^\circ$ as desired. Furthermore, if $\mathcal{M}_1$ is finite, then the set $\mathcal{T}^*$ can be chosen finite, thus proving the last statement. □

Remark 2.18: [158] contains a proof for a similar statement about $\oplus$-updates in the *finite* case. However, the generalization to infinite Kripke models does not hold for the $\oplus$-update.

Remark 2.19: Note that the model $\mathcal{M}'$ constructed in the right-to-left direction of the prove of Lemma 2.13 is in general *not* a $\mathcal{L}_{EL}$ model that is simply interpreted as an $\mathcal{L}_{EL}^{\mathcal{T}}$ model. That is: There is in general some $\varphi \in \mathcal{T}$ and some $w \in \mathcal{M}'$ such that $\mathcal{M}', w \vDash \check{\varphi}$ but $\mathcal{M}', w \nvDash \varphi$ (where $\check{\varphi}$ is an atom and φ is a formula evaluated in $\mathcal{M}'$ interpreted as a Kripke model over At (i.e. only containing atoms from $\{\check{p} \mid p \in \mathsf{At}\}$). In order to gain the expressive power of updating with an arbitrary event model, that is, one needs the class of *all* $\mathcal{L}_{EL}^{\mathcal{T}}$-models. Interestingly enough, this is no longer true when we restrict ourselves

to the class of *finite* Kripke structures. There, the full expressive power of the class of all $\oplus$-updates is already given by the class of all finite Kripke models over $\mathcal{L}_{EL}$ together with the set of all $\times_{N_n}$ products for $n \in \omega$. More formally, we have the following theorem (whose slightly tedious proof is relegated to the appendix).

Theorem 2.20. *Let $\mathcal{K} = \langle W, (R_i)_i, V\rangle$ and $\mathcal{L} = \langle W', R'_i, V'\rangle$ be finite Kripke structures such that $\mathcal{L}$ is obtainable from $\mathcal{K}$ by an update. Then there is some $\mathcal{T} \supseteq \mathsf{At}$ and some Kripke model $\mathcal{M}$ over the ground language $\mathcal{L}_{EL}$ such that $\mathcal{K} \times_{\mathcal{T}} \mathcal{M} = \mathcal{L}$.*

2.3.2 Characterization Result

As discussed in the previous section, every $\oplus$-update can be written as a $\times_{\mathcal{T}}$-update over a language in which the formulas in $\mathcal{T}$ are treated as atomic propositions. This will help us represent the product update in knowledge structures.

First, we need an equivalent to the extension of atomic propositions on types: For $n \in \mathbb{N}$ let S_n denote the set of all possible n-worlds, thus $S_n = \mathcal{F}_n(S)$ and $S_0 = S$). Technically, this is redundant, though it helps conceptually to distinguish $\mathcal{F}_n(S)$ as a type space generated by S and S_n which is the same type space reinterpreted as new set of atoms. By switching between those interpretations, every $n + k$ world over S can be seen as a k-world over S_n and thus there is a canonical embedding $\mathcal{F}_\omega(S) \to \mathcal{F}_\omega(S_n)$.[19]

For any two Kripke models $\mathcal{K}, v$ and $\mathcal{L}, w$ we have defined the product update $(\mathcal{K} \times \mathcal{L}, (v, w))$ over the unextended language $\mathcal{L}_{EL}$ above. Furthermore, we have seen that there is some κ such that r_κ is a bisimulation of $\mathcal{K}$ onto its image. Since $r_\kappa(v)$ is obviously in the image of r_κ this implies that parts of $\mathcal{K}$ are somehow coded in $r_\kappa(v)$. The idea of the following definition is that we can unravel enough information about $\mathcal{K}$ and $\mathcal{L}$ from $r_\kappa(v)$ and $r_\kappa(w)$ to determine $r_\kappa((v, w))$. We define a product $\times_0$ below and we will show later (lemma 2.23) that $r_\kappa((v, w)) = r_\kappa(v) \times_0 r_\kappa(w)$. As with the original definition of a κ-world (see 2.6), the definition is by induction.

Definition 2.21. Suppose that $n \in \mathbb{N}$ and $\mathbf{f}, \mathbf{g} \in \mathcal{F}_\omega(S)$. Then the $\times_0$-product $(\mathbf{f} \times_0 \mathbf{g}) \in \mathcal{F}_\omega(S) \cup \{\emptyset\}$ is defined as follows:

- $(\mathbf{f} \times_0 \mathbf{g})_0 = \langle f_0\rangle$ iff $f_0 = g_0$ and $\emptyset$ otherwise.
- $(\mathbf{f} \times_0 \mathbf{g})_m(i) = \{(\mathbf{f}' \times_0 \mathbf{g}')_{m-1} \mid \mathbf{f}' \in f_m(i), \mathbf{g}' \in g_m(i)\}$

[19] Note that this map is not surjective for $n \geq 1$: For instance the introspection conditions of $\mathcal{F}_{k+1}(S)$ gives some limitations on which elements of $\mathcal{F}_2(S_k)$ can come from $\mathcal{F}_{k+1}(S)$.

This definition can be lifted to an analogue of the generalized product update: The operator $\times_n$ will correspond to a product update with $\mathcal{T} = N_n$. First observe that the above definition of $\times_0$ works equally well if all S are replaced by S_n. As in the case of the generalized product update, the $\times_n$ update implicitly consists of two steps: First a product update between two elements of $\mathcal{F}_\kappa(S_n)$ followed by a removal of information, i.e. a projection from S_n to S. As with general product updates, the definition contracts these two steps into one:

Definition 2.22 ($\times_n$-Product)**.** Let $\bar{\pi} : S_n \to S$ be the projection map sending the tuple $\langle f_0 \dots f_{n-1} \rangle$ to f_0. Define $\times_n : \mathcal{F}_\omega(S_n) \times \mathcal{F}_\omega(S_n) \to \mathcal{F}_\omega(S)$ as follows:

- $(\mathbf{f} \times_n \mathbf{g})_0 = \langle s_0 \rangle$ iff $\bar{\pi}(f_0)) = \bar{\pi}(g_0) = s_0$, and $\emptyset$ otherwise.
- $(\mathbf{f} \times_n \mathbf{g})_m(i) = \{(\mathbf{f}' \times_n \mathbf{g}')_{m-1} \mid \mathbf{f}' \in f_m(i), \mathbf{g}' \in g_m(i)\}$. ◁

The following lemma describes the relationship between the $\times_{N_n}$-product and the $\times_n$-product. Basically, the $\times_{N_n}$ product of two Kripke models $(\mathcal{K}, w)$ and $(\mathcal{L}, v)$ carries the same information as the $\times_n$-product on the types $r(v)$ and $r(w)$.

For technical convenience we need a definition before we state the lemma: Recall that $N_n \setminus \mathsf{At}$ was chosen to be a set of characteristic formulas for $\mathcal{F}_n(S)$. Therefore, every state w in a Kripke structure $\mathcal{K}$ over $\mathcal{L}_{EL}$ satisfies exactly one formula of $N_n \setminus \mathsf{At}$. In particular for any Kripke model $\mathcal{L}$ over $\mathcal{L}_{EL}^{N_n}$ we have that $(v, w) \in \mathcal{K} \times_{N_n} \mathcal{L}$ implies that there is exactly one $\breve{\varphi} \in N_n \setminus \mathsf{At}$ with $w \in V(\breve{\varphi})$. We call Kripke models over $\mathcal{L}_{EL}^{N_n}$ satisfying this property **admissible**. Since every $e \in \mathcal{F}_n(S)$ satisfies exactly one formula from $N_n \setminus \mathsf{At}$ we have that every state of nature in an admissible Kripke model corresponds to exactly one $e \in \mathcal{F}_n(S)$ and we can define a map r' from admissible Kripke models to $\mathcal{F}_\omega(S_n)$ in the same way as we defined r.

Lemma 2.23. *Let $n \in \omega$ and let $\mathcal{K}, \mathcal{L}$, be Kripke models over $\mathcal{L}_{EL}^{N_n}$. Let $v \in \mathcal{K}$, $w \in \mathcal{L}$ satisfying the same N_n-formulas. Let $(v, w) \in \mathcal{K} \times_{N_n} \mathcal{L}$ denote the product of v and w in $\mathcal{K} \times_{N_n} \mathcal{L}$. Then we have $r((v, w)) = r'(v) \times_n r'(w)$, i.e., the following diagram commutes:*

$$\begin{array}{ccc} \mathcal{K}, \mathcal{L} & \xrightarrow{\times_{N_n}} & \mathcal{K} \times_{N_n} \mathcal{L} \\ \downarrow{\scriptstyle r', r'} & & \downarrow{\scriptstyle r} \\ \mathcal{F}_\omega(S_n), \mathcal{F}_\omega(S_n), & \xrightarrow{\times_n} & \mathcal{F}_\omega(S) \end{array}$$

Proof. Let $n \in \mathbb{N}$ and $v \in \mathcal{K}$, $w \in \mathcal{L}$ satisfying the same N_n-formulas. We inductively show that $(r'(v) \times_n r'(w))_k = r(v,w)_k$. For $k = 0$ this is trivial: If v and w satisfy the same atomic propositions over $\check{N}_n$ we have $(r'(v) \times_n r'(w))_0 = r((v,w))_0 = \{p \in \mathsf{At} : v \in V^{\mathcal{K}}(p)\}$. If they satisfy different atomic propositions we have $(v,w) \notin \mathcal{K} \times_{N_n} \mathcal{L}$ and $r'(v) \times_n r'(w) = \emptyset$. Now assume the statement holds for $k-1$ and let $i \in I$ (the set of agents).

First, we show $r(v,w)_k(i) \subseteq (r'(v) \times_n r'(w))_k(i)$. Let $x \in r((v,w))_k(i)$, thus x is a $k-1$-world. By construction of the map r there is some $\tilde{x}$ in $\mathcal{K} \times_{N_n} \mathcal{L}$ such that $\tilde{x} R_i (v,w)$ and $r(\tilde{x})_{k-1} = x$. Thus there are $x_1 \in \mathcal{K}$ and $x_2 \in \mathcal{L}$ such that the product of x_1 and x_2 in $\mathcal{K} \times_{N_n} \mathcal{L}$ is $\tilde{x}$ - in particular $x_1 R_i v$ and $x_2 R_i w$ and x_1 and x_2 satisfy the same N_n-formulas. In particular, $r'(x_1) \times_n r'(x_2) \neq \emptyset$ and by induction we have that $(r(x_1,x_2))_{k-1} = (r'(x_1) \times_n r'(x_2))_{k-1}$. On the other hand, we have $r'(x_1)_{k-1} \in r'(v)_k(i)$ and similarly for x_2 and w by the construction of r'. In particular, we have $x = (r'(x_1) \times_n r'(x_2))_{k-1} \in (r'(v) \times_n r'(w))_k(i)$ as desired, thus proving the first direction.

The argument for the reverse inclusion $r(v,w)_k(i) \supseteq (r'(v) \times_n r'(w))_k)(i)$ is similar: Let $x \in (r('v) \times_n r'(w))_k(i)$. Then there are $\tilde{x}_1 \in r'(v)$ and $\tilde{x}_2 \in r'(w)$ such that $(r'(\tilde{x}_1) \times_n r'(\tilde{x}_2))_{k-1} = x$ and such that there are $x_1 \in \mathcal{K}, x_2 \in \mathcal{L}$ such that $r'(x_i) = \tilde{x}_i$ and $x_1 R_i v$ and $x_2 R_i w$ hold. Since $\tilde{x}_1 \times_n \tilde{x}_2$ exists, x_1 and x_2 satisfy the same N_n-formulas. In particular there is some (x_1,x_2) in $\mathcal{K} \times_{N_n} \mathcal{L}$ with $(x_1,x_2) R_i (v,w)$. By construction of r we have $r((x_1,x_2))_{k-1} \in r((v,w))_k$ and by induction we have $r((x_1,x_2))_{k-1} = x$, thus proving the reverse direction. □

Note that the calculation of $\mathbf{f} \times_n \mathbf{g}$ from types $\mathbf{f}$ and $\mathbf{g}$ is computationally efficient: In order to calculate the k-th level of $\mathbf{f} \times_n \mathbf{g}$ only the first $n+k$ levels of $\mathbf{f}$ and $\mathbf{g}$ are required.

The above definition of $\times_n$ updates gives a way of modeling dynamics on a type space – thus opening up the field of epistemic game theory to belief dynamics. Event models were designed as a very intuitive and natural tool for representing epistemic events in a multi agent setting. The translation of event models into the corresponding pair of Kripke models and a product relation $\times_{N_n}$, and further into a type and a relation $\times_n$ allows us to calculate the change of epistemic status brought about by an event model $\mathcal{E}$.

On the other hand, every product update with a finite event model can be written as a $\times_n$-update, thus it suffices to understand the structure of $\times_n$ to study product updates. Thus, $\mathcal{F}_\omega(\wp(S))$ is not only a universal Kripke model

in the static sense, together with the products $\times_n$ is also universal in that it incorporates all potential updates.

On Kripke structures, translating event models into types allows us to study updating events as separate entities without any reference to a ground type. Furthermore, the translation blurs the distinction between types as static descriptions of epistemic states and knowledge changing events.

One natural and important question is: Given two types $\mathbf{f}$ and $\mathbf{g}$, is there a possible piece of incoming information that transforms $\mathbf{f}$ into $\mathbf{g}$? The intuition behind the answer given by the following theorem is: In the entire model, the agents are assumed to be omniscient and non-forgetting. Thus, an event cannot add any uncertainty about the state of nature, it can only remove some states from the sets of possible states. In contrast, for the higher order information, essentially anything is possible as long as it is compatible with individuals gaining new information about the state of nature. In particular, an epistemic event may increase the uncertainty about other agents' types. This idea is captured by the following definition.

Definition 2.24 (Admissibility of Types)**.** For a type $\mathbf{f} \in \mathcal{F}_\alpha(S)$ we say that a type $\mathbf{g}$ is **admissible** for $\mathbf{f}$ iff

- $f_0 = g_0$;
- for all agents i: $g_1(i) \subseteq f_1(i)$; and
- for $\alpha > 1$: If $\mathbf{h} \in g_\alpha(i)$ then there is some $\mathbf{h}' \in f_\alpha(i)$ such that h is admissible for $\mathbf{h}'$. ◁

Our characterization theorem is similar to Theorem 2.17.

Theorem 2.25. *Let $\mathbf{f}, \mathbf{g} \in \mathcal{F}_\alpha(S)$ be types such that $\mathbf{g}$ is obtainable by an update from $\mathbf{f}$, i.e. there is some n and some $\mathbf{h} \in \mathcal{F}_\alpha(S_n)$ such that $\mathbf{f} \times_n \mathbf{h} = \mathbf{g}$. Then $\mathbf{g}$ is admissible for $\mathbf{f}$. If the submodel of $\mathcal{F}_\omega(S)$ generated by $\mathbf{f}$ is finite also the converse holds true.*

Before we can prove this theorem, we recall the following result from infinite combinatorics.

Theorem 2.26. *(König's Lemma) Let T be an infinite, finitely branching tree. Then, T has an infinite branch.*

Proof. Construct an infinite branch $\langle x_0, x_1, \ldots \rangle$ as follows: x_0 is the root. For $i > 0$: If $x_0, \ldots x_i$ are already in the branch, pick a successor x_{i+1} of x_i that

has itself infinitely many successors (since the tree is finitely branching such a successor always exists). Then $\langle x_0, x_1, \ldots \rangle$ is an infinite branch. □

Proof of Theorem 2.25. The first statement is straightforward: Let $\mathcal{F}$ and $\mathcal{G}$ be the epistemic submodels of $\mathcal{F}_\omega(S)$ induced by $\mathbf{f}$ and $\mathbf{g}$, respectively. Assume that there is some $\mathbf{h} \in \mathcal{F}_\omega(S_n)$ such that $\mathbf{f} \times_n \mathbf{h} = \mathbf{g}$. By Lemma 2.23, this is equivalent to saying that $\mathcal{F} \times_{N_n} \mathcal{H} = \mathcal{G}$, where $\mathcal{F}, \mathcal{G}, \mathcal{H}$ are the generated Kripke models (over $\mathcal{L}_{EL}^{N_n}$) from $\mathbf{f}, \mathbf{g}$, and $\mathbf{h}$. By Theorem 2.17 there is a total simulation S from G to F. We inductively show that every $\mathbf{g}' \in G$ is admissible for every $\mathbf{f}' \in F$ with $\mathbf{f}' S \mathbf{g}'$. The 0th-level is clear by the definition of a simulation. Now it suffices to show that the definition of admissibility is fulfilled at the 1st level: Since we do this for all $\mathbf{g}' \in G$, the rest follows from the inductive definition of admissibility and the map r. To see that admissibility is fulfilled at the $1st$ level, let $\mathbf{h} \in G$ with $\mathbf{g}' \sim_i \mathbf{h}$. By definition, there is a $\mathbf{h}' \in \mathcal{F}$ with $\mathbf{f}' \sim_i \mathbf{h}'$. Thus, every state of nature that is conceivable for agent i in $\mathcal{G}$ via $\mathbf{h}$ is also conceivable in $\mathcal{F}$ via $\mathbf{h}'$ - this is exactly the definition of being admissible in the first level.

For the second statement let $\mathbf{g}$ be admissible for $\mathbf{f}$ and let the submodel of $\mathcal{F}_\omega(\wp(S))$ generated by $\mathbf{g}$ be finite. Again, let $\mathcal{F}$ and $\mathcal{G}$ be the Kripke submodels of $\mathcal{F}_\omega(S)$ induced by $\mathbf{f}$ and $\mathbf{g}$. Define the Relation Z between $\mathcal{F}$ and $\mathcal{G}$ as $\mathbf{f}' Z \mathbf{g}'$ iff $\mathbf{g}' \in \mathcal{G}$ is admissible for $\mathbf{f}' \in \mathcal{F}$. We will show that Z is a total simulation from $\mathcal{G}$ to $\mathcal{F}$, thus showing that $\mathcal{G}$ is obtainable by $\mathcal{F}$ via update (again using Theorem 2.17 and Lemma 2.23).

By assumption, $\mathbf{g}$ is admissible for $\mathbf{f}$. We show that whenever $\mathbf{g}' \in \mathcal{G}$ is admissible for $\mathbf{f}' \in \mathcal{F}$ and $\tilde{\mathbf{g}} \sim_i \mathbf{g}'$, then there is some $\tilde{\mathbf{f}} \sim_i \mathbf{f}'$ such that $\tilde{\mathbf{g}}$ is admissible for $\tilde{\mathbf{f}}$. This proves that Z is a left simulation. To see that Z is total, note that for every $\mathbf{g}'$ in $\mathcal{G}$ there is a chain $\mathbf{g} \sim_{i_1} \mathbf{g}_1 \sim_{i_2} \ldots \sim_{i_n} \mathbf{g}'$ connecting $\mathbf{g}$ with $\mathbf{g}'$. Let $\mathbf{g}' \in \mathcal{G}$ be admissible for $\mathbf{f}' \in \mathcal{F}$ and $\tilde{\mathbf{g}} \sim_i \mathbf{g}'$. We construct an ω-tree $(T, \prec)$ as follows: The k-th level consists of all those types in $f'_{k+1}(i)$ that enlarge $\tilde{g}_k$. The $\prec$-relation is defined as $\mathbf{r} \prec \mathbf{s}$ iff $\mathbf{r}$ is an initial segment of $\mathbf{s}$. By definition of the admissibility relation, every finite level of T is non-empty. Since the state of nature is considered finite, every nonempty level is also finite. Thus, by König's lemma T has an infinite path P. By construction, $\tilde{\mathbf{f}} = \bigcup_{\mathbf{r} \in P} \mathbf{r}$ is a type and $\tilde{\mathbf{g}}$ is admissible for $\tilde{\mathbf{f}}$. Since $\mathcal{F}$ is the substructure of $\mathcal{F}_\omega(S)$ induced by $\mathbf{f}$ (and thus by $\mathbf{f}'$) we have $\tilde{\mathbf{f}} \in \mathcal{F}$, thus the simulation Z relates $\tilde{\mathbf{g}}$ to $\tilde{\mathbf{f}}$. □

Again, there is an obvious counterpart of Remark 2.19 allowing us to update with $\mathcal{F}(S)$ worlds rather than $\mathcal{F}(S_n)$ worlds, provided all the induced Kripke structures involved are finite. To be precise, we can show the following: Let $\mathbf{f}, \mathbf{g} \in \mathcal{F}_\omega(S)$ be such that the epistemic submodels of $\mathcal{F}_\omega(S)$ induced by $\mathbf{f}$ and $\mathbf{g}$ are finite. Then $\mathbf{g}$ is admissible for $\mathbf{f}$ if and only if there is some natural number n and some $\mathbf{h} \in \mathcal{F}_\omega(S)$ such that $\mathbf{f} \times_n \mathbf{h} = \mathbf{g}$.

2.4 Conclusion and Future Work

Many different formal models have been used to describe the players knowledge and beliefs in game-theoretic situations. The variety of models reflect different mathematical conventions used by the various sub-communities, as well as competing intuitions about how best to describe the players' beliefs and reasoning in a game situation. It is important to understand the precise relationship between the alternative modeling paradigms. In this chapter, we focused on the two most prominent models found in the literature on the epistemic foundations of game theory: Kripke or Aumann structures and knowledge structures (non-probabilistic variants of Harsanyi type spaces).

There are two main contributions in this chapter. The first is to initiate a study of "information dynamics" for knowledge structures in the style of recent work on *dynamic epistemic logic* (cf. [154]). Such a theory would further illustrate the subtle relationship between type spaces and Kripke structures (updating the discussion initiated in [52, 54]). In particular, it allows us to combine the strengths of both approaches and use event models as a tool to describe epistemic events. The main technical contribution is the definition of a product operation $\times_n$ on the type space $\mathcal{F}_\omega(S)$. We provide a procedure that allows us to translate arbitrary event models into types. Furthermore, we show that the $\times_n$ product is powerful enough to simulate all updates by event models. Furthermore, we prove a characterization theorem (Theorem 2.25) showing when a type can be transformed into another type by updates with an event model.

This is only an initial study. We see our work here opening up many different avenues of future research. In particular, we plan on investigating the following issues in the future.

- What happens if we allow only updating types from a certain subclass of $\mathcal{F}_\alpha(S_n)$ (for example, finite epistemic models $\langle \mathcal{F}_\alpha(S_n), \{\sim_i\}_{i\in I}, V\rangle$)?
- What are the "behavioral" implications of our main characterization theorem (Theorem 2.25)? For example, if a strategy is rational for a type

f in a game G, does that strategy remain rational for all types that are admissible for **f**?

- How do we extend the ideas developed in this work to Harsanyi type spaces where the beliefs are represented by probability measures? The first step is to generalize the dynamic epistemic logic framework to settings where beliefs are represented by probabilities. Fortunately, this has largely been done (see [2, 155] for details). A very interesting direction for future research is to explore how to use the probabilistic event models and product update operation of [155] to prove a result analogous to our main characterization theorem (Theorem 2.25) for Harsanyi type spaces.

- The relation "obtainable by an update" together with our extended theorem (see Remark 2.19) turns the set of finite induced submodels of $\mathcal{F}_S(w)$ into an algebra. Can we characterize this algebra?

2.5 Appendix

Before we can prove Theorem 2.20, we need the following lemma showing that every n-world can be extended in at least $n+2$ different ways.

Lemma 2.27. *Let* $|\mathsf{At}| \geq 2$ *and* $|I| \geq 2$. *Let* $n \in \mathbb{N}$ *and let* $f \in \mathcal{F}_n(S)$ *be an* n*-world. Then* f *has at least* $n+2$ *different extensions to* $n+1$*-worlds.*

To prove both, the above lemma and the theorem, we construct a special Kripke Model $\mathcal{W}^f$, the *induced Kripke model* of f:

For the rest of the construction fix some n-world f. We inductively construct a finite set of points W together with partial partitions[20] P_i of W for every agent $i \in I$. In the construction every $w \in W$ is labeled with a k-world $l(w)$ for some $k \leq n$ To do this, we define the following operations:

For a point labeled with a k-world g, that is not contained in any P_i-partition cell yet, we define the **i-extension** of a point w to be the following:

- if $k = 0$ no new point is added and a single partition cell only containing w is added to P_i

- if $k > 0$ we add one new point for every $k-1$-world $h \in g_k(i)$ and label it with h. We add a new partition cell to P_i containing w and all newly added points.

[20] A *partial partition* of a set X is a set $K \subset Pow(X)$ such that $X_1 \neq X_2 \in K$ implies $X_1 \cap X_2 = \emptyset$. That is, a partial partition is a set that can be extended to a partition.

We then construct the set W^f as follows: We start with a set containing one point x labeled with f and all P_i empty. Inductively for all constructed points w and all agents i we do the following: If w is not contained in any partition cell of P_i yet, execute the i-extension of w. It is easy to see that the set W^f that is constructed by this procedure fulfills the following:

Fact 2.28.
- *W is finite*
- *P_i is a partition on W^f for every i (i.e. every $w \in W$ is contained in exactly one partition cell). Note that reading the partition sets as equivalence classes induces an equivalence relation on W^f. By a slight abuse of notation, we will denote both the partition and the induced equivalence relation by P_i*
- *The actual result does not depend upon the order of executing the i-updates*

Note that by definition $r_0(l(w)) \in S = \wp(\mathsf{At})$. Thus we can define a valuation V^f on W^f via: $w \in V(p) :\Leftrightarrow p \in r_0(l(w))$. Since all P_i are equivalence relations, W^f together with V^f defines a Kripke structure $\mathcal{W}^f = \langle W^f, (P_i)_{i\in I}, V^f\rangle$. We call $\mathcal{W}^f$ the **induced Kripke model of f.**

By induction on $k \leq n$ it is not too difficult to see that the following holds:

Fact 2.29. *For all $w \in W^f$ we have: $r_k(w) = r_k(l(w))$ whenever the latter is defined, i.e. $l(w)$ is a k'-world for some $k' \geq k$.*

Note that W^f has a pseudo-tree structure in the following sense: Let w be the initial point labeled with f and let $w' \in W^f$ be different from w. Then there are vectors $(i_1 \dots i_n)$ and $(v_1 \dots v_n)$ and a chain: $w = v_0 P_{i_1} v_1 P_{i_2} v_2 \dots P_{i_n} v_n = w'$. Furthermore, if we demand that $i_j \neq i_{j+1}$ for all j and $v_j \neq v_r$ for all $j \neq r$ then both vectors are unique - we denote them by $\mathbf{i}(w')$ and $\mathbf{v}(w')$. Furthermore, if f is an n-world and $l(w')$ is an n'-world then $i(w')$ has length $n - n'$.

Now we have all prerequisites to prove Lemma 2.27.

Proof of Lemma 2.27**.** We show the following: Let f be an n-world and let $v_0 \in W^f$ be such that $l(v_0)$ is a k-world for some $k \geq 1$. Then there are at least $k + 2$ extensions $\mathcal{W}_1, \dots \mathcal{W}_{k+2}$ of the Kripke-Model $\mathcal{W}^f$ such that

- each v' still satisfies $r_k(v) = r_k(l(v))$ whenever this is defined
- $r_{k+1}(v_0)(i)$ calculated in $\mathcal{W}_i$ is different from $r_{k+1}(v_0)(i)$ calculated in $\mathcal{W}_j$ for all $i \neq j$.

We will construct the $\mathcal{W}_i$ such that

- $v_0 \in \mathbf{v}(x)$ whenever xP_iy for some $x \in W^f$ and $y \in W_i/W^f$.
- For all $x \in W_i/W$ there is a chain $v_0K_{i_1}v_1K_{i_2}\dots K_{i_r}v_r = w$ of at most length $k+1$.

We prove the lemma by induction over k. To start, let $k = 1$. We have to construct three different Kripke models $\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3$ as claimed in the statement. Pick an agent j such that all elements of $P_j[v_0]$ other than v_0 itself are labeled with 0-worlds (and there is at least one such element). Note that such an j exists by the construction of $\mathcal{W}^f$.

Let $\mathcal{W}_1 = \mathcal{W}^f$. By construction, we have the following: All $v \in P_j[v_0]$ with $v \neq v_0$ are labeled with a 0-world. By construction of $\mathcal{W}^f$ it holds that for all these v and all agents $j' \neq j$ the singleton $\{v\}$ is a partition cell of $P_{j'}$ - thus j' knows the state of nature in v'. Thus in all but possibly one of the 1-worlds (f itself) that j considers possible at v_0 all other agents know the state of nature. Furthermore, j considers at least one such 1-world possible, since there is some $v \neq v_0$ in $P_j[v_0]$

Now we expand W^f to a new set of worlds W_2 in the following way: For every $v \in P_j[v_0]$ that is labeled with a 0-world we add another world v'.

We expand the partitions P_i to partitions of W_2 as follows: $\{v'\}$ is a new partition cell in P_j. For $i \neq j$ the partition cell $\{v\}$ of P_i is replaced by $\{v, v'\}$. Furthermore, for every new v' we pick a 0-world $l(v')$ such that $l(v') \neq l(v)$. The valuation induced by this labeling makes W_2 again a Kripke model $\mathcal{W}_2$. Furthermore this Kripke model satisfies the following:

- Still, we have $r_k(v) = r_k(l(v))$ whenever $l(v)$ is a k'-world for some $k' \leq k$
- $r_2(v_0)$ satisfies: For all 1-worlds that j considers possible (apart from possibly $l(v_0)$ itself) every agent $l \neq j$ considers exactly two different states of nature possible. In addition j considers at least one such 1-world possible.

In the same style, we construct a model $\mathcal{W}_3$ in the following way: Instead of adding only one new point w' for every $w \in P_j[v_0]/\{v_0\}$ we add two new points w' and w''. The new partition cells are adapted as above, i.e. $\{w', w''\}$ is a new cell in P_j and for all agents $i \neq j$ we replace the singleton $\{w\}$ in P_i by $\{w, w', w''\}$. Note that $|\mathsf{At}| \geq 2$, thus $|S| \geq 4$, since $S = \wp(\mathsf{At})$ In particular we can label the new points with 0-worlds such that $l(w), l(w')$ and $l(w'')$ are all mutually distinct. Again this gives a new Kripke-model $\mathcal{W}_3$ such that:

- We still have $r_k(v) = r_k(l(v))$ whenever $l(v)$ is a k'-world for some $k' \leq k$

- $r_2(v_0)$ satisfies: For all 1-worlds that j considers possible (apart from possibly $l(v_0)$ itself) every agent $l \neq j$ considers exactly three different states of nature possible. In addition j considers at least one such 1-world possible.

It is easy to see that the $\mathcal{W}_l, l \in \{1,2,3\}$ are as claimed.

Now assume inductively that we have already shown that for every v with $l(v)$ a k world there are $\mathcal{W}_1, \ldots, \mathcal{W}_{k+2}$ as claimed. We want to show the statement for $k+1$. Assume that v_0 is labeled with a $k+1$ world $l(w') = f'$. Fix some $j \in \mathsf{At}$ such that all $x \in P_j[v_0]$ other than v_0 itself are labeled with k-worlds. (Again, the construction of $\mathcal{W}^f$ ensures that such a j exists). By the correctness axiom, we have $r_k(l(v_0)) \in l(v_0)_{k+1}(j)$. Thus, by our construction there is some $v_1 \in P_j[v_0]$ labeled with $r_k(l(v_0))$. Furthermore, these two are the only points $x \in P_j[v_0]$ satisfying $r_k(l(x)) = r_k(l(v_0))$ Thus if we calculate $r_{k+2}(v_0)$ in $\mathcal{W}^f$ there are at most two different $k+1$-worlds $g_1, g_2 \in r_{k+2}(v_0)(j)$ with $r_k(g_i) = r_k(l(v_0))$ By induction, there are extensions $\mathcal{W}_1 \ldots \mathcal{W}_{k+2}$ of $\mathcal{W}^f$ such that each $\mathcal{W}_i$ comes along with a different extension of v_1 to a $k+1$-world. By our inductive construction, all these $\mathcal{W}_i$ have the same set $P_j[v_0]$.

In particular, in $\mathcal{W}_i$ we have that $r_{k+2}(v_0)(j)$ contains $l(v_0)$ and $r_{k+1}(v_1)$ as realized in $\mathcal{W}_i$ and these are the only k-worlds h in $r_{k+2}(v_0)(j)$ satisfying $r_{k+1}(g) = l(v_0)$. In particular, since all $\mathcal{W}_i$ induce different extension $r_{k+1}(v_1)$ of $l(v_1)$ we have that all $\mathcal{W}_i$ induce different extension $r_{k+2}(v_0)$. Furthermore, by induction all, these $\mathcal{W}_i$ fulfill the extra claims we made about them.

Thus, we need to find one more $\tilde{W}_{k+3}$ offering a different extension of $l(v_0)$ to an $n+2$-world. We first define the set of states W_{k+3}. By induction there are $\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3$ such that all $r_{k+1}(v_1)$ calculated in W_i are different. We can assume without loss of generality that W_1 and W_2 are such that $r_{k+1}(v_1)$ calculated in W_1 and W_2 are different from $l(v_0)$. Let $X = (W_2/W^f) \cup \{x \in W^f : v_1 \in \mathbf{v}(x)\}$. Note that X has induced partitions P_i coming from $\mathcal{W}_j$.

We define $\mathcal{W}_{k+1}$ as follows. The set of worlds of W_{k+1} is $W_1 \dot{\cup} X$. We denote the elements v_1 of W_{k+1} coming from $\mathcal{W}_1$ and X by v_1^1 and v_1^2. For the partitions note that

- For every agent $i \neq j$ and all $x \in X$ the $\mathcal{W}_2$ partition-cell $P_i[x]$ is completely contained in X
- For all $x \neq v_0$ in X the $\mathcal{W}_2$-partition cell $P_j[x]$ is completely contained in X.

We define the partition cells P_i as follows: For all $i \neq j$ the partitions P_j are

just the union of the partitions on $\mathcal{W}_1$ and on X. For P_j take the union of all partition cells of $\mathcal{W}_1$ and X not containing v_1^1 resp v_1^2. We add one further partition cell $P_j[v_1^1] \cup \{v_1^2\}$. (where $P_j[v_1^1]$ is the partition cell of v_1^1 in $\mathcal{W}_i$). It is not too difficult to see that the P_i thus defined are indeed total partitions.

By construction, $P_i[v_0]$ contains v_0, v_1^1 and v_1^2 thus $r_{k+2}(v_0)(j)$ has three different elements x_i satisfying $r_k(x_i) = r_k(l(v_0))$. In particular, the value of $r_{k+2}(v_0)(j)$ in $\mathcal{W}_{k+3}$ is different from the value in all other extensions. Furthermore, since $v_1 \in \mathbf{v}(x)$ implies $v_0 \in \mathbf{v}(x)$ the extra claim for the W_{k+3} is fulfilled.

Finally, for w labeled with our initial type f the $W_1 \dots W_{n+2}$ witness that there are $n+2$ different extensions of f to a $n+1$-world

□

Now we have all ingredients for the proof of Theorem 2.20

Proof of Theorem 2.20**.** Without loss of generality we can assume that $\mathcal{K}$ does not contain any bisimilar states. Furthermore, by Theorem 2.17 there is a functional left simulation S from $\mathcal{L}$ to $\mathcal{K}$.
Next, pick n_0 such that for $v \neq w \in W$ holds $r_{n_0}(v) \neq r_{n_0}(w)$. Such an n_0 exists since $\mathcal{K}$ is finite and has no bisimilar state. Let $n = 3 \cdot (n_0 + |W|^2 + |W'|)$. We will show that there is some epistemic model $\mathcal{M}$ such that $\mathcal{K} \times_{N_n} \mathcal{M} = \mathcal{L}$. We construct $\mathcal{M}$ as follows: By assumption, every $v \in W'$ has exactly one $w \in W$ with wSv. We define a labeling l on W' by $l(v') := r_n(w)$. In particular, we have $l(v_1) = l(v_2')$ iff $w_1 R_i w_2$ in $\mathcal{K}$ (Where $w_i \in W$ are such that $w_i S v_i$). Since S is a left simulation we have $l(x) = l(y)$ whenever $xR_i'y$ for $x, y \in W'$. Inductively, we construct a labeled set $W'' \supseteq W'$ as follows: For every i and every partition cell C of R_i intersecting the image of S we pick some $v \in c$ such that xSv for some $x \in W'$. For every $g \in l(v)$ we add a new point v_g to W''. To label this point, we pick some extension $\bar{g}$ of g to an n-world such that $\bar{g} \notin \{r_n(v) | v \in W\}$. (Note that this is possible by lemma 2.27 since $n > |W|^2$.) Furthermore, we construct these extensions such that for all $w, w' \in W'$ holds that if $\neg(wR_i'w')$ then also $\{l(v_g) | g \in l(w)_n(i)\} \neq \{l(v_g) | g \in l(w')_n(i)\}$. Again, this is possible by lemma 2.27 and the choice of n. For $x \in W$ and $i \in I$ let $P_{x,i} = R_i'[x] \cup \{v_g | g \in l(x)_n(i)\}$ and let $P_i = \{P_{x,i} | x \in W'\}$. Thus P_i is a partial partition of W''. As in construction of induced graphs, we construct a set $\tilde{W}$ by iteratively applying i-extensions to all labeled points not yet contained in a partition cell of P_i. (Note that all $x \in W'$ are in P_i-partition-cells for every i).

As above, we turn W'' into a Kripke model $\mathcal{M}'$ in the standard way. It is not difficult to see that in $\mathcal{M}'$ we have $r_s(v) = r_n(l_s(v))$ whenever $l(v)$ is a k-world for some $k \geq s$. Furthermore we have constructed $\mathcal{M}'$ such that $\{(x, y) \in \mathcal{K} \times_{N_n} \mathcal{M}' | y \in W'\} = \mathcal{L}$. Thus we only have to ensure that no pair (x, y) with $y \notin W'$ appears in $\mathcal{K} \times_{N_n} \mathcal{M}'$.

To ensure this, we enlarge $\mathcal{M}'$ to a model $\mathcal{M}$. Consider the set X of all points labeled with $n/3$-worlds. By our choice of n we can extend every such $n/3$ world to an $n/3 + 1$ world g with $g \notin \{r_{n/3+1}(w) | w \in W\}$. The proof of 2.27 shows that for every $v \in X$ there is an extension $\mathcal{M}(v)$ of $\mathcal{M}'$ such that

- $r_{n/3+1}(v) = g$ where g is as chosen above
- It still holds for all $v \in \mathcal{M}'$ and $i \in \mathbb{N}$ that $r_k(v) = r_k(l(v))$ whenever this is defined
- New points added in $\mathcal{M}(v)$ are only related to points that were created by iteratedly applying i-extensions to v
- For all new points x added there is a sequence $i_1, \ldots, i_k$ of length at most $n/3 + 2$ such that there are v_i with $v = v_0 P_{i_1} v_1 P_{i_2} v_2 \ldots P_{i_k} v_k = x$.

Applying the same technique to $\mathcal{M}(v)$ again for some $v' \neq v$ we get some $\mathcal{M}(v, v')$ with the same properties. Iterating this problem for all of X we get a model $\mathcal{M}(X) =: \mathcal{M}$ satisfying

- For all $v \in X$ holds: $r_{n/3+1}(v) =\notin \{r_{n/3+1}(w) | w \in W\}$
- till for every $v \in \mathcal{M}'$ and all $i \in \mathbb{N}$ holds $r_k(v) = r_k(l(v))$ whenever this is defined
- For all $y \in \mathcal{M}/\mathcal{M}'$ there is some $x \in X$ and a sequence $i_1, \ldots, i_k$ of length at most $n/3+2$ such that there are v_i with $x = v_0 P_{i_1} v_1 P_{i_2} v_2 \ldots P_{i_k} v_k = y$.

Since it holds for all $g \in X$ that $g \notin \{r_{n/3+1}(w) | w \in W\}$, we have for every k and every k-type h: If *some agent i considers in h possible a $k-1$ type where some agent j considers possible a $k-2$-type where... where some agent k considers the ($\frac{n}{3} + 1$) type g possible,* then $h \notin \{r_k(w) | w \in W\}$.

On the other hand, notice that for every $v \in \mathcal{M}'$ that is labeled with a k-world for some $k < n$ there is some vector $(i_1, \ldots, i_r)$ of length at most $2/3n - 1$ and some $(w_0 \ldots w_r)$ such there is a chain $v = w_0 P_{i_1} w_1 P_{i_2} w_2 \ldots p_{i_r} w_r = x$ for some $x \in X$. The same holds for all points in $\mathcal{M}/\mathcal{M}'$: Thus, by induction we have: For every world $x' \in \mathcal{M}$ holds: If x' is not labeled with an n-world,

then $r_n(x') \notin \{r_n(w)|w \in W\}$. By our initial choice the same holds for all $w \in W'/W''$. In particular

$$\{w \in \mathcal{M}|r_n(w) \in \{r_n(w)|w \in W\}\} = W'$$

Thus we have for all $(x, y) \in \mathcal{K} \times_{N_n} \mathcal{M}$ that $y \in W'$, therefore $\mathcal{K} \times_{N_n} \mathcal{M} = \mathcal{L}$ as desired. □

Chapter 3

Levels of Knowledge and Belief

Information is a central driving force for social interaction. The epistemic and doxastic states of the various parties not only decide whether some social interaction happens at all, but also how it pans out. To sharpen our intuitions, we start with two examples.

Example 3.1: Our first example is about a pedestrian trying to cross a road. As it so happens, a car is heading down that same road at the very same time. These two are bound to collide unless one of them is aware of the situation and acts accordingly. Thus, we would assume both parties to collect as much information about the situation as possible or, to say it in other words, to pay attention to the traffic. However, in his work on social software [127], Rohit Parikh reports a seemingly paradoxical observation. In many countries, pedestrians ostentatiously look *away* before crossing a street, rather than concentrating on the approaching car. Parikh then continues to explain this behavior as strategically rational. While both parties, the pedestrian and the motorist, prefer not to crash, an accident would have far worse consequences for the former. The pedestrian will thus be more likely to back off from a conflict situation and, knowing this, the driver can speed up, leaving it to the pedestrian to avoid the accident. Obviously, if the traffic was dense enough and if all drivers adopted this reasoning, the pedestrian would be stuck on her side on the road forever. One possible escape route from this strategic disadvantage for the pedestrian is to undermine the driver's reasoning. If the driver can't be sure that the pedestrian is aware of him, he can no longer safely assume that the latter would

This chapter is based on joint work with E. Pacuit. We thank J. van der Meeren for valuable discussions.

refrain from crossing. Thus, if the pedestrian is sufficiently positive that the driver is paying attention, it is a dominant strategy for her to signal *not* to be paying attention, leaving it to the driver to prevent the accident.

Example 3.2: For our second example, consider a group of mountaineers having gotten lost in a snow storm while it is already turning dark. After analyzing the situation thoroughly, the group identifies two major options. First, they could hold out, spend the night and hope for rescue in the morning or, second, they could attempt to climb up a risky route and try to arrive at a safe place where help could reach them easily. In either case, the mountaineers decide to stay together as a group, so they jointly deliberate on what to do. While discussing different options, the group finds it helpful to know how many of them are supporting either option. However, since the stakes are high and the situation is stressful, the group fears the risk of some members blindly following others, or saying whatever seems to avoid discussion, rather than reporting their true beliefs. In order to minimize this risk and incentivize truthful reports, the group decides that it should become common knowledge *how many* members opted for the different options, but not *who* did so.[1,2]

The driving factor in both these examples is the higher order *information* available to the different agents. For a successful social interaction, it can be important to have *the right amount* of first and higher order information among the parties involved. Obviously, a lack of relevant information can lead to undesirable results, a car accident or an ill-chosen decision among the mountain climbers. But also *too much information* can lead to unwanted results, strategizing motorists or climbers reporting their preferences untruthfully.

In this chapter, we will inquire into the various informational states that can occur in social interaction, and in how to represent these. The two central questions thus are *what* information we are interested in precisely and *how* this information should be represented. To start, we give a first, tentative, answer to the "how". We use this chapter to explore different logical frameworks for representing informational states, mainly stemming from epistemic and doxastic logic. Our main goal here is to compare these different frameworks with respect to two criteria, their ability to adequately represent the relevant aspects of some informational setting, and the difficulty of realizing or bringing about some informational distribution formulated within some framework. We will

[1] We silently ignore those cases where all or all but one group member share the same opinion. In these cases the goal is obviously not or only partially realizable.

[2] How to do so without pens, paper or a ballot box is a central topics in the theory of secure communication. See [140, chapter 6] for a solution.

explain both of these criteria later in greater detail.

Regarding the "what", we will focus our inquiry on the informational attitudes towards a *single proposition* φ. We should emphasize here that we do not restrict the content of this proposition φ. In general, φ could be of arbitrary length, possibly conjoining a multitude of simpler statements about the matter of fact and the beliefs and intentions of the various agents. We hold that focusing on a single proposition only does not pose a major restriction in studying social situations. For a broad range of situations, the relevant informational setting can be described by the attitudes towards a potentially complex proposition φ. For instance, in both of the above examples all the relevant information consists in different epistemic attitudes towards the propositions *the car and the pedestrian are heading towards a collision* and *we should camp over night* respectively. In the road crossing case, the relevant epistemic attitudes will be of first order, the motorist and the pedestrian both knowing that φ, but also of second order, the motorist not knowing that the pedestrian knows that φ. The second example even needs to refer to third order epistemic states, each individual climber knowing that no other group member will know whether that agent believes φ.

This leads us straight to a second, more fine-grained characterization of what we want to represent. In discussing informational attitudes, we will not be interested in a *complete* description of all properties that hold true about φ. Just on the contrary, we will solely focus on those informational attitudes about φ that are *relevant* for the situation in question and the behavior of the different agents. For instance, in the traffic example the relevant propositions will be *"The driver knows that φ"*, *"The pedestrian knows that φ"* and so on, while propositions such as *"The nearby shop owner believes that φ"* are irrelevant for the further development of that situation. We will call the set of all expressions relevant for the situation, i.e., of formulas whose truth or falsity might influence the panning out of the situation, the *reasoning language*. Of course, different situations will require different reasoning languages. But more than that is true. Different situations may even require different *types* of reasoning languages. For instance, all the relevant properties of the first example are of the form *A knows that B knows that φ*, requiring only knowledge modalities to express them. In contrast, the second example additionally requires disjunctive formulas of the forms *A knows that either B_1 or B_2 believes that φ*. These examples are indicative of the fact that various *types* of situations require different *types* of reasoning languages. To give two more examples, all formulas appearing in the analysis of

bounded reasoning [89] will have a bounded quantifier depth, whereas the logical analysis of distributed computing [16, 38, 71, 129] is primarily interested in *positive* knowledge, formulas that only involve φ and the knowledge modalities.

In this chapter, we will inquire into some salient choices of reasoning languages and their respective properties. We mainly focus on two properties of the different reasoning languages. The first of these, *expressive power*, is quite intuitive. The expressive power of a framework or a reasoning language describes its ability to distinguish situations and express the bits and pieces of information relevant for some social situation. Our second criterion of interest is *realizability.* To illustrate this criterion, assume that we use some reasoning language to describe a given social situation such as the case of pedestrians crossing a road. Assume further, that we have identified some informational setting, a distribution of first and higher order information about φ, that would foster a desirable outcome of that situation. We will call such a particular informational setting a *level of information.* A natural question to ask then is whether such a level of information, a favorable informational setting, could realistically occur and, if so, *how* we could bring it about, that is how we could *realize* it. Of course, this definition depends upon how we flesh out the "realistically". We will come back to this issue later. Relating this discussion back to reasoning languages, we will be interested in whether and how levels of information given in a particular reasoning language are realizable. To be a bit more precise, our second criterion is to ask which reasoning languages *guarantee* that every consistent level of information expressable in that language could arise in a realistic scenario.

As is easy to see, these two criteria, expressive power and realizability, draw in different directions. The more fine grained distinctions a language allows to make, that is the more expressive it is, the more difficult it is to realize a given level formulated in that reasoning language. In the most extreme case, the full multi-agent epistemic language allows for a highly detailed description of any situation. However, the price to pay is that most such descriptions, even if consistent, cannot be realized in a sufficiently small model. On the other extreme, a highly impoverished reasoning language, only allowing to express whether some agent i believes φ or not, makes it extremely easy to realize any given level of information: Since such a level can only express whether i believes φ or not, we just need to convince that agent of φ or its opposite. However, this language is, of course, too poor to discuss most social situations adequately. In general, we will have to decide case by case which reasoning language to pick, depending upon the situation we want to model and its relevant features.

Before proceeding, we should clarify how our criteria, expressive power and realizability, relate to standard discussions of models. To begin with, we should emphasize that we do *not* apply these criteria to particular, concrete models of some situation. Rather, they serve to discuss reasoning languages or logics, formal frameworks for composing and describing models of concrete situations.[3] Notably, the reasoning languages themselves do not need to be models of anything, they simply are formal languages with some internal structure. In a sense, the choice between different reasoning languages or, more generally, between different logics, can be seen as a choice between general modeling paradigms for any concrete situation to come. As we will elaborate later, in chapter 7, this choice is guided by a variety of factors, two of them being expressive power and realizability. The first of these two, expressive power, is one of the most widely used criteria in comparing different logical frameworks for a given target domain, see for instance the discussion in [130]. The expressive power of a reasoning language is loosely related to the precision of its models. The more expressive a logical language, the more precisely it can describe how the target system is or is not. The second criterion, realizability, roughly is a measure of internal coherence of the corresponding reasoning language. Intuitively, some (consistent) set of formulas is easily realizable if all formulas in that set cohere well, if they are interrelated smoothly.

Now, it is time to be a bit more precise about what we mean with a *realistic* situation in discussing realizability. For this paper, we will adapt a rather weak criterion about which situations could possibly arise in realistic settings. Arguably, a necessary condition for an informational setting to be realizable in a real-world situation is that it must rest on a *finite* amount of options or states the different agents take into account. Thus, we are interested in which levels of information could arise within such a finitely represented situation. Going back to reasoning languages, this leads us to the following criterion: We want to distinguish reasoning languages that *guarantee* that every consistent level of information expressible in these languages is representable in a finite setting from those that do not. In the first part of this paper, we will focus on a closely related question. We will distinguish languages that allow for only few, countably many, different consistent levels from those that allow for uncountably many.[4]

[3] A model of a concrete situation will then correspond to a subset of a reasoning language describing what is and is not the case in that situation. This is exactly what levels of information are supposed to do.

[4] To clarify the connection: There are only countably many finite models of some given, finitely generated language, thus only countably many different levels of information could, in principle, be represented within a finite model.

The main contribution of this chapter is to unify and extend two previous results about the cardinality of levels of information in the literature. The first result by Parikh and Krasucki [129, Theorem 3] shows that there are only countably many *levels* of knowledge. A related result by Hart *et al.* [74, Theorem 2.2.] proves that there are, in fact, uncountably many *states* of knowledge. Of course, these results are not contradictory as the two papers are, in fact, counting different sets. The crucial difference is that Parikh and Krasucki are interested in *knowledge that*, that is their reasoning language consists of all statements of the form *A knows that B knows that* φ, whereas Hart *et al.* analyze the case of *knowledge whether*, that is statements of the form *A knows whether B knows whether* φ. Building on these results, we aim at identifying which fragments of the standard epistemic language *guarantee* that there are only countably many different levels of information and which parts cause an explosion to uncountably many levels. As it will turn out, the dividing line between countably and uncountably many different state descriptions can be sharp and unexpected. We will show that the difference can be as subtle as adding an additional connective to the reasoning language or even simply adding a further agent to the situation. The rest of this chapter is structured as follows: We start by introducing our formal definitions and motivating the problem in section 3.1. In the next section 3.2 we then provide several results about which languages allow for countably many levels of information and which do not. In the third section, we come back to the question of realizability, identifying which levels of information identified in the previous section are realizable within finite models under various conditions. In section 3.4 we then conclude and outline some directions for future research. Finally, the appendix contains some proofs and constructions.

3.1 The Framework

Our analysis will be conducted in the framework of multi-agent epistemic logic. To start, we give a quite general definition of the reasoning language we are going to employ. Since we restrict ourselves to the informational attitudes towards a single proposition, our language will have a single propositional variable x.[5] For some given index set I, let $\mathcal{L}$ be the logical language generated by the following Backus-Naur normal form:

[5] In the following, the proposition of interest φ will be represented by the atomic formula x. Of course, φ could, in principle, be itself a complex proposition such as $K_i x$ or $x \wedge \neg x$. We hold that the case of an atomic proposition is the most fundamental case and all other cases follow from there.

$$\varphi ::= x|\varphi \wedge \varphi|\neg\varphi|\Box_i\varphi \text{ for } i \in I$$

where the derived operators such as $\rightarrow, \vee$ or $\Diamond_i$ are defined in the usual way. Of course, I is intended to be a set of agents, in which case the modal operators will be knowledge or belief modalities, denoted by K_i and B_i respectively. However, much of the subsequent analysis applies to more general modal logics and we wish to maintain the flexibility to accommodate a broader class of languages. For instance, knowledge-belief logics equip each agent with two modal operators, one representing her knowledge the other one representing the belief states. In the following, we will write K_i whenever we mean a knowledge operator and B_i for the respective belief operators. We will use the standard axiomatizations for these, that is

normality	$K_i(\varphi \rightarrow \psi) \rightarrow (K_i\varphi \rightarrow K_i\psi)$
factivity	$K_i\varphi \rightarrow \varphi$
positive introspection	$K_i\varphi \rightarrow K_iK_i\varphi$
negative introspection	$\neg K_i\varphi \rightarrow K_i\neg K_i\varphi$

for the knowledge operator and normality plus the last two, positive and negative introspection, for the belief operator. As usual, we will refer to these axiomatizations as S5 and KD45 respectively. We will further assume that there are no additional interaction rules between the operators for different agents. As we have motivated above, we won't be interested in the entire logical language, but only in some selected subset of formulas relevant for our purpose. To define that formally, we have

Definition 3.3 (reasoning language)**.** A **reasoning language** is any fragment $\mathcal{L}_{reas} \subseteq \mathcal{L}$. The reasoning language **generated** by a set of operators $\mathcal{O} = \{O_1 \dots O_k\}$ definable in $\mathcal{L}$, where each $\mathcal{O}_i$ has arity n_i, is defined by the Backus-Naur normal form:

$$\varphi ::= x|O_1(\underbrace{\varphi \dots \varphi}_{n_1})|O_2(\underbrace{\varphi \dots \varphi}_{n_2})| \dots |O_k(\underbrace{\varphi \dots \varphi}_{n_k})$$

◁

To give an example, the reasoning language of positive knowledge $\mathcal{L}_K$, generated by the set $\{K_1, \dots, K_n\}$ is the set of all formulas of the form $K_{i_1} \dots K_{i_r}x$ with all $i_j \in \{1, \dots, n\}$, that is, formulas that only use x and any number of knowledge operators. Next, we can define our primary target of research, levels of information of some given reasoning language $\mathcal{L}_{\mathcal{O}}$.

Definition 3.4 (Level of Information)**.** A **level of $\mathcal{L}_{\mathcal{O}}$-information** or, short, a level of information, is a set $T \subseteq \mathcal{L}_{\mathcal{O}}$ such that the set

$$T \cup \{\neg\varphi | \varphi \in \mathcal{L}_{\mathcal{O}} \setminus T\}$$

is consistent. We will denote the set of all levels of information for $\mathcal{L}_{\mathcal{O}}$ by $\mathcal{T}_{\mathcal{L}_{\mathcal{O}}}$. ◁

With other words, a level of information is a *complete* list of all and exactly those formulas of $\mathcal{L}_{\mathcal{O}}$ that are or should be made *true* in a given situation. While this definition is fully syntactic, a part of our original question was semantic, asking about the *realizability* of a given level of information within a model. Of course, these definitions are equivalent, as witnessed by the several completeness theorems of modal logic (see for instance [124]). Thus, an equivalent definition of a level of information would be:[6]

Alternative Definition (Level of Information)**.** A **level of $\mathcal{O}$-information** or, short, a level of information, is a set $T \subseteq \mathcal{L}_{\mathcal{O}}$ such that for some Kripke or Neighborhood Model $\mathcal{M}$ and some $w \in \mathcal{M}$ we have

$$T = \{\varphi \in \mathcal{L}_{\mathcal{O}} | (\mathcal{M}, w) \vDash \varphi\}$$

In this case, we say that the pointed model $(\mathcal{M}, w)$ **realizes** the level T. ◁

Since we are especially interested in the cases of knowledge and belief, we briefly recall two particular completeness results that we will need later in this chapter. The logic S5, and thus also multi-agent S5 is sound and strongly complete with respect to the class of Kripke frames where all accessibility relations are equivalence relations ([22, Theorem 4.29]), while KD45 with its multi-agent variants is the logic of Kripke frames where all accessibility relations are transitive, serial and Euclidean.

So let us come back to our main question, how many different levels of knowledge there are for a given reasoning language $\mathcal{L}_{\mathcal{O}}$. If the reasoning language is infinite,[7] there are uncountably many subsets $S \subseteq \mathcal{L}_{\mathcal{O}}$, so why isn't it trivial that there are uncountably many levels of information? The answer lies, of course, in the consistency conditions. Not every subset $S \subseteq \mathcal{L}_{\mathcal{O}}$ is consistent and can thus be (part of) a level of information. The following is an example of

[6] For a definition of Kripke structures see chapter 2, Definition 2.2. Neighborhood structures are a generalization of Kripke structures where the accessibility relations are replaced by a map $N : W \to \mathcal{P}(\mathcal{P}(W))$. See [124] for details

[7] If $\mathcal{L}_{\mathcal{O}}$ is finite, there are only finitely many levels of information since $\mathcal{T}_{\mathcal{L}_{\mathcal{O}}} \subseteq \mathcal{P}(\mathcal{L}_{\mathcal{O}})$ and the latter is finite.

an inconsistent subset of the language with two knowledge operators:[8]

$$\{x, K_1x, \neg K_2K_1x, \neg K_1\neg K_2K_1x, K_2\neg K_1\neg K_2K_1x\}$$

A second reason why it is not obvious that there are uncountably many levels of information of some proposition ϕ is that two sets of formulas X and Y may, while consistent, correspond to the same level of information. That is, we might have that $X \subseteq T \Leftrightarrow Y \subseteq T$ for every level of information $T \in \mathcal{T}_{\mathcal{L}_\mathcal{O}}$. The following is an example of two subsets representing the same information:

$$X = \{x, K_1x, K_3x, K_1K_2K_3x\} \quad \text{and} \quad Y = \{x, K_1x, K_2x, K_3x, K_1K_2K_3x\}$$

The reason for why X and Y represent the same information leads us straight to the countability result by Parikh and Krasucki that the language $\mathcal{L}_K$ generated by $\{K_1 \dots K_n\}$ has only countably many levels of information. In line with their use of notation, we will refer to levels of $\mathcal{L}_K$ information as levels of *knowledge*. To begin with, let us study the case of a single agent. As can be easily seen, there are three different possible levels of knowledge towards a single proposition x. The proposition x could be false, it could be true but our only agent doesn't know this or it could be true and the agent knows about this. By positive introspection, the agent in the latter case also knows that she knows, knows that she knows that she knows and so on. Thus, the three possible levels of knowledge are:[9]

$$T_1 = \{\} \qquad T_2 = \{x\} \qquad T_3 = \{x, Kx, KKx \dots\}$$

As a first step towards a multi-agent account of knowledge, note that the factivity axiom $K_i\varphi \to \varphi$ for agent i, inserted into the left side of the normality axiom $K_j(\chi \to \psi) \to (K_j\chi \to K_j\psi)$ for agent j yields the following rule

$$K_jK_i\varphi \to K_j\varphi$$

Applying this rule and the factivity axiom several times gives us the following general result:

[8] This example is from [74]. To see that it is not a level of knowledge, note that $K_2K_1x \to K_1x$ is derivable in S5, so by propositional reasoning we have that $\neg K_1x \to \neg K_2K_1x$ is derivable, by basic modal reasoning we can derive $K_1\neg K_1x \to K_1\neg K_2K_1x$. Applying the negative introspection axiom we can derive $\neg K_1x \to K_1\neg K_2K_1x$. Then, by propositional reasoning we have $\neg K_1\neg K_2K_1x \to K_1x$ is derivable. By basic modal reasoning, we then derive $K_2\neg K_1\neg K_2K_1x \to K_2K_1x$. This shows the set is inconsistent.

[9] In the belief case there are four possible levels. The three above with each K replaced by a B and a fourth level in which our agent believes x even though x is false. Thus, this level is given by $\{Bx, BBx \dots\}$.

Lemma 3.5. *Let $\varphi \in \mathcal{L}_K$, say $\varphi = K_{i_1} \dots K_{i_m} x$. Further, let $s_1 < \dots < s_r \in \{1, \dots m\}$. Then the following holds:*

$$K_{i_1} \dots K_{i_m} x \to K_{i_{s_1}} \dots K_{i_{s_r}} x$$

Or to say it differently: if $\varphi = K_{i_1} \dots K_{i_m} x$ is contained in some level of knowledge T, then so is every formula ψ that is obtained by removing any number of knowledge operators for φ and leaving the remaining operators in the original order. This shows that the sets X and Y from above contain the same information which is, in fact, the information already contained in the singleton $\{K_1 K_2 K_3 x\}$. To explore this property further, we let the property of the above lemma define a relation $\preceq$. That is, we define:

> Let $K_{j_1} \dots K_{j_r} x \preceq K_{i_1} \dots K_{i_m} x$ iff there is an order preserving embedding from the first to the second formulas, that is, a sequence $s_1 < \dots < s_r$ such that $K_{i_{s_l}} = K_{j_l}$

This pre-order has some intriguing structural properties, as has been found by Higman is his 1952 combinatorial lemma [80]. To state this result, recall that a **well-partial order** on a set X is an order in which all strictly descending sequences as well as all antichains, that is sets of mutually $\preceq$-incompatible elements, are finite.

Theorem 3.6 (Higman's Lemma)**.** *The order $\preceq$ is a well quasi order.*

While Higman derived this lemma from a more general result, we will provide an elementary derivation in the appendix. From Higman's lemma, we can immediately derive that there are only countably many levels of knowledge.

Theorem 3.7 (Theorem 3 of [129])**.** *There are only countably many levels of knowledge.*

Proof. By Lemma 3.5, every level $T \in \mathcal{T}_{\mathcal{L}_K}$ is downward closed under $\preceq$ and thus the complement $C = \mathcal{L}_K \setminus T$ is upward closed. Since $\preceq$ has no infinitely descending sequences, C is characterized by the set M of its minimal elements. Naturally the different minimal elements are mutually incomparable, thus M is an antichain and therefore finite. In particular, every level of knowledge is uniquely characterized by a finite subset of $\mathcal{L}_K$. Since $\mathcal{L}_K$ has only countably many finite subsets, there can be at most countably many different levels of information. □

Note that the only properties used in the proof of the countability result were normality and the factivity axiom. Thus we immediately get the following generalization:

Corollary 3.8. *Let $\Box_1 \dots \Box_n$ be normal modal operators that satisfy the factivity axiom. Then the language $\mathcal{L}_\Box$ generated by the set $\{\Box_1 \dots \Box_n\}$ has only countably many levels of knowledge.*

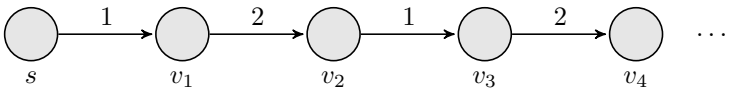
Next, we will show that Theorem 3.7 and Corollary 3.8 capture minimal conditions for ensuring that levels of information remain countable. If either the normality or the factivity axiom is given up, there will, in general, be uncountably many levels of information already for the most basic case of only two agents. We illustrate this along two examples, the case of belief as a normal operator violating the factivity axiom and the case of knowing whether as a non-normal modal operator. For the case of belief we have

Lemma 3.9. *The language $\mathcal{L}_B$ generated by $\{B_1, B_2\}$ has uncountably many levels of information.*[10]

Proof. We will show that the formulas φ_n defined by

$$\varphi_n := \underbrace{B_1 B_2 B_1 B_2 \dots}_{n \text{ operators}} x$$

are mutually independent. That is, for every $I \subseteq \mathbb{N}$ there is some level of information $T \in \mathcal{T}_{\mathcal{L}_B}$ such that $\varphi_n \in T \Leftrightarrow n \in I$. In particular, there are uncountably many levels of $\mathcal{L}_B$-information. To finish the proof, consider the following frame $\mathcal{F}$ where reflexive relations are omitted.



For any $I \subseteq \mathbb{N}$, let the model $\mathcal{M}^I$ based on $\mathcal{F}$ be defined by the valuation $Val(x) = \{v_i | i \in I\}$. Then it is not difficult to see that $\mathcal{M}^I, s \vDash \varphi_n \Leftrightarrow n \in I$. □

Thus, as soon as levels of information concern the doxastic states of several agents, they immediately become uncountable. In the rest of this chapter, we will therefore focus our attention on different aspects of *knowing* about a situation. To start, we consider the case of knowledge *whether*, analyzed by Hart et al. in [74]. Here, knowledge whether φ, written $J_i\varphi$, means that agent i knows the truth value of φ while being silent about that truth value, formally:

[10] This result has already been mentioned by Parikh in [128, Fact 5], however without giving a proof. To the best of our knowledge, no proof of this result has been published so far.

$J_i\varphi := K_i\varphi \vee K_i\neg\varphi$. Thus, knowledge whether is a non-normal modal operator[11] definable within multi-agent S5. The result obtained by Hart et al. is:

Theorem 3.10 (Theorem 2.2. of [74]). *Let $\mathcal{L}_J$ be the reasoning language generated by $\{J_1, J_2\}$. Then there are uncountably many levels of $\mathcal{L}_J$-information.*

They key idea of the proof provided in [74] is again to show that all formulas of the form $J_1J_2J_1 \ldots x$ are mutually independent. To see this, they construct a universal model, realizing uncountably many levels of information at once, as follows (we only sketch the main idea here, see the full paper for more details): Let $\mathcal{M}$ consist of all pairs (a, b); where $a, b \in \{0, 1\}^\omega$. Define the relations as $(a, b) \sim_1 (a', b')$ iff $a = a'$ and $a_k = a'_k = 1 \Rightarrow b_{k-1} = b'_{k-1}$ and similarly for $\sim_2$. Let φ denote the fact $a_0 = b_0 = 0$. Then an induction argument shows that $J_1J_2J_1 \ldots \varphi$ holds at $(\mathcal{M}, (a, b))$ iff $a_k = 1$ for k the number of quantifiers in $J_1J_2J_1 \ldots \varphi$ and similarly for $J_2J_1 \ldots x$ and $b_k = 1$. In particular, all those formulas are mutually independent, thus there are uncountably many levels of knowing whether.

3.2 Results

In the last section we have seen that the countability of levels of knowledge is closely tied to certain properties of the knowledge operator. But of course, not every reasoning language based on n knowledge operators keeps the levels of information countable. As illustrated by Theorem 3.10, the knowing whether modality, definable from the knowledge modality, negation and disjunction, makes for uncountably many levels of information. In this section, we explore the behavior of several extensions of $\mathcal{L}_K$ to slightly richer reasoning languages by incorporating conjunctions, disjunctions or negations.

Not every informational state we might be interested in can be expressed with the knowledge modality alone. For instance, recall our second example about the mountain climbers. There, some of the relevant propositions were of the form $K_1(K_2\varphi \vee K_3\varphi)$ expressing that agent 1 knows that either agent 2 or 3 supports φ, but might not necessarily know which of the two does. Thus, the appropriate reasoning language needs to contain some form of disjunctions. Yet other social situations might call for further operators such as negation, conjunction or the epistemic possibility operator.

[11] To give a bit more detail: Starting from a multi-agent knowledge model W, the neighborhood model for *knowing whether* is defined by the neighborhood relations: $n_i : W \to \mathcal{P}(\mathcal{P}(W))$ with $n_i(w) = \{X \subseteq W | X \supseteq K_i[w] \text{ or } X \cap K_i[w] = \emptyset\}$ where $K_i[w]$ is the knowledge cell of agent i containing w.

To give an example, assume that some researcher just learned about a highly interesting talk to be delivered in twenty minutes time. She has, however, a lunch appointment at the same time with two colleagues from her faculty whom she does not want to let down. These colleagues would, in fact, also attend the talk if they knew about it, thus a necessary condition for her to go is to know that her colleagues c_1 and c_2 know. The natural way to express this condition is $K(K_{c_1}x \wedge K_{c_2}x)$, requiring the conjunction operator next to the knowledge modalities. Or, for a second example, consider a slight variation of this setting. This time, instead of coordinating with her colleagues, the agent is trying to avoid yet another colleague at all price. This colleague is, at it stands, also interested in that talk, thus a necessary condition for her going is to be sure that said colleague is not aware of the talk, or formally $K\neg K_a x$.

As these examples show, there are plenty of situations that cannot be adequately described with the knowledge modalities alone, but we might need to add further operators such as conjunction, disjunction or negation. Sometimes, these operators only occur in some highly constrained surroundings such as the definition of the knowing-whether operator form negation and disjunction ($J\varphi := K\varphi \vee K\neg\varphi$) or of epistemic possibility ($L = \neg K\neg$), at other times these operators might be added unrestrictedly. In the following we inquire into adding these conjunctions, disjunctions and negations both restrictedly and unrestrictedly.

We begin our inquiry with the negation operator. For cases such as the researcher trying to avoid some colleague we need to incorporate negations, the knowledge that somebody else does *not* know something, into our reasoning language. We start our inquiry with a cautious introduction of negation through the epistemic possibility operator L, definable from K as $L = \neg K\neg$. Note that for instance in our very first example about the pedestrian trying to cross the road, we do not need the driver to know that the pedestrian does not *know* about him approaching. It would suffice for the driver to consider it possible that the pedestrian might not be paying attention, i.e., $L_{driv.}L_{pedes.}x$. The following is, in a certain sense, a dual of Theorem 3.7.

Lemma 3.11. *Let $\mathcal{L}_L$ be the reasoning language generated by $\{L_1, \ldots, L_n\}$. Then there are at most countably many levels of $\mathcal{L}_L$-information.*

Proof. For a given $\varphi \in \mathcal{L}_L$ let $\varphi^{\#} \in \mathcal{L}_K$ be the formula resulting by replacing every L_i with the corresponding K_i. Then we claim that the map f sending $T \in \mathcal{T}_{\mathcal{L}_L}$ to $f(T) = \mathcal{L}_K \setminus \{\varphi^{\#} | \varphi \in T\}$ is a bijection between $\mathcal{T}_{\mathcal{L}_L}$ and $\mathcal{T}_{\mathcal{L}_K}$. To see this, let T be a level of $\mathcal{L}_L$ information and let $(\mathcal{M}, w)$ be a model realizing

T, i.e., for every $\varphi \in \mathcal{L}_L$ holds that $\varphi \in T \Leftrightarrow (\mathcal{M}, w) \vDash \varphi$. Then, using the identity $L_i = \neg K_i \neg$, we have for every $\varphi \in \mathcal{L}_L$ with $\varphi = L_{i_1} \dots L_{i_r} x$

$$\varphi \in T \Leftrightarrow \mathcal{M}, w \vDash L_{i_1} \dots L_{i_r} x \Leftrightarrow \mathcal{M}, w \vDash \neg K_{i_1} \dots K_{i_r} \neg x \Leftrightarrow \mathcal{M}, w \nvDash \varphi^{\#}[\frac{\neg x}{x}].$$

Thus, $\mathcal{M}'$, the model $\mathcal{M}$ where the valuation $V(x)$ is replaced by $V'(x) = W \setminus V(x)$, satisfies for all $\psi \in \mathcal{L}_K$ that $\mathcal{M}', w \vDash \psi \Leftrightarrow \varphi \in f(T) \Leftrightarrow \psi \notin \{\varphi^{\#} | \varphi \in T\}$, thus showing that f is indeed an injective function. The same construction applied to $\mathcal{L}_K$ shows that the inverse of f is also well defined and injective, thus f is a bijection. □

Next, we show that slightly less generosity about negation operators, expressed by allowing for K_i and L_i to appear simultaneously, blows up the expressive power drastically.

Lemma 3.12. *Assume there are at least two agents and let $\mathcal{L}_{L,K}$ be the language generated by $\{L_1 \dots L_n, K_1 \dots K_n\}$. Then there are uncountably many levels of $\mathcal{L}_{L,K}$-information.*

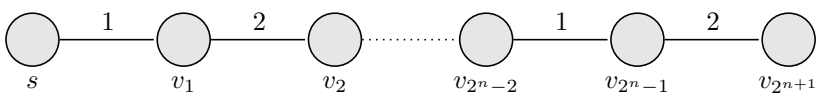
Of course, we immediately get the following generalization, using that $L_i = \neg K_i \neg$:

Corollary 3.13. *Assume there are at least two agents and let $\mathcal{L}_{K,\neg}$ be the language generated by $\{\neg, K_1 \dots K_n\}$. Then there are uncountably many levels of $\mathcal{L}_{K,\neg}$-information.*

Proof of Lemma 3.12. The proof strategy is similar as in the proof of Lemma 3.9. We will show that the formulas φ_n defined by

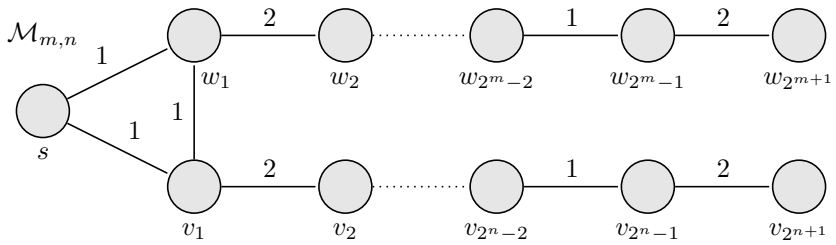
$$\varphi := \underbrace{L_1 L_2 \dots L_1 L_2}_{2^n \ (L_1 L_2) \text{ blocks}} \underbrace{K_1 K_2 \dots K_1 K_2}_{n \ (K_1 K_2) \text{ blocks}} x$$

are mutually independent. That is for every $I \subseteq \mathbb{N}$ there is some model $\mathcal{M}_I, s$ such that $\mathcal{M}_I, s \vDash \varphi_n$ iff $n \in I$. To this end, let the model $\mathcal{M}_n$ be defined as:



with the valuation given by $V(x) = \{v_i | 2^{n+1} - 2n \leq i \leq 2^{n+1}\}$, thus x is true at the $2n+1$ right-most worlds. It is not difficult to see that $\mathcal{M}_n, s \vDash \varphi_n$ and $\mathcal{M}_n, s \nvDash \varphi_m$ for all $m \neq n$. Now, for any subset $I \subseteq \mathbb{N}$ we define the model

$\mathcal{M}_I$ in the following way: We take all $\mathcal{M}_n$ for $n \in I$, identify all the points s from the different $\mathcal{M}_n$ and replace the equivalence relations by their transitive closure. For the case $I = \{m, n\}$, we thus get the following picture:



Now, it is not difficult to see that $(\mathcal{M}_I, s) \vDash \varphi_n$ iff $n \in I$, which completes our proof. □

The last lemma finishes our analysis of negations and we can shift our attention to conjunctions. In the first part of the above example, we used conjunctions to express the fact that some researcher knows that both of her colleagues are aware of some interesting talk. Intuitively, this information is already contained in her level of knowledge. Saying that the agent knows that both her colleagues know about the talk is nothing else but saying that she knows that the first colleague knows about the talk and that she also knows that the second colleague knows about the talk. This intuition is made precise in the following lemma, showing that conjunctions do not change the expressive power of our reasoning language.

Lemma 3.14. *Let $\mathcal{L}_{K,\wedge}$ be the language generated by $\{K_1, \ldots, K_n, \wedge\}$. Then there are only countably many levels of $\mathcal{L}_{K,\wedge}$-information. Furthermore, there is a bijection $f : \mathcal{T}_{\mathcal{L}_{K,\wedge}} \to \mathcal{T}_{\mathcal{L}_K}$ such that any pointed Kripke model $(\mathcal{M}, w)$ realizes some level $T \in \mathcal{T}_{\mathcal{L}_{K,\wedge}}$ if and only if $(\mathcal{M}, w)$ realizes $f(T)$.*

Proof of Lemma 3.14. Let $f : \mathcal{T}_{\mathcal{L}_{K,\wedge}} \to \mathcal{T}_{\mathcal{L}_K}$ be the map sending every $T \in \mathcal{T}_{\mathcal{L}_{K,\wedge}}$ to $T \cap \mathcal{L}_K$. To see that f is well-defined, note that every model $(\mathcal{M}, w)$ realizing T also realizes $f(T)$. For surjectivity let $S \in \mathcal{T}_{L_K}$ and pick some model $(\mathcal{N}, w)$ realizing S. Let S' be the $\mathcal{L}_{K,\wedge}$ level defined as $S' = \{\varphi \in \mathcal{L}_{K,\wedge} | (\mathcal{M}, w) \vDash \varphi\}$, thus $f(S') = S$. Finally, for injectivity, note that the rule $K_i(a \wedge b) \Leftrightarrow K_i a \wedge K_i b$ is valid on all knowledge models. Thus, by applying this rule repeatedly, every $\varphi \in \mathcal{L}_{K,\wedge}$ is equivalent to a formula of the form $\psi_1 \wedge \ldots \wedge \psi_n$ where the ψ_i do not contain the symbol $\wedge$, i.e., $\psi_i \in \mathcal{L}_K$. Thus, we have for all $T \in \mathcal{T}_{\mathcal{L}_{K,\wedge}}$ that $\varphi \in T \Leftrightarrow \psi_1 \wedge \ldots \wedge \psi_n \in T \Leftrightarrow \psi_1, \ldots, \psi_n \in T$, i.e., every level $T \in \mathcal{T}_{\mathcal{L}_{K,\wedge}}$ is already determined by $T \cap \mathcal{L}_K$. □

Finally, we turn to the analysis of disjunctions. The disjunction operator can be used to weaken the concept of knowledge in various ways. Knowledge whether, for instance, states the presence of some information while being silent about its content, while distributed knowledge expresses that some member knows φ without saying who does. Or, even more generally, we might use disjunctions to express the fact that some agent knows that one out of a list of propositions is true without knowing which. As it will turn out, the disjunction connective is the most delicate of the extensions studied. For a start, we consider a cautions introduction of disjunctions, used only to define distributed knowledge. For any subset $J \subseteq I$ of agents, we define the distributed knowledge operator D_J as

$$D_J\varphi := \bigvee_{i \in J} K_i\varphi.$$

Thus, D_J expresses that some members of J know φ without specifying who. Note that $D_{\{i\}} = K_i$, thus the reasoning language $\mathcal{L}_D$ defined by $\{D_J \mid J \subseteq I\}$ is an extension of $\mathcal{L}_K$. In fact, $\mathcal{L}_D$ is a proper extension of L_K and it is even more expressive than $\mathcal{L}_K$, that is there are two models $(\mathcal{M}, w)$ and $(\mathcal{M}', v)$ that satisfy exactly the same formulas from $\mathcal{L}_K$, that is $\mathcal{M}, w \vDash \varphi \Leftrightarrow \mathcal{M}', v \vDash \varphi$ for all $\varphi \in \mathcal{L}_K$, but these models are distinguishable in $\mathcal{L}_D$, i.e., there is some $\psi \in \mathcal{L}_D$ such that $\mathcal{M}, w \vDash \psi$ and $\mathcal{M}', v \nvDash \psi$. Nevertheless, the language $\mathcal{L}_D$ still only allows for countably many levels of knowledge.

Lemma 3.15. *Let $\mathcal{L}_D$ be the reasoning language defined by $\{D_J \mid J \subseteq I\}$. Then $\mathcal{L}_D$ has only countably many levels of knowledge.*

Proof. We derive this statement from corollary 3.8. First, we note that, by a simple case distinction, the modal operators D_J all satisfy the factivity axiom. The D_J are, however, no normal modal operators, but they satisfy the property

$$D_J D_{J'}\varphi \to D_J\varphi$$

which is all that was needed for the proof of 3.8. To see this, let $D_J = \bigvee_{i \in J} K_i$ and assume that $D_J D_{J'}\varphi$ holds. Then there is some $i \in J$ such that $K_i D_{J'}\varphi$ holds, which, by the normality of $D_{J'}$, implies that $K_i\varphi$. By $\vee$-introduction we finally get that $\bigvee_{i \in J} K_i\varphi = D_J\varphi$, as desired.

□

Next, we turn to a less cautious use of disjunctions. We study the full language obtained by adding disjunctions unrestrictedly, i.e., the language generated by

$\{K_1, \ldots, K_n, \vee\}$. As it turns out, the properties of this language crucially depend upon the number of agents. It is a well known phenomenon in modal logic that the behavior of knowledge models for two agents can differ radically from models with three or more agents in certain aspects. The following two lemmas show that levels of information are such a case:

Lemma 3.16. *Let $\mathcal{L}_{\vee 2}$ be the language generated by $\{K_1, K_2, \vee\}$. Then $\mathcal{L}_{\vee 2}$ has only countably many levels of information.*

Lemma 3.17. *Let $\mathcal{L}_{K,\vee}$ be the language generated by $\{K_1, \ldots, K_n, \vee\}$ for $n \geq 3$. Then $\mathcal{L}_{K,\vee}$ has uncountably many levels of information.*

We relegate the slightly lengthy proofs of these lemmas to the appendix. We want to emphasize here that Lemma 3.17 in a certain sense represents minimal conditions for arriving at uncountably many levels. Several slight restrictions, such as adding an upper bound to the number of disjunctions allowed to appear in any formulas or an upper limit for the quantifier depth under which conjunctions may occur, turns levels of information countable again. To conclude this chapter, we collect the results we have presented so far in a theorem:

Theorem 3.18. *i) The following reasoning languages have countably many levels of information:*

Reasoning language	*generated by*
$\mathcal{L}_K$	$\{K_1 \ldots K_n\}$ *(Parikh/Krasucki)*
$\mathcal{L}_L$	$\{L_1 \ldots L_n\}$
$\mathcal{L}_{K,\wedge}$	$\{K_1, \ldots, K_n, \wedge\}$
$\mathcal{L}_D$	$\{D_J \mid J \subseteq I\}$ *where* $D_J\varphi := \bigvee_{i \in J} K_i\varphi$
$\mathcal{L}_{\vee 2}$	$\{K_1, K_2, \vee\}$

ii) The following reasoning languages have uncountably many levels of information:

Reasoning language	*generated by*
$\mathcal{L}_B$	$\{B_1 \ldots B_n\}$
$\mathcal{L}_{L,K}$	$\{K_1, \ldots, K_n, L_1 \ldots L_n\}$
$\mathcal{L}_{K,\neg}$	$\{K_1 \ldots K_n, \neg\}$
$\mathcal{L}_J$	$\{J_1, \ldots, J_n\}$ *where* $J_i\varphi = K_i\varphi \vee K_i\neg\varphi$ *(knowing whether, Hart et al.)*
$\mathcal{L}_{K,\vee}$	$\{K_1, \ldots, K_n, \vee\}$ *for* $n \geq 3$

3.3 Realizing Levels of Information

Having discussed the different reasoning languages and their expressive power abstractly, it is now time to relate levels of information back to social situations where they could arise. In the two examples we gave in the introduction, a

pedestrian crossing a street and strayed mountaineers discussing their options, we used levels of information to denote an *ideal* state of information that some of the agents want to achieve. So let us assume that we identified some level of information T that we wish to bring about. The natural question to ask here is whether this is possible at all and, if so, how. In the broadest sense, the answer to the first question is trivial. By the completeness theorem, for a level of information T, being consistent just means that T is realizable in some Kripke model $(\mathcal{M}, w)$. However, so one could argue, there are limits to which we can realize Kripke models in a realistic multi-agent setting. Given that all our information is finite, as are the agents whose reasoning the Kripke model represents, a natural criterion would be that the Kripke models realizing some levels of information should also be finite. If our reasoning language was finite, the finite model property (see [22, section 3.4] for details)would guarantee exactly that: Every consistent formula, and thus also every finite set of consistent formulas of the knowledge or belief language is realizable in a finite Kripke model. This does, however, not hold true anymore for infinite reasoning languages. By a well-known fact from infinite combinatorics, there are only countably many finite Kripke Models of a given language. Thus, if our reasoning language allows for uncountably many levels of information, we cannot hope for all of them being realizable in finite Kripke Models. The following result shows that for the reasoning languages we studied in the last section, the converse holds also true. If we pick a reasoning language that allows for only countably many different levels of information, every such level can be represented in a finite Kripke Model. Thus, the following theorem can be seen as an extension of the finite model property mentioned above. The proof is again relegated to the appendix.

Theorem 3.19. *Let $\mathcal{L}_c$ be any of the reasoning languages in part i) of Theorem 3.18 and let T be a level of $\mathcal{L}_c$ information. Then T is realizable in a finite model.*

Also this answer might not be fully satisfactory for some practical purposes. In most cases, the agents will have some prior information about the situation or about each others' information. For instance, some of the mountaineers might have already stated their positions publicly or the pedestrian may have visibly reacted to the car. Thus, rather than producing some informational situation from scratch, our efforts to bring about T will have to start with the prior informational setting of the agents. And of course, this prior information may limit the levels of information that are still realizable. For instance, we can provide agents with additional information, truthful or not, but we cannot force

them to forget whatever they already know or believe.[12] Thus, the question to ask here is: Under which circumstances can some given level of information, describing the initial information of the agents, be transformed into some other level of information that we wish to bring about? We give a partial answer to this question for the case of $\mathcal{L}_K$. Note that this analysis immediately extends to the reasoning languages $\mathcal{L}_{K,\wedge}, \mathcal{L}_{\vee 2}$ and $\mathcal{L}_L$: Each level of information T for the former two languages is determined by $T \cap \mathcal{L}_K$ (see Lemmas 3.14 and 3.16), while each level of $\mathcal{L}_L$ information realized in some model $(\mathcal{M}, w)$ is uniquely determined by the level of knowledge of $\neg x$ realized at $(\mathcal{M}, w)$ (see Lemma 3.11). For the following theorem, recall the definition of event models (Definition 2.3 of chapter 2). Before stating our result we need to introduce one more piece of notation. For some level of knowledge T, we will denote the set of $\preceq$-minimal elements of $\mathcal{L}_K \setminus T$ by $M(T)$. By Theorem 3.7 the set $M(T)$ is finite for each level of knowledge T, and T is completely determined by $M(T)$.

Theorem 3.20. *Let $L_1 \neq \emptyset$ and L_2 be consistent $\mathcal{L}_K$ levels of knowledge for some atomic or boolean proposition x. Let $\mathcal{M}, w$ be a finite Kripke model realizing L_1. Then there is an event model (E, e) such that $(\mathcal{M}, w) \oplus (\mathcal{E}, e)$ realizes L_2 if for every $\varphi \in M(L_2)$ there is some $\psi \in M(L_1)$ such that*

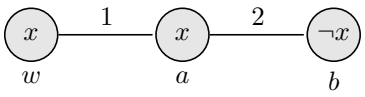
i) $\psi \preceq \varphi$

ii) Let $\varphi = K_{i_1} \dots K_{i_r} x$ *and* $\psi = K_{j_1} \dots K_{j_s} x$. *Then* $K_{i_r} = K_{j_s}$.

The first condition simply expresses that $L_1 \subseteq L_2$, that is, the knowledge about φ can only increase. However, the following example 3.21 shows that L_1 may not contain all the information in $(\mathcal{M}, w)$ relevant for which levels of knowledge could be obtained by product updates from $(\mathcal{M}, w)$. Condition *ii)* of the theorem precisely excludes this type of counterexamples. As is well known, no positive knowledge can be lost through product updates. Thus, condition *i)*, stating precisely this, is also a necessary condition for which levels of knowledge are reachable from $(\mathcal{M}, w)$. This condition is even a maximal necessary condition in the following sense. For every $L_1 \subseteq L_2$ there is some model $(\mathcal{M}, w)$ realizing L_1 and some event model $(\mathcal{E}, e)$ such that $(\mathcal{M}, w) \oplus (\mathcal{E}, e)$ realizes L_2.

Example 3.21: To see that $\mathcal{M}, w$ may restrict the levels of knowledge realizable through product updates beyond what is expressed in L_1, consider the following three(!) agent model $\mathcal{M}$, where reflexive relations are, as usual, omitted.

[12] We could, however, try to make some agent *overwrite* their information. This falls in the realm of belief revision that we do not touch upon in this work.



It is easy to see that the level of information T realized in $\mathcal{M}, w$ is characterized by $M(T) = \{K_1K_2x\}$, thus T consists of exactly those formulas in $\mathcal{L}_K$ that do not contain the letters K_1 and K_2 in this order. Next, consider the type T' characterized by $M(T') = \{K_1K_2K_3x\}$. Since $K_1K_2x \prec K_1K_2K_3x$ we have that $T \subset T'$. Yet, T' cannot be realized in any update model of $\mathcal{M}, w$. To see this, observe that the formula $K_3x \vee K_3\neg x$ is valid on $\mathcal{M}$ and thus also on all product updates of $\mathcal{M}$. Thus, using factivity, $x \to K_3x$ is valid on all updates of $\mathcal{M}$, which, in turn, implies that $K_1K_2x \to K_1K_2K_3x$ is also a validity. By contraposition we get $\neg K_1K_2K_3x \to \neg K_1K_2x$ implying that no level of knowledge $\tilde{T}$ obtainable through an update can have $K_1K_2K_3x \in M(\tilde{T})$.
A next direction of research now would be to ask which levels of knowledge could be realized if the means of communication are limited. For instance, Parikh and Krasucki [129] study which levels of knowledge can arise under private communication with delays and which levels require public announcements. As their analysis shows, a combination of both methods suffices to bring about every level of knowledge. For our general levels of information, this is no longer true. For instance, the level of information desired by the mountain climbers in our initial example, everybody knowing *how many* group members support x, but not which, is reachable through communication among the mountain climbers, but only if there is common knowledge in some underlying protocol, a case that is excluded by Parikh and Krasucki. Thus, the question arises as to which means of communication are necessary to realize all possible levels of information. The last lemma reduces this question for many reasoning languages to an analysis of product updates and the question which communication channels are needed to produce certain update models. We leave this question for future work.

3.4 Conclusions and Outlook

Information is an important building block in our understanding of social procedures. As is emphasized by a vast body of work, ranging from epistemic game theory [131] to the analysis of conventions [106] and social norms [19], the epistemic and doxastic states of the various agents decisively shape the way in which interactive situations pan out. In this chapter, we have concentrated on an abstract representation of the epistemic and doxastic states present in a situation. Our main goal was to compare several languages we could use to represent epis-

temic states, all of them fragments of the full epistemic language $\mathcal{L}$, with respect to their expressive power and their realizability, that is the question of whether an epistemic state is realizable in a *finite* situation or whether it requires an infinite model. On the expressive power side, we distinguish languages allowing for only countably many levels of information, consistent descriptions of a situation, from those that give rise to uncountably many different levels. Our first main result is to identify which operators and junctors of the logical vocabulary cause an explosion in expressive power and which do not. In particular, we show that negations and an unlimited use of disjunctions raise the expressive power drastically while conjunctions and a limited use of disjunctions have a slight or no effect on the expressive power. Our second main result is related to the realizability of levels of information, that is to the question whether *every* consistent level of information is realizable in a finite Kripke Model. Here, we have shown that all those languages studied that allow for at most countably many different levels of information guarantee in return that every possible such level is already realized in some finite situation. We end this chapter with indicating some possible topics for future research.

A first direction for extended research was already indicated at the end of the previous section, namely to clarify the connection between communicational means and levels of information. In their attempt to realize a certain level of information, a set of agents may only have access to a limited amount of communication channels. A group of card players sitting at a table will only have access to public announcements, while a group of friends at a distance, communicating over text messages, are limited to asynchronous messaging with uncertain success. In the first case it is hard to exchange information between two parties without the rest of the table learning about it [43], in the second case it is impossible to generate common knowledge among any subgroup of agents [71]. Thus, the main question in this research stream is which levels of information can be realized within a given, limited set of communicational tools or, conversely, what kind of communication is needed to bring about some desired level of information.

The second direction of future research is to extend our analysis to more general reasoning languages. So far, we have primarily focused on reasoning languages that are based on the knowledge or belief modalities. Of course, there are interesting extensions to this. First of all, it would be interesting to see how an interaction of these modalities work for instance studying the beliefs of one agent about the knowledge of another agent. Here, even the

basic case, that is the reasoning language generated by $\{K_1, B_2\}$ makes for uncountably many levels of information. A related direction of research would be to introduce relations between the different agents. For instance, if some agent i gets accepted as a guru by agent j, the formula $B_j B_i \varphi \to B_j \varphi$ would be part of the underlying axioms. Also here the question would be which and how many correlations between agents can make for countably or uncountably many levels. Some other cases will call for different modal operators altogether. To give a prominent example, Bicchieri's analysis of conventions and social norms [19] can be translated into levels of information for a single proposition φ, if we add additional modal operators D_i, where $D_i\varphi$ denotes that agent i is prepared to do φ in certain types of situations. With this language, the fact that some agent i has a convention about φ in Bicchieri's sense translates to

$$D_i\varphi \leftrightarrow B_i \left(\bigvee_{\substack{J \subseteq I \\ \text{big enough}}} \bigwedge_{j \in J} D_j \varphi \right)$$

meaning that agent i adheres to a descriptive norm regulating φ whenever she is willing to follow φ if and only if she expects a large group of agents, denoted by the *"big enough"*, to do likewise. In this same style, the fact that i has a social norm about φ translates to

$$D_i\varphi \leftrightarrow B_i \left(\bigvee_{\substack{J \subseteq I \\ \text{big enough}}} \bigwedge_{j \in J} D_j \varphi \wedge B_j D_i \varphi \right)$$

That is, agent i has a social norm concerning φ if she is willing to do φ whenever she expects enough people to also follow φ in such circumstances and if she expects these people to expect her to follow φ under the appropriate circumstances.

In light of corollary 3.8, we expect that much of the analysis presented here carries over to a quite general class of modal operators. But, of course, details need to be checked.

Finally, a third and last direction of future research is how to combine different levels of information. Within some given situation, we might have learned about the available information concerning some proposition p and also about the level of information about some q. But what does this imply about the information concerning more complex formulas such as $p \to q$ or $p \wedge q$? Let's consider the reasoning language $\mathcal{L}_K$ as an example. There, the level of knowledge about $p \wedge q$ is exactly the intersection of the level of knowledge of p and

q. However, this is not true for disjunction: The level of knowledge of $p \vee q$ can be a proper superset of the union of the levels of knowledge of p and q. The general question is how the levels of information of ϕ and ψ are related to the levels of information of $\phi \vee \psi, \phi \wedge \psi, \phi \to \psi$, etc. More generally, we can define a map Ψ assigning each formula of our language $\mathcal{L}$ a level of information over some reasoning language $\mathcal{L}_r$. Can we characterize such maps, depending on $\mathcal{L}_c$? For example, for $\mathcal{L}_K$ we have $\Psi(\varphi) \neq \emptyset$ implies $\Psi(\neg\varphi) = \emptyset$, whereas, for $\mathcal{L}_B$ we only have that $\Psi(\varphi) \cap \Psi(\neg\varphi) = \emptyset$.

3.5 Appendix: Proofs

Proof of Theorem 3.6. To start, we introduce a simplifying assumption. Throughout this proof, we will identify every formula $K_{i_1}K_{i_2}\dots K_{i_n}x$ with the corresponding word $K_{i_1}K_{i_2}\dots K_{i_n}$, omitting the x. To introduce a bit of notation: For a given set $\Sigma = \{K_1 \dots K_n\}$ let Σ^* denote the set of finite words in Σ, i.e., finite sequences where all members are from Σ. Now, define the relation $\leq$ on $\Sigma*$ by: $x \leq y$ iff there is an order preserving injection from x to y. Thus, we have to show that $\leq$ is a well quasi order on Σ^*. The proof will proceed in several steps.

Step 1: The fact that $\leq$ is a well quasi order is equivalent to the statement that every infinite set X in Σ^* contains an infinite $\leq$-increasing sequence $S = \langle x_i | i \in \mathbb{N}\rangle$. To show this, we start with the direction from right to left. Assume that $\leq$ was not a well quasi order. Then it has either an infinite antichain or an infinite $\leq$-decreasing subsequence. Neither of these contain an infinite increasing subsequence. For the direction from left to right assume $X \subseteq \Sigma^*$ is an infinite subset. Assume to the contrary that every increasing sequence in X is finite. Let $Y \subseteq X$ be the set of elements appearing as maximal elements in some maximally increasing subsequence of X. Then Y is an antichain and thus finite. On the other hand, the relation $x \leq y$ implies that the word x is not longer than y, thus $\{x \in X | x \leq y \text{ for some } y \in Y\}$ is finite. But this contradicts the fact that Y consists of all maximal elements of maximally increasing subsequences.

Step 2: Let $n \in \mathbb{N}$ and define $\leq^n$ on $(\Sigma^*)^n$ componentwise. Then $\leq$ is a well quasi order on Σ iff $\leq^n$ is a well quasi order on Σ^*. We prove this step by induction on n. The case $n = 1$ is trivial. Assume the claim holds for n, and we have to show it for $n + 1$. Let $X \subseteq (\Sigma^*)^{n+1}$ be infinite. By Step 1, there is an infinite sequence $X' \subseteq X$ such that $\leq$ is a linear order on the last components of X'. Applying the induction assumption to the first n components of X yields the desired.

Step 3: The fact that $\leq$ is a well quasi order is equivalent to the claim that for every infinite $X \subseteq \Sigma^*$ there is some $x \in X$ such that the set $X' = \{y \in X | x \leq y\}$ is infinite. For the direction from left to right pick X' as in step 1 and $x = min(X')$. For the direction from right to left pick such x and X'. Then, applying the claim to the infinite set X', we can pick some x' such that $X'' \in \{y \in X | x' \leq y\}$ is infinite. Iterating this construction infinitely often yields an increasing sequence $\langle x, x' \ldots \rangle$.

Step 4: Now we can finally prove the theorem. We will do so by induction over n, the size of alphabet. The case $n = 1$ is trivial, thus assume we have shown it for $n-1$ and want to show it for n. Let X be infinite and let x be a shortest word in X, let k be its length. Let w be the word.

$$w = \underbrace{K_1 K_2 \ldots K_n K_1 K_2 \ldots K_n \ldots K_1 K_2 \ldots K_n}_{k \text{ times}}$$

Thus we have $x \leq w$, though not necessarily $w \in X$. Now, for every $y \in X$, we define a partial map $f : w \to y$ with the following algorithm: Send the the first letter of w, K_1, to the first K_1 appearing in y. Then, send the second letter of w, K_2 to the first K_2 that appears in y after the image of K_1 just chosen, and so forth. Stop this algorithm whenever some letter of w cannot be mapped successfully because it doesn't appear in y after the image of the previous letter. Furthermore, let $m(y)$ denote the number of letters that could be mapped successfully before the algorithm stopped, that is $m(y) \in [0; k \cdot n]$. Now, pick some infinite set $X' \subseteq X$ such that $m(\cdot)$ is constant on X'. Here we distinguish two cases:

Case 1: $m(y) = n \cdot k$ for all $y \in X'$. Thus the algorithm described above defined an order preserving injection $f_y : w \to y$ for every $y \in X'$. In particular we have $w \leq y$ and thus also $x \leq y$ and therefore (x, X') is as in step 3.

Case 2: $m(y) = r < n \cdot k$ for all $y \in X'$. In this case let $w_1 \ldots w_{r-1}$ be the first $r-1$ letters of the word w. Thus, every $y \in X'$ can be uniquely written in the form

$$M_1^y w_1 M_2^y w_2 \ldots w_{r-1} M_r^y$$

where each M_i^y does not contain the letter w_i. Trivially, we have $y < y'$ iff $M_i^y < M_i^{y'}$ for every $i \leq r$. But since each M_i^y does not contain letter w_r, it is already defined over a vocabulary with $n-1$ words. Thus, by the induction hypothesis and step 2 applied to the set $\{(M_i^y)_{i \leq r} | y \in X'\}$, there is some infinitely increasing $\leq$-sequence X'' in X'. $\square$

Proof of Lemma 3.16. We show that every $\varphi \in \mathcal{L}_{\vee 2}$ is equivalent to some $\psi \in \mathcal{L}_K$. In particular, every $T \in \mathcal{T}_{L_{\vee 2}}$ is uniquely determined by $T \cap \mathcal{L}_K$ and thus $\mathcal{T}_{L_{\vee 2}}$ is countable. To this end, we note that for any $\varphi \in \mathcal{L}_{\vee 2}$ of the form $\varphi = \psi_1 \vee \ldots \vee \psi_n$ and for every level of information $T \in \mathcal{T}_{\mathcal{L}_{\vee 2}}$, it holds that $\varphi \in T$ iff $\psi_i \in T$ for some i. Thus, every $T \in \mathcal{T}_{\mathcal{L}_{\vee 2}}$ is uniquely determined by its elements φ of the form $\varphi = K_i\psi$ for some $i \leq n$ and $\psi \in \mathcal{L}_{\vee 2}$. It therefore suffices to show that for every φ of that form there is some $\psi \in L_K$ with $\varphi \leftrightarrow \psi$. Before we proceed, note that the following three rules are valid:

$$\begin{array}{ll} (I) & K_i\varphi \vee K_i\psi \to K_i(\varphi \vee \psi) \\ (II) & K_i(K_i\varphi \vee \psi) \to K_i(\varphi \vee \psi) \\ (III) & K_i(x \vee \psi) \to K_i x \text{ for } \psi \in \mathcal{L}_{\vee 2} \end{array}$$

The first two are general validates, the third follows the fact that $\varphi \to x$ holds for all $\psi \in \mathcal{L}_{\vee 2}$, which in turn follows from the factivity axiom by induction over the complexity of φ. Now we can finally prove that for every $\varphi \in L_{\vee 2}$ of the form $\varphi = K_i\psi$ there is some $\psi \in \mathcal{L}_K$ with $\varphi \leftrightarrow \psi$. Without loss of generality, we can assume that $\varphi \notin \mathcal{L}_K$. We will construct a sequence

$$\varphi = \psi_0 \to \psi_1 \to \ldots \to \psi_n = \psi$$

Assume ψ_i is given and $\psi_i \notin \mathcal{L}_K$. To construct ψ_{i+1} write ψ_i in the form $K_{i_1} \ldots K_{i_r}(\varphi_1 \vee \ldots \vee \varphi_n)$. Now, if any of the φ_i is x, rule III from above implies $\psi_i \to K_{i_1} \ldots K_{i_r} x$. Set $K_{i_1} \ldots K_{i_r} x = \psi_{i+1}$ and end the construction. If none of the φ_i is x, check whether there are φ_l, φ_m of the form $K_a\chi_l$ resp. $K_a\chi_m$ for some $a \in \{1, 2\}$. If this is the case, apply rule (I) to $\varphi_i \vee \varphi_j$, that is $\psi_{i+1} = K_{i_1} \ldots K_{i_r}\left(K_a(\chi_i \vee \chi_j) \vee \bigvee_{i \neq l,m} \varphi_i\right)$. Finally, if there are no φ_l, φ_m of this form, then some of the φ_i must be of the form $K_{i_r}\chi$, wlog φ_1 is of that form. In this case apply rule (II) and let $\psi_{i+1} = K_{i_1} \ldots K_{i_r}(\chi_1 \vee \ldots \vee \varphi_n)$.

Note that every application of rules (I) or (II) reduces the number of K_i operators by one, thus there can be only finitely many applications of these rules before applying rule (III) and stopping the algorithm. Thus, we have $\varphi \to \psi$. To finish the proof, we need to show the converse, that is $\psi \to \varphi$. To this end, we need to introduce a bit of vocabulary. For $\varphi, \psi \in \mathcal{L}_{\vee 2}$, say that ψ is a *pruning* of φ iff ψ can be obtained from φ by repeatedly replacing some disjunctions $(\chi_1 \vee \chi_2)$ appearing in φ by one of the disjuncts χ_1 or χ_2. By $\vee$-introduction we have that $\psi \to \varphi$ whenever ψ is a pruning of φ. Further call ψ a *complete* pruning of φ if ψ does not contain any further $\vee$-operators and let S^φ denote the set of complete prunings of φ. Now, it is not difficult to see that rules (I) and (II) leave S^φ invariant, that is $S^\varphi = S^{\psi_0} = \ldots = S^{\psi_{n-1}}$. Further,

by our application of rule (III), we have that $\psi_n \in S^{\psi_{n-1}} = S^{\varphi}$. Thus ψ is a pruning of φ and thus $\psi \to \varphi$.

□

Proof of Lemma 3.17. We will show that the reasoning language $\mathcal{L}_{K,\vee}$ generated by $\{K_1, K_2, K_3, \vee\}$ has uncountably many levels of information. In the following, we will mimic substantial parts of the proof for the belief case (Lemma 3.9). In particular, we will again define a set of formulas φ_n such that all φ_n are mutually independent, i.e., such that for all $I \subseteq \mathbb{N}$ there is some model $\mathcal{M}, w$ such that $\mathcal{M}, w \vDash \varphi_n \Leftrightarrow n \in I$. To be a bit more precise, we will construct two unary operators B_1 and B_2 and a formula χ, all definable in $\mathcal{L}_{K,\vee}$, such that the formulas of the form

$$\varphi_n := \underbrace{B_1 B_2 \dots}_{n \text{ operators}} \chi$$

are all mutually independent. Just as in the belief case, we would like to build a linear Kripke frame, based on an infinite set of worlds $\{v_1, v_2, v_3 \dots\}$. However, in order to define our operators, we need to connect an additional five-node cluster to each v_i. Thus, we will work will the following Kripke Frame $\mathcal{F}$ depicted in table 3.1 where reflexive arrows are omitted.

Now, let $I \subseteq \mathbb{N}$ be any subset. We define a model $\mathcal{M}^I$ on $\mathcal{F}$ by picking the following valuation V.

$v_i, u_i \in V(x)$	for all $i \in \mathbb{N}$	$x_i \in V(x)$	iff i even
$y_i, z_i \in V(x)$	iff $i \in I$	$w_i \in V(x)$	iff i odd

Thus x is true on the darkly shaded areas above and each lightly shaded area labeled with y_i or z_i are in $V(x)$ iff $i \in I$. We now define the operators B_1 and B_2 as:

$$B_1\varphi := K_1\left(K_3K_1x \vee \varphi\right) \qquad B_2\varphi := K_2\left(K_3K_2x \vee \varphi\right)$$

First, note that on the set of worlds labeled with v_i, the operator B_i behaves like a belief operator. That is, we have for i even $\mathcal{M}^I, v_i \vDash B_1\varphi$ iff $\mathcal{M}^I, v_i \vDash \varphi$ and $\mathcal{M}^I, v_i \vDash B_2\varphi$ iff $\mathcal{M}^I, v_{i+1} \vDash \varphi$ and vice versa for i odd. Now, define $\chi = K_3(K_1K_3p \vee K_2K_3p)$, then it is not difficult to see that $\mathcal{M}^I, v_i \vDash \chi$ iff $i \in I$. Combining these insights, we get that

$$\mathcal{M}^I, v_1 \vDash \varphi_n \text{ iff } n \in I$$

where φ_n is as defined above. Thus, all φ_n are independent, this finishes our proof. □

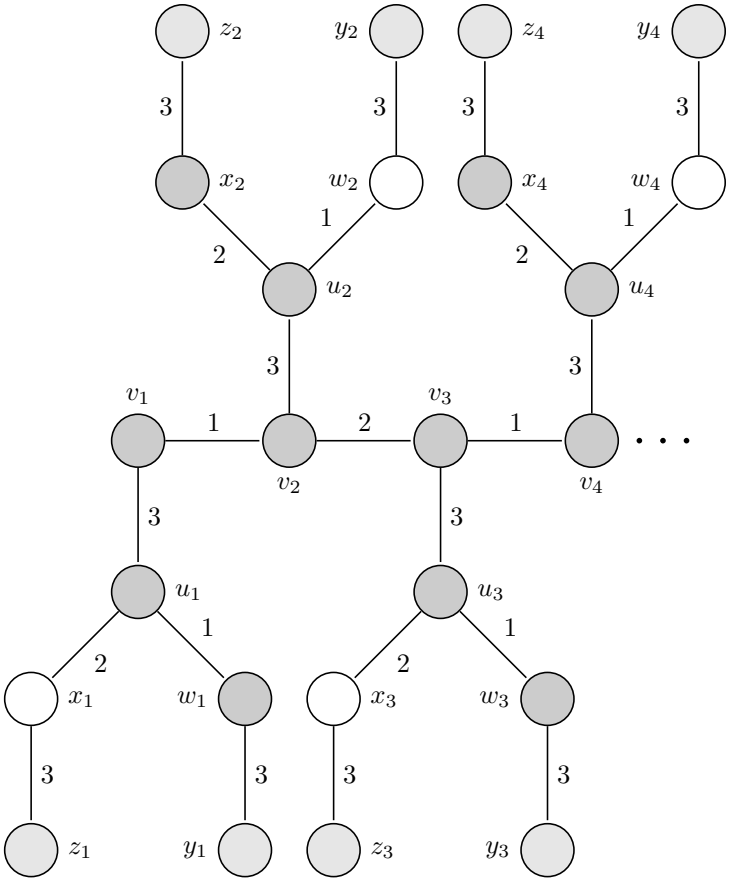


Table 3.1: The Kripke frame $\mathcal{F}$

Proof of Theorem 3.19. Recall that any type in $T \in \mathcal{T}_c$ for $c \in \{\mathcal{L}_{K,\wedge}; \mathcal{L}_{\vee 2}\}$ is already determined by $T \cap \mathcal{L}_K$. Further, the proof of Lemma 3.11 shows that every T in $\mathcal{T}_{\mathcal{L}_L}$ is realizable in a finite model iff every $\mathcal{T}' \in \mathcal{T}_{\mathcal{L}_K}$ is. Finally, note that $\mathcal{L}_K \subseteq \mathcal{L}_D$, thus it suffices to show the claim for $\mathcal{L}_D$.

The proof will use some of the constructions from chapter 2. Let $T \in \mathcal{T}_D$ be the type we want to realize and let $\mathcal{N}, v$ be some (not necessarily finite) model realizing T. By the proof of Lemma 3.15, T is characterized by the finite set M of minimal elements of $\mathcal{L}_D \setminus T$. Denote the maximal length of any of the formulas in M, that is the number of D_I in that formula, by r and let $k = r+2$. By part *ii*) of Observation 2.8, there is some ordinal $\gamma \geq \omega$ such that the map $r : \mathcal{N} \to \mathcal{F}_\gamma(S)$ defined in Definition 2.7 is a bisimulation onto its image, where $S = \{\emptyset, \{x\}\}$. By a slight abuse of notation, we will identify $\mathbb{N}, v$ with its image under r, that is we will assume that $\mathbb{N}, v \subset \mathcal{F}_\gamma(S)$.

Let $\pi_k : \mathcal{F}_\gamma(S) \to \mathcal{F}_k(S)$ be the projection sending every type $t \in \mathcal{F}_\gamma$ to its initial segment of length k. Then define $\mathcal{M} := \{\pi_k(x) | x \in \mathcal{N}\}$ and let $w := \pi_k(v)$. Since $\mathcal{M} \subseteq \mathcal{F}_k(S)$, we can turn $\mathcal{M}$ into a Kripke Model with the equivalence relation induced by $\mathcal{F}_k(S)$. We claim that $\mathcal{M}, w$ realizes T. First, we show that $\mathcal{M}, w \nvDash \varphi$ whenever $\mathcal{N}, v \nvDash \varphi$ To prove this, first note that for all $\varphi \in \mathcal{L}_D$ of quantifier depth at most $k-1$ we have that $\mathcal{M}, w \vDash \varphi \Leftrightarrow \mathcal{N}, v \vDash \varphi$. This follows by an induction over the quantifier depth of φ, see [54, Lemma 2.5.] for details. In particular, this implies $\mathcal{M}, v \nvDash \varphi$ for all $\varphi \in M$ and thus also $\mathcal{M}, v \nvDash \psi$ for all $\psi \in \mathcal{L}_D \setminus T$, since for every such ψ there is a $\varphi \in M$ with $\psi \to \varphi$.

Next, we show the converse direction, $\mathcal{N}, v \vDash \varphi \Rightarrow \mathcal{M}, w \vDash \varphi$. By the definition of the Kripke structures $\mathcal{F}_\kappa$ the map π_k is a functional simulation, that is $\mathbf{f} \sim_i \mathbf{g}$ for $\mathbf{f}, \mathbf{g} \in \mathcal{F}_\gamma(S)$ implies that $\pi_k(\mathbf{f}) \sim_i^k \pi_k(\mathbf{g})$, where $\sim_i$ and $\sim_i^k$ are the equivalence relations for agent i in $\mathcal{F}_\gamma(S)$ and $\mathcal{F}_k(S)$ respectively. Thus, since $\mathcal{L}_D$ only contains instances of positive knowledge, we get $\mathcal{N}, v \vDash \varphi \to \mathcal{M}, w \vDash \varphi$ for all $\varphi \in L_D$, thus finishing our proof. □

Remark: Alternatively we could have given a constructive proof of Theorem 3.19, giving an explicit construction of a model of $\mathcal{M}, w$ realizing T. This construction is lengthy and slightly tedious, but basically straightforward.

Proof of Theorem 3.20. Let L_1, L_2 be as stated in the theorem and let $\mathcal{M}, w$ be a finite Kripke Model realizing L_1. To find the desired event model, we use a characterization result from [158]: For two multi-agent S5-models $\mathcal{M}', s$ and

$\mathcal{M}'', t$ there is an event model $\mathcal{E}, e$ such that

$$\mathcal{M}', s \oplus \mathcal{E}, e \leftrightarrows \mathcal{M}'', t$$

if and only if there is a total simulation from $\mathcal{M}'', t$ to $\mathcal{M}', s$.

Thus, it suffices to construct a model $\mathcal{N}, v$ that realizes level L_2 such that there is a total simulation from $\mathcal{N}, v$ to $\mathcal{M}, w$. Note again that L_1 and L_2 are completely determined by $M(L_1)$ and $M(L_2)$ respectively and let $k = max_{x \in M(L_1)}(qd(x)) + max_{x \in M(L_2)}(qd(x)) + 3$, where $qd(x)$ stands for the quantifier depth of x. To begin our construction of $\mathcal{N}, v$, we construct a finite tree $\mathcal{T}$ of height k with root v. In this tree, every node will be labeled with a world from $\mathcal{M}$ and every edge will be labeled with one of the agents. We define this tree inductively as follows:

- The root v is labelled with w
- To construct the first level do the following: For every pair (i, x) where i is an agent and $x \in \mathcal{M}$ with wR_ix (where R_i is the equivalence relation for agent i in $\mathcal{M}$), add a new node and label it with x. Further, add an edge between this new node and v and label it with i.
- Assume the l-th level has been constructed for $1 \leq l < k$. For every vertex v in the l-th level do the following: Let i_v be the label of the edge connecting e to some edge in the $l-1$st level and let m_v be the label of v. For every pair (j, x) where j is an agent and x a world with xR_jm_v add a new node and a new edge from that node to v. Label the node with x and the edge with j.

In the following we write $l(v)$ for the label of some node $v \in T$. We turn this tree $\mathcal{T}$ into a Kripke Model $\mathcal{N}$ in the following way: The accessibility relations $R_i^{\mathcal{N}}$ are the equivalence relations generated by the edge labeling on $\mathcal{T}$. The valuation $V^{\mathcal{N}}$ on $\mathcal{N}$ is generated by the valuation $V^{\mathcal{M}}$ of $\mathcal{M}$ by the formula $v \in V^{\mathcal{N}}(x) :\Leftrightarrow l(v) \in V^{\mathcal{M}}(x)$. Now it is not difficult to see that the labeling on $\mathcal{N}$, i.e., the map $l : \mathcal{N} \to \mathcal{M}$ sending every $v \in \mathcal{N}$ to $l(v)$ is a functional and hence total simulation from $(\mathcal{N}, v)$ to $(\mathcal{M}, w)$. First, we note that $(\mathcal{N}, v)$ also realizes L_1, the level of knowledge of $\mathcal{M}, w$.

To see this, first note that since l is a simulation, we have that $\mathcal{N}, v \nVdash \varphi$ implies $\mathcal{M}, w \nVdash \varphi$ for all $\varphi \in \mathcal{L}_K$. For the converse direction, we introduce the following notation: For $\varphi \in \mathcal{L}_K$ of the form $\varphi = K_{i_1} \dots K_{i_r}x$ let $\bar{\varphi} = L_{i_1} \dots L_{i_r}\neg x$, thus $\mathcal{M}, w \nVdash \varphi \Leftrightarrow \mathcal{M}, w \vDash \bar{\varphi}$. Note that for all $\varphi \in M(L_1)$, with $\varphi = K_{i_1} \dots K_{i_r}x$, there is some sequence $wR_{i_1}w_1R_{i_2} \dots R_{i_r}w_r$ in $\mathcal{M}$ witnessing

that $\mathcal{M}, w \vDash \bar{\varphi}$. By the choice of k, the height of our tree, we have $r < k$, thus there is a corresponding sequence $vR_{i_1}v_1R_{i_2}\dots R_{i_r}v_r$ with $l(v_i) = w_i$ in $\mathcal{N}$ witnessing that $\mathcal{N}, v \vDash \bar{\varphi}$. Therefore, $\mathcal{M}, w \nvDash \varphi$ implies that $\mathcal{N}, v \nvDash \varphi$ for all $\varphi \in M(L_1)$ and thus also for all $\varphi \in \mathcal{L}_K$.

Next, we thin out the tree $\mathcal{T}$ and thus the model $\mathcal{N}$ to a submodel $\mathcal{N}'$ such that $(\mathcal{N}', v)$ realizes L_2. Before we do so, note that after removing any set of nodes from $\mathcal{N}$, the labeling function l still is a functional simulation on whatever remains of $\mathcal{N}$. Thus, we only have to ensure that our thinning out is performed in such a way that $(\mathcal{N}', v)$ realizes L_2. To do so let Ψ be the set of all $\varphi \in L_2$ of length at most k. For each $\psi \in \Psi$ of the form $\psi = K_{i_1}\dots K_{i_r}x$ we do the following: If $\mathcal{N}, v \nvDash \psi$ there are some witnesses for $\bar{\psi} \in \mathcal{N}, v$, i.e., sequences of the form $vR_{i_1}v_1R_{i_2}\dots R_{i_r}v_r$ with $x \notin V^{\mathcal{N}}(x)$. We remove all nodes $v \in \mathcal{T}$ that appear as a last node in any such sequence for any $\psi \in \Psi$ and call the resulting graph $\mathcal{T}^{\#}$. We turn $\mathcal{T}^{\#}$ into a tree by removing all connected components that do not contain v. We call the resulting tree $\mathcal{T}'$ and the corresponding induced model $\mathcal{N}'$. Thus, we have that $\mathcal{N}', v \vDash \psi$ for every $\psi \in \Psi$. Next, we show that $\mathcal{N}', v \vDash \varphi$ for all $\varphi \in L_2$. Assume not and assume that ψ is a counterexample. Since we have shown this claim already for Ψ, we can assume that the quantifier depth of ψ is $r > k$. Let $vR_{i_1}v_1R_{i_2}\dots R_{i_r}v_r$ be a witness that $\mathcal{N}', v \vDash \bar{\varphi}$. Since $r > k$ and since $\mathcal{T}$ is a tree of height k, there are some $l < j$ with $v_l = v_j$ such that $vR_{i_1}v_1\dots R_{i_l}v_lR_{i_j}\dots R_{i_r}v_r$ (i.e., the string $v_{l+1}R_{l+1}\dots R_j$ has been removed) has length $r' \leq k$. This shortened sequence is a witness of the fact that $\mathcal{N}', v \nvDash \varphi'$, where $\varphi' = K_{i_1}\dots K_{i_l}K_{i_{j+1}}\dots K_{i_r}$, yet $\varphi' \in L_2$ since $\varphi' \preceq \varphi$. But this is an contradiction, since $\varphi' \in \Psi$ and we have already shown the claim for Ψ.

Last, we show that $\mathcal{N}', v \nvDash \varphi$ for every $\varphi \in \mathcal{L}_K \setminus L_2$. Again, it suffices to show this for $\varphi \in M(L_2)$. Let thus $\varphi \in MLT_2)$, say φ of the form $K_{j_1}\dots K_{j_s}x$. By our assumption we can pick some $\psi \in M(T_1)$ of the form $K_{i_1}\dots K_{i_r}x$ with $\psi \preceq \varphi$ and $K_{i_r} = K_{j_s}$. Thus, by the definition of the order $\preceq$, there is an increasing function $f\{1\dots r\} \to \{1,\dots s\}$ witnessing that there is an embedding form ψ to φ, that is $K_{i_r} = K_{j_{f(r)}}$. Since $K_{i_r} = K_{j_s}$, we can also assume that $f(r) = s$. Next, we define an adjoint function $\bar{f} : f\{1\dots s\} \to \{1,\dots r\}$ sending every j to $max\{i \leq r || f(i) \leq s\}$. By construction, there is some path $p = vR_{i_1}v_1R_{i_2}\dots R_{i_s}v_s$ in $\mathcal{N}$ witnessing that $\psi \notin L_1$, that is $v_r \notin v^{\mathcal{N}}(x)$. Since $\psi \in M(L_1)$ we also have $v_i \in V^{\mathcal{N}}(x)$ for $i \leq r$. Now, let $p' = vR_{j_1}y_1R_{i_2}\dots R_{j_s}y_s$ be the sequence defined by $y_i = v_{\bar{f}(i)}$. We claim that p' is a path in $\mathcal{T}'$. First, we show the weaker claim that p' is in $\mathcal{T}$. To see this, observe that p' was

constructed by p by inserting some reflexive arrows into p, that is every segment $y_{j-1}R_{i_j}y_j$ of p' either satisfies $y_{j-1} = y_j$ or it is of the form $v_{\bar{f}(j)-1}R_{\bar{f}(j)}v_{\bar{f}(j)}$. In either case the associated labels $l(y_{j-1})$ and $l(y_j)$ satisfy $l(y_{j-1})R_{i_j}l(y_j)$ in $\mathcal{M}$ and thus $p' \in \mathcal{T}$. To see that p' is also in $\mathcal{T}'$, observe that $y_{s-1} = v_{r-1}$ by our construction of $\bar{f}$, thus $y_i \in V^{\mathcal{N}}(x)$ for all $i < s$, which implies that none of these y_i got removed in the transition from $\mathcal{T}$ to $\mathcal{T}'$. Since $\varphi \in M(L_2)$, also y_r did not get removed, thus $p' \in \mathcal{T}'$. Hence, $\mathcal{N}', v \vDash \bar{\varphi}$ which finishes our proof.

□

Chapter 4

Modeling Individual Expertise in Group Judgments

4.1 Introduction

Groups frequently make judgments that are based on aggregating the opinions of its individual members. A panel of market analysts at Apple or Samsung may estimate the expected number of sales of a newly developed cell phone. A group of conservation biologists may assess the population size of a particular species in a specific habitat. A research group at the European Central Bank may evaluate the merits of a particular monetary policy. Generally, such problems occur in any context where groups have to combine various opinions into a single group judgment [for a review paper, see 40].

Even in cases of fully shared information, the assessment of the evidence will generally vary among the agents and depend on factors such as professional training, familiarity with similar situations in the past, and personal attitude toward the results. Thus, it will not come as a surprise that the individual judgments may differ. But how shall they be aggregated?

Often, some group members are more competent than others. Recognizing these *experts* may then become a crucial issue for improving group performance. Research in social psychology and management science has investigated the ability of humans to properly assess the expertise of other group members in such contexts [26, 40, 100]. Most of this research stresses that recognizing experts is no easy task: perceived and actual expertise need not agree, data are noisy,

This chapter is based on joint work with J. Sprenger. It is an extended version of [93].

questions may be too hard, and expertise differences may be too small to be relevant [e.g., 113]. This motivates a comparison of two strategies for group judgments: (i) deferring to the agent who is perceived as most competent, and (ii) taking the straight average of the estimates [78, 146]. The overall outcomes suggest that the straight average is often surprisingly reliable, apparently being one of those "fast and frugal heuristics" [66] that help boundedly rational agents to make cost-effective decisions.

On the other hand, even if not explicitly recognized as such, experts tend to exert greater influence on group judgments than non-experts [26]. This motivates a principled epistemic analysis of the potential benefits of expertise-informed group judgments. We characterize conditions under which differentially weighted averages, fed by incomplete and perhaps distorted information on individual expertise, ameliorate group performance, compared to a straight average of the individual judgments. Our paper approaches this question from an analytical perspective, that is, with the help of a statistical model. We follow the social permutation approach [e.g., 24] and model the agents as unique entities with different abilities. This differs notably from more traditional social combination research where individual agents are modeled as interchangeable [e.g., 44]. Our main result – that individual expertise makes a robust contribution to group performance – is not without surprise, given the generality of our conditions that also allow for perturbations such as individual bias or correlations among the group members. Therefore, our analytical results provide theoretical support to research on the recognition of experts in groups [e.g., 14], and they directly relate to empirical comparisons of differentially weighted group judgments to "composite judgments", such as the group mean or median [25, 49, 81, 108].

Our work is also related to two other research streams. First, there is a thriving epistemological literature on peer disagreement and rational consensus, where consensus is mostly reached by deference to (perceived) experts. However, this debate either focuses on social power and mutual respect relations [e.g., 103], or on principled philosophical questions about resolving disagreement [e.g., 50]. By means of a performance-focused mathematical model, we hope to bring this literature close to its primary target: the truth-tracking abilities of various epistemic strategies. There is also a vast literature on group decisions preference and judgment aggregation [e.g., 112], but two crucial features of our inquiry—the aggregation of numerical values and the particular role of experts—do not play a major role in there.

Second, there is a fast increasing body of literature on expert judgment and forecasting, which has emerged from applied mathematics and statistics and became a flourishing interdisciplinary field. This strand of research deals with the theoretical modeling of expert judgment, most notably the (Bayesian) reconciliation of probability distributions [111], but it also includes more practical questions such as comparison of calibration methods, choice of seed variables, analyses of the use of expert judgment in the past [42], and the study of general forecasting principles, such as the benefits of opinion diversity [7, 126]. We differ from that approach in pooling individual (frequentist) estimators instead of subjective probability distributions, but we study similar phenomena, such as the impact of in-group correlations.

Admittedly, our baseline model is very simple, but due to this simplicity, we are able to prove a number of results regarding the behavior of differentially weighted estimates under correlation, bias and benchmark uncertainty. Here, our paper builds on analytical work in the forecasting and social psychology literature [13, 83], following the approach of Einhorn et al. [49].

The rest of the paper is structured as follows: we begin with explaining the model and stating conditions where differentially weighted estimates outperform the straight average (Sect. 4.2). In the sequel, we show that this relation is often preserved even if bias or mutual correlations are introduced (Sect. 4.3 and 4.4). Subsequently, we assess the impacts of over- and underconfidence (Sect. 4.5). Finally, we discuss our findings and wrap up our conclusions (Sect. 4.6).

4.2 The Model and Baseline Results

Our problem is to find a good estimate of an unknown quantity μ. For reasons of convenience, we assume without loss of generality that $\mu = 0$.[1]

We model the group members' individual estimates X_i, $i \leq n$, as independent random variables that scatter around the true value $\mu = 0$ with variance σ_i^2. The X_i are *unbiased* estimators of μ, that is, they have the property $\mathbb{E}[X_i] = \mu$. This baseline model is inspired by the idea that the agents try to approach the true value with a higher or lower degree of precision, but have no systematic bias in either direction. The competence of an agent is explicated as the degree of precision in estimating the true value. No further assumptions on the distributions of the X_i are made—only the first and second moments are fixed.

[1] Rewriting our results for the general case $\mu \neq 0$ is just a matter of affine transformation, but comes with some notational baggage. Therefore we focus without loss of generality on $\mu = 0$.

To illustrate our assumptions further, it might be instructive to compare our approach to the famous James-Stein estimator. This estimator, presented by James and Stein in their seminal 1961 paper [86], is a powerful classical example for how differential weights can improve upon the quality of an estimator under quite general conditions. James and Stein treat the case of estimating several, at least three, parameters p_i at once. The only information given about these parameters is a single draw w_i from a normal distribution centered around each parameter. These normal distributions are all assumed to share the same variance. Notably, all p_i are independent of each other, thus one could naturally assume $\langle w_1, \ldots, w_n \rangle$ to be the best estimator for the vector $\langle p_1, \ldots, p_n \rangle$. Surprisingly, James and Stein showed that this is not true. There is some weighted average, the James-Stein estimator, that *always* outperforms $\langle w_1, \ldots, w_n \rangle$ as an estimator in terms of mean square error. In its basic setting, the James-Stein estimator shares some properties with ours. Also we deal with a vector of random draws, each taken from a different distribution. However, all these distributions are centered around the same parameter p. That is, they all share the same mean. We do, on the other hand, allow our distributions to differ in their variances, mirroring the varying degrees of expertise of the individual agents. In a certain sense, the settings of both approaches are orthogonal to each other. While the James-Stein estimator deals with a single judgment on many different, at least three, variables at once, our case consists of an entire group of judgments about a single variable. That is, both approaches need to make different types of assumptions on the underlying estimator. As described above, the James-Stein estimator assumes all distributions to share the same variance, while in our case the distributions all need to be centered around the same mean. Further, the weights in our approach only depend on the variances of the individual estimators, not on the outcomes of the random draw, and we allow for any type of input distributions. In the James-Stein case, on the other hand, the distributions all need to be normal and the weights depend upon the outcomes of the random draw. Of course, our weaker assumptions come at a price. While the James-Stein estimator is *guaranteed* to outperfom the initial observation vector *under any conditions*, we merely aim to identify a (broad) range of conditions under which differential weighting fares better than straight averaging.

In our model, the question of whether the recognition of individual expertise is epistemically advantageous translates into the question of which convex combination of the X_i, $\hat{\mu} := \sum_{i=1}^{n} c_i X_i$, outperforms the straight average

$\bar{\mu} := \frac{1}{n}\sum_{i=1}^{n} X_i$. Standardly, the quality of an estimate is assessed by its mean square error (MSE) which can be calculated as

$$\begin{aligned} \mathrm{MSE}(\hat{\mu}) := , \mathcal{E}[(\hat{\mu}-\mu)^2] &= \mathcal{E}\left[\left(\sum_{i=1}^{n} c_i X_i\right)^2\right] \\ &= \sum_{i=1}^{n} c_i^2\, \mathcal{E}\left[X_i^2\right] + \sum_{i=1}^{n}\sum_{j\neq i} c_i c_j\, \mathcal{E}[X_i]\,\mathcal{E}[X_j] \\ &= \sum_{i=1}^{n} c_i^2\, \sigma_i^2 \end{aligned}$$

which is minimized by the following assignment of the c_i [cf. 103, 139]:

$$c_i^* = \left(\sum_{j=1}^{n} \frac{\sigma_i^2}{\sigma_j^2}\right)^{-1} . \tag{4.1}$$

Thus, naming the c_i^* as the "optimal weights" is motivated by two independent theoretical reasons:

1. As argued above, for independent and unbiased estimates X_i with variance σ_i^2, mean square error of the overall estimate is minimized by the convex combination $X = \sum_i c_i^* X_i$. Thus, for a standard loss function, the c_i^* are indeed the optimal weights.

2. Even when the square loss function is replaced by a more realistic alternative [76], the c_i^* can still define the optimal convex combination of individual estimates. In that case, we require stronger distributional assumptions.[2]

The problem with these optimal weights is that each agent's individual expertise would have to be known in order to calculate them. Given all the biases that actual deliberation is loaded with, e.g., ascription of expertise due to professional reputation, age or gender, or bandwagon effects, it is unlikely that the agents succeed at unraveling the expertise of all other group members [cf. 7, 121].

Therefore, we widen the scope of our inquiry:

> **Question:** Under which conditions will differentially weighted group judgments outperform the straight average?

A first answer is given by the following result where the differential weights preserve the expertise ranking:

[2] Hartmann and Sprenger [76] prove the optimality of the c_i^* for the case of normally distributed independent and unbiased estimates with variance σ_i^2 and the loss function family $L_\alpha(x) = 1 - \exp(-x^2/2\alpha^2)$. That paper also contains an elaborate justification for choosing this family of loss functions.

Theorem 4.1 (First Baseline Result)**.** *Let $c_1, \ldots, c_n > 0$ be the weights of the individual group members, that is, $\sum_{i=1}^n c_i = 1$. Without loss of generality, let $c_1 \leq \ldots \leq c_n$. Further assume that for all $i > j$:*

$$1 \leq \frac{c_i}{c_j} \leq \frac{c_i^*}{c_j^*} \tag{4.2}$$

Then the differentially weighted estimator $\hat{\mu} := \sum_{i=1}^n c_i X_i$ outperforms the straight average. That is, MSE($\hat{\mu}$) $\leq$ MSE($\bar{\mu}$), with equality if and only if $c_i = 1/n$ for all $1 \leq i \leq n$.

This result demonstrates that relative accuracy, as measured by pairwise expertise ratios, is a good guiding principle for group judgments as long as the relative weights are not too extreme.

The following result extends this finding to a case where the benefits of differential weighting are harder to anticipate: we allow the c_i to lie in the entire $[1/n, c_i^*]$ (or $[c_i^*, 1/n]$) interval, allowing for cases where the ranking of the group members is not represented correctly. One might conjecture that this phenomenon adversely affects performance, but this is not the case:

Theorem 4.2 (Second Baseline Result)**.** *Let $c_1 \ldots c_n \in [0, 1]$ such that $\sum_{i=1}^n c_i = 1$. In addition, let $c_i \in [\frac{1}{n}; c_i^*]$ respectively $c_i \in [c_i^*; \frac{1}{n}]$ hold for all $1 \leq i \leq n$. Then the differentially weighted estimator $\hat{\mu} := \sum_{i=1}^n c_i X_i$ outperforms the straight average. That is,* MSE($\hat{\mu}$) $\leq$ MSE($\bar{\mu}$)*, with equality if and only if $c_i = 1/n$ for all $1 \leq i \leq n$.*

Note that none of the baseline results implies the other one. The conditions of the second result can be satisfied even when the ranking of the group members differs from their actual expertise, and a violation of the second condition (e.g., $c_i^* = 1/n$ and $c_i = 1/n + \varepsilon$) is compatible with satisfaction of the first condition. So the two results are really complementary.

We have thus shown that differential weighting outperforms straight averaging under quite general constraints on the individual weights, motivating the efforts to recognize experts in practice. The next sections extend these results to the presence of correlation and bias, thereby transferring them to more realistic circumstances.

4.3 Biased Agents

The first extension of our model concerns *biased* estimates X_i, that is, estimates that do not center around the true value $\mu = 0$, but around $B_i \neq 0$. We still

assume that agents are honestly interested in getting close to the truth, but that training, experience, risk attitude or personality structure bias their estimates into a certain direction. For example, in assessing the impact of industrial development on a natural habitat, an environmentalist will usually come up with an estimate that significantly differs from the estimate submitted by an employee of an involved corporation—even if both are intellectually honest and share the same information.

For a biased agent i, the competence/precision parameter σ_i^2 has to be re-interpreted: it should be understood as the *coherence* (or non-randomness) of the agent's estimates instead of the accuracy. This value is indicative of accuracy only if the bias B_i is relatively small.

Under these circumstances, we can identify an intuitive sufficient condition for differential weighting to outperform straight averaging.

Theorem 4.3. *Let $X_1, \ldots, X_n$ be random variables with bias $B_1, \ldots, B_n$.*

(a) Suppose that the c_i in the estimator $\hat{\mu} = \sum_{i=1}^n c_i X_i$ satisfy one of the conditions of the baseline results (i.e., either $1 \leq c_i/c_j \leq c_i^*/c_j^*$ or $c_i \in [1/n, c_i^*]$ respectively $c_i \in [c_i^*, 1/n]$). In addition, let the following inequality hold:

$$\left(\sum_{i=1}^n c_i B_i\right)^2 < \left(\sum_{i=1}^n \frac{1}{n} B_i\right)^2 \tag{4.3}$$

Then differential weighting outperforms straight averaging, i.e.,f $\mathrm{MSE}(\hat{\mu}) < \mathrm{MSE}(\bar{\mu})$.

(b) Suppose the following inequality holds:

$$\left(\sum_{i=1}^n c_i B_i\right)^2 > \left(\sum_{i=1}^n \frac{1}{n} B_i\right)^2 + \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 \tag{4.4}$$

Then differential weighting does worse than straight averaging, that is, $\mathrm{MSE}(\hat{\mu}) > \mathrm{MSE}(\bar{\mu})$.

Intuitively, condition (4.3) states that the differentially weighted bias is smaller or equal than the average bias. As one would expect, this property favorably affects the performance of the differentially weighted estimator. Condition (4.4) states, on the other hand, that if the difference between the mean square biases of the weighted and the straight average exceeds the mean variance of the agents, then straight averaging performs better than weighted averaging.

When the group size grows to a very large number, both parts of Theorem 4.3 collapse into a single condition, as long as the biases and variances are both

bounded. This is quite obvious since the last term of (4.4) is of the order $\mathcal{O}(1/n)$. Theorem 4.3 applies in particular in the case where agents are biased into the same direction and less biased agents make more coherent estimates (that is, with smaller variance):

Corollary 4.4. *Let $X_1, \ldots, X_n$, be random variables with bias $B_1, \ldots, B_n \geq 0$ such that $c_i \geq c_j$ implies $B_i \geq B_j$ (or vice versa for $B_1, \ldots, B_n \leq 0$). Then, with the same definitions as above:*

- $\mathrm{MSE}(\bar{\mu}) \geq \mathrm{MSE}(\hat{\mu})$.
- *If there is a uniform group bias, that is, $B := B_1 = \ldots = B_n$, then $\mathrm{MSE}(\bar{\mu}) - \mathrm{MSE}(\hat{\mu})$ is independent of B.*

So even if all agents have followed the same training, or have been raised in the same ideological framework, expertise recognition does not multiply that bias, but helps to increase the accuracy of the group's judgment. In particular, if there is a uniform bias in the group, the relative advantage of differential weighting is independent of the size of the bias. All in all, these results demonstrate the importance of expertise recognition even in groups where the members share a joint bias—a finding that is especially relevant for practice.

4.4 Independence Violations

We turn to violations of independence between the group members. Consider first the following fact that compares two groups with different degrees of correlation:

Fact 4.5. *If $0 \leq \mathcal{E}[X_i X_j] \leq \mathcal{E}[Y_i Y_j]$ $\forall i \neq j \leq n$ and $\mathcal{E}[X_i^2] = \mathcal{E}[X_j^2]$, then both straight averaging and weighted averaging on X_i yield a lower mean square error than the same procedures applied to Y_i.*

Fact 4.5 shows that less correlated groups perform better, ceteris paribus. For practical purposes, this suggests that heterogeneity of a group is an epistemic virtue since strong correlations between the agents are less likely to occur, making the overall result more accurate [cf. 126].

Regarding the comparison of straight and weighted averaging, we can show the following result:

Theorem 4.6. *Let $X_1, \ldots, X_n$ be unbiased estimators, that is, $\mathcal{E}[X_i] = \mu = 0$, and let the c_i satisfy the conditions of one of the baseline results, with $\hat{\mu}$ defined*

as before. Let $I \subseteq \{1, \ldots, n\}$ *be a subset of the group members with the property*

$$\forall i, j \in I : c_i \geq c_j \;\Rightarrow\; \forall k \in I, k \neq i, j : \mathcal{E}[X_j X_k] \geq \mathcal{E}[X_i X_k] \geq 0. \tag{4.5}$$

(*i*) **Correlation vs. Expertise** *If* $I = \{1, \ldots, n\}$*, then weighted averaging outperforms straight averaging, that is,* $\mathrm{MSE}(\hat{\mu}) \leq \mathrm{MSE}(\bar{\mu})$.

(*ii*) **Correlated Subgroup** *Assume that* $\mathcal{E}[X_i X_j] = 0$ *if* $i \notin I$ *or* $j \notin I$*, and that*

$$\frac{1}{|I|} \sum_{i \in I} c_i \leq \frac{1}{n} \sum_{i=1}^{n} c_i. \tag{4.6}$$

Then weighted averaging still outperforms straight averaging, that is, $\mathrm{MSE}(\hat{\mu}) \leq \mathrm{MSE}(\bar{\mu})$.

To fully understand this theorem, we have to clarify the meaning of condition (4.5). Basically, it expresses that an expert i is less correlated with any given group member k than a non-experts j.[3]

Once we have understood this condition, the rest is straightforward. Part (i) states that if I equals the entire group, then differential weighting has an edge over averaging. That is, the benefits of expertise recognition are not offset by the perturbations that mutual dependencies may introduce. Arguably, the generality of the result is surprising since condition (4.5) is quite weak. Part (ii) states that differential weighting is also superior whenever there is no correlation with the rest of the group, and as long as the average competence in the subgroup is lower than the overall average competence (see equation (4.6)).

It is a popular opinion [e.g., 150] that correlation of individual judgments is one of the greatest dangers for relying on experts in a group. To some extent, this opinion is vindicated by Fact 4.5 in our model. However, expertise-informed group judgments may still be superior to composite judgments, as demonstrated by Theorem 4.6. The interplay of correlation and expertise is subtle and not amenbale to broad-brush generalizations.

4.5 Over- and Underconfidence

We now consider a specific family of c_i's in order to study how group members' self-assessment in terms of quality affects group performance as a whole, modeled again as unbiased estimates X_i with variance σ_i^2.

[3]Recall that $\mathcal{E}[X_i, X_k] \leq \mathcal{E}[X_j, X_k]$ can be rewritten as $\sigma_i/\sigma_j \leq \rho_{jk}/\rho_{ik}$ with ρ_{ij} defined as the Pearson correlation coefficient $\rho_{ij} := \mathcal{E}[X_i X_j]/\sigma_i \sigma_j$. Also, if $c_i \geq c_j$ then automatically $\sigma_i \leq \sigma_j$.

Suppose that the group members have some idea of their own competence. That is, they are able to position themselves in relation to a commonly known *benchmark*: they are able to assess how much better or worse they expect themselves to perform compared to a default agent, modeled as a unbiased random variable with variance s^2. Such a scenario may be plausible when agents have a track record of their performance, or obtain performance feedback. The agents then express how much weight they should ideally get in a group of $n-1$ default agents. Using equation 4.1, this ideal weight c_i in a group of benchmark agents (i.e. $\sigma_j^2 = s^2$ for all $j \neq i$) is given by

$$c_i = \left(1 + \sum_{j \neq i} \frac{\sigma_i^2}{\sigma_j^2}\right)^{-1} = \frac{s^2}{s^2 + (n-1)\sigma_i^2}. \tag{4.7}$$

Assume further that every agent uses the same benchmark, that these weights also determine to what extent a group member compromises his or her own position, and that decision-making takes place on the basis of the normalized c_i. It can then be shown (proof omitted) that the differentially weighted estimator $\hat{\mu}$ defined by equation (4.7) outperforms straight averaging—in fact, this is entailed by the Second Baseline Result (Theorem 4.2).

Here, we want to study how over- and underestimating the competence of a "default agent" will affect group performance. Is it always epistemically detrimental when the agents misguess the group competence?

The answer is, perhaps surprisingly, no. To explain this result, we first observe that the less confidence we have in the group (= s^2 is large), the more does the weighted average resemble the straight average. Recalling equation (4.7), we note that all c_i will be very close to 1. This implies that the expertise-informed average will roughly behave like the straight average.

Conversely, if the group is perceived as competent (=small value of s), then the c_i will typically *not* be close to 1 such that differential weights will diverge significantly from the straight average. This intuitive insight leads to the following theorem:

Theorem 4.7. *Let $\hat{\mu}_{s^2}$ and $\hat{\mu}_{\tilde{s}^2}$ be two weighted expertise-informed estimates of μ, defined according to equation* (4.7) *with benchmarks s^2 and $\tilde{s}^2$, respectively. Then MSE($\hat{\mu}_{s^2}$) $\leq$ MSE($\hat{\mu}_{\tilde{s}^2}$) if and only if $s^2 \leq \tilde{s}^2$.*

It can also be shown (proof omitted) that this procedure approximates the *optimal* weights c_i^* if the perceived group competence approaches perfection, that is, $s \to 0$. In other words, as long as the group members judge themselves

accurately, optimism with regard to the abilities of the other group members is epistemically favorable. On the other hand, overconfidence in one's own abilities relative to the group typically deteriorates performance.

4.6 Discussion

We have set up an estimation model of group decision-making in order to study the effects of individual expertise on the quality of a group judgment. We have shown that, in general, taking into account relative accuracy positively affects the epistemic performance of groups. Translated into our statistical model, this means that differential weighting outperforms straight averaging, even if the ranking of the experts is not represented accurately.

The result remains stable over several representative extensions of the model, such as various forms of bias, violations of independence, and over- and underconfident agents (Theorems 4.3–4.7). In particular, we demonstrated that differential weighting is superior (i) if experts are, on average, less biased; (ii) for a group of uniformly biased agents; (iii) if experts are less correlated with the rest of the group than other members. We also showed that uniform overconfidence in one's own abilities is detrimental for group performance whereas (over)confidence in the group may be beneficial. These properties may be surprising and demonstrate the stability and robustness of expertise-informed judgments, implying that the benefits of recognizing experts may offset the practical problems linked with that process.

Our model can in principle also be used for describing how groups actually form judgments. In that case, the involved tasks should neither be too intellective (that is, there is a *demonstrable* solution) or too judgmental [101]: in highly intellective tasks, group will typically not perform better than the best individual (=the one who has solved the task correctly). This differs from our model where any agent has only partial knowledge of the truth. On the other hand, if the task is too judgmental, any epistemic component will be removed and the individual weights may actually be based on the *centrality* of a judgment, such as in Hinsz's (1999) SDS-Q scheme.

Finally, we name some distinctive traits of our model. First, unlike other models of group judgments that are detached from the group members' individual abilities [44, 47, 82, 103], it is a genuinely epistemic model, evaluating

the performance of different ways of making a group judgment.[4] Thus, our model can be used normatively, for supporting the use of differential weights in group decisions, but also descriptively, for fitting the results of group decision processes.

Second, we did not make any specific distributional assumptions on how the agents estimate the target value. Our assumptions merely concern the first and second moment (bias and variance). We consider this parsimony a prudent choice because those distributions will greatly vary in practice, and we do not have epistemic access to them. Classical work in the social combination literature makes much more specific distributional assumptions (e.g., the multinomial distributions in Thomas and Fink 1961 and Davis 1973), restricting the scope of that analysis.

Third, we are not aware of other analytical models that take into account important confounders such as correlation, bias and over-/underconfident agents. Thus, we conclude that our model makes a substantial contribution to understanding the epistemic benefits of expertise in group judgments.

4.7 Appendix: Proofs

We will need the following inequalities repeatedly in the subsequent proofs. Let $c_1, \ldots, c_n > 0$. Then

$$\sum_{i=1}^{n} \frac{1}{c_i} \geq \frac{n^2}{\sum_{i=1}^{n} c_i} \tag{4.8}$$

with equality if and only if $c_1 = \ldots = c_n$. Moreover

$$n \sum_{i=1}^{n} c_i^2 \geq \left(\sum_{i=1}^{n} c_i \right)^2 \tag{4.9}$$

again with equality if and only if $c_1 = \ldots = c_n$. Both inequalities are special cases of the Power Mean Theorem [cf. 165, 258].

For the First Baseline Result, we need the following:

Lemma 4.8. *Let $k < n$ and let $(c_1, \ldots, c_n)$ be a sequence such that*

(1) $\sum_{i=1}^{n} c_i = s$ for some $s > 0$ and all c_i are positive;

(2) $c_1 = \ldots = c_k$ and $c_{k+1} = \ldots = c_n$;

[4] Lehrer and Wagner also defend their model from a normative point of view, but their arguments for this claim are not particularly persuasive, see e.g., Martini et al. [114].

(3) $c_k \leq c_{k+1}$ *and* $1 \leq \frac{c_{k+1}}{c_k} \leq \frac{c^*_{k+1}}{c^*_k}$.

Further assume that $\sigma_1 \geq \ldots \geq \sigma_n$. *Then*

$$\sum_{i=1}^{n} \left(\frac{s}{n}\right)^2 \sigma_i \geq \sum_{i=1}^{n} c_i^2 \sigma_i.$$

Furthermore, we show that under the above conditions (i.e. $\sum_{i=1}^n c_i = s$*), the value of the sum* $\sum_{i=1}^n c_i^2\sigma_i$ *decreases as the quotient* $\frac{c_{k+1}}{c_k}$ *increases.*

Proof of Lemma 4.8. Fix r such that

- $c_i = \frac{s}{n} - \frac{r}{k}$ for $i \leq k$
- $c_i = \frac{s}{n} + \frac{r}{n-k}$ for $i > k$.

Then we have to show that:

$$\sum_{i\leq k} \left(\frac{s}{n} - \frac{r}{k}\right)^2 \sigma_i + \sum_{i>k} \left(\frac{s}{n} + \frac{r}{n-k}\right)^2 \sigma_i - \sum_{i=1}^{n} \left(\frac{s}{n}\right)^2 \sigma_i \leq 0.$$

The above equation reduces to:

$$r^2 \left(\sum_{i\leq k} \frac{1}{k^2}\sigma_i + \sum_{i>k} \frac{1}{(n-k)^2}\sigma_i\right) - \frac{2s}{n} r \left(\sum_{i\leq k} \frac{1}{k}\sigma_i - \sum_{i>k} \frac{1}{n-k}\sigma_i\right) \leq 0. \quad (4.10)$$

Now the left hand side of the above equation is a quadratic function in r with zeros at 0 and

$$r_0 = \frac{2s}{n} \frac{\sum_{i\leq k} \frac{1}{k}\sigma_i - \sum_{i>k} \frac{1}{n-k}\sigma_i}{\sum_{i\leq k} \frac{1}{k^2}\sigma_i + \sum_{i>k} \frac{1}{(n-k)^2}\sigma_i}. \quad (4.11)$$

Since the σ_i are ordered decreasingly we get

$$r_0 \geq \frac{2s}{n} \frac{\sum_{i\leq k} \frac{1}{k}\sigma_i - \sigma_{k+1}}{\sum_{i\leq k} \frac{1}{k^2}\sigma_i + \frac{1}{(n-k)}\sigma_{k+1}}.$$

Now this is a function of the form $\frac{kx-a}{x+b}$ with $a, b > 0$. Since these functions are increasing for $x > -b$, the inequality above can be strengthened to

$$r_0 \geq \frac{2s}{n} \frac{\sigma_k - \sigma_{k+1}}{\frac{1}{k}\sigma_k + \frac{1}{(n-k)}\sigma_{k+1}}.$$

Recall that $\frac{c_{k+1}}{c_k} \leq \frac{c^*_{k+1}}{c^*_k} = \frac{\sigma_k}{\sigma_{k+1}} =: \sigma$ Inserting this transforms the above equation into:

$$r_0 \geq \frac{2s}{n} \frac{(\sigma-1)\sigma_{k+1}k(n-k)}{\sigma_{k+1}((n-k)\sigma+k)}.$$

Our assumptions about the c_i translate into

$$\frac{\frac{s}{n}+\frac{r}{n-k}}{\frac{s}{n}-\frac{r}{k}} \leq \frac{c^*_{k+1}}{c^*_k} = \frac{\sigma_k}{\sigma_{k+1}}.$$

This transforms to

$$r \leq \frac{s}{n}\frac{(\sigma-1)k(n-k)}{(n-k)-\sigma k}.$$

In particular $r < r_0$, finishing the proof of (4.10). For the last statement of Lemma 4.8, observe that the left hand side of (4.10) is a quadratic function with minimum $\frac{1}{2}r_0$, and that $r \leq \frac{1}{2}r_0$. □

Proof of Theorem 4.1. By assumption the c_i are ordered increasingly, thus the σ_i are ordered decreasingly. For a vector of weights $\mathbf{w} \in \mathbb{R}^n$ (i.e. all w_i positiv and $\sum_i w_i = 1$), we denote the mean square error of the estimator $\sum w_i X_i$ by $\Psi(\mathbf{w})$: That is,

$$\Psi(\mathbf{w}) := \sum w_i^2 \sigma_i.$$

Thus for $\mathbf{c} = (c_1 \dots c_n)$ as in the theorem we have to show $\Psi(\mathbf{c}) \leq \Psi(\mathbf{e})$, where $\mathbf{e}$ is the equal weight vector $(\frac{1}{n}, \dots, \frac{1}{n})$. To this end we will construct a sequence of weight vectors $\mathbf{e} = \mathbf{d}_0, \dots, \mathbf{d}_{n-1} = \mathbf{c}$ such that

(i) each $\mathbf{d}_i$ satisfies the assumptions of Theorem 1;

(ii) for $\mathbf{d}_i = (d_1 \dots d_n)$, there is some $k \in \mathbb{N}$ such that

$$d_1 = \dots = d_k \text{ and } d_1 > c_1; \dots; d_k > c_k;$$
$$d_j = c_j \text{ for } k < j \leq k+i \qquad \text{(where } i \text{ is the index of } \mathbf{d}_i\text{)};$$
$$d_{k+i+1} = \dots = d_n \text{ and } d_{k+i+1} \leq c_{k+i+1}; \dots; d_n \leq c_n;$$

(iii) $\Psi(d_{i-1}) \geq \Psi(d_i)$.

Thus $\mathbf{d}_{i-1} = \mathbf{c}$ and $\Psi(\mathbf{c}) \leq \Psi(\mathbf{e})$ as desired. The $\mathbf{d}_i$ are constructed inductively as follows: Assume $\mathbf{d}_{i-1} = (d'_1 \dots d'_n)$ has already been constructed. If $i = 1$ let k be the unique index such that $c_k < \frac{1}{n}$ and $c_{k+1} \geq \frac{1}{n}$. If $i > 1$ let k be as in the above conditions for $\mathbf{d}_{i-1}$. First note that if $k = 0$, then $d'_j \leq c_j$ for all j and thus $\mathbf{d}_{i-1} = \mathbf{c}$ since both are weight vectors and we are done. Thus assume $k \geq 1$ for the rest of the proof. With a similar argument, we can show that $k + i + 1 \leq n$. Now choose the maximal $r \in \mathbb{R}$ that satisfies

$$d'_k - c_k \geq \frac{r}{k} \qquad c_{k+i+1} - d'_{k+i+1} \geq \frac{r}{n-k-i-1}. \tag{4.12}$$

By the above conditions, $r \geq 0$. Then define $\mathbf{d}_i = (d_1, \dots, d_n)$ by

- $d_j = d'_j - \frac{r}{k}$ for $j \leq k$;
- $d_j = c_j$ for $k < j \leq k+i$;
- $d_j = d'_j + \frac{r}{n-k-i-1}$ for $j \geq k+i+1$.

To see that $\mathbf{d}_i$ satisfies conditions (i)-(iii), first note that since r was chosen to be maximal, one of the two inequalities in (4.12) has to be an equality. Thus we either have $d_k = c_k$ or $d_{k+i+1} = c_{k+i+1}$ and condition (ii) is satisfied. Further note that

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} d'_i - \sum_{i \leq k} + \frac{r}{k} + \sum_{i \geq k+i+1} \frac{r}{n-k-i-1} = 1.$$

Using that the c_i are ordered increasingly, it is easy to see that $\mathbf{d}_i$ satisfies the assumptions of Theorem 4.1. Furthermore, applying the monotonicity part of Lemma 4.8 to the set of indices $I := \{1, \ldots, k\} \cup \{i+k+1, \ldots, n\}$, we get $\sum_I d_i \sigma_i^2 \leq \sum_I d'_i \sigma_i^2$. Thus $\Psi(\mathbf{d}_i) \leq \Psi(\mathbf{d}_{i-1})$ since $\mathbf{d}_{i-1}$ and $\mathbf{d}_i$ coincide outside I. This finishes the proof. □

Proof of Theorem 4.2. We would like to show that the mean square error of the straight average $\bar{\mu} := (1/n)\sum_{i=1}^{n} X_i$ exceeds the mean square error of the weighted estimate $\hat{\mu}$. The MSE difference can be calculated as

$$\begin{aligned} \Delta(c_1, \ldots, c_n) &:= \mathrm{MSE}(\bar{\mu}) - \mathrm{MSE}(\hat{\mu}) = \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n} \sigma_i^2 - \sum_{i=1}^{n} c_i^2 \sigma_i^2 \\ &= \tfrac{1}{n^2}\left(\textstyle\sum_{j=1}^{n} \frac{1}{\sigma_j^2}\right)^{-1} \sum_{i=1}^{n} \frac{1}{c_i^*}\left(1 - n^2 c_i^2\right) \end{aligned}$$

where we have made use of $\mathcal{E}[X_i\, X_j] = 0$, $\forall i \neq j$, and of $c_i^* = \left(\sum_{j=1}^{n} \frac{\sigma_i^2}{\sigma_j^2}\right)^{-1}$ (cf. equation (4.1)). Thus, instead of considering Δ, it suffices to show that

$$\Delta'(c_1 \ldots c_n) := \sum_{i=1}^{n} \frac{1}{c_i^*}\left(1 - n^2 c_i^2\right) \geq 0.$$

To this end, let $I_i := [1/n; c_i^*]$ (respectively $[c_i^*; 1/n]$) and let $\mathcal{Q} := I_1 \times \ldots \times I_n$. Then,

$$\mathcal{D} := \mathcal{Q} \cap \{(c_1, \ldots, c_n) | \sum_{i=1}^{n} c_i = 1\}$$

defines the "domain" of our theorem, and it is a polygon. Moreover, since $\sum_i \frac{n^2}{c_i^*} c_i^2$ is a positive determinate quadratic form in the c_i, we get that ${\Delta'}^{-1}([0; \infty))$ is convex. Thus, it suffices to show that Δ' is positive on the vertices of $\mathcal{D}$. Note that since $\{x | \sum x_i = 1\}$ is of dimension $n-1$, the vertices of $\mathcal{D}$ are of

the form $\mathbf{v} = (c_1^*, \ldots, c_{k-1}^*, c_k, 1/n, \ldots, 1/n)$ – the ordering is assumed for convenience, and c_k is defined such that $||\mathbf{v}||_1 = 1$. Thus we have to show that $\Delta'(c_1^*, \ldots, c_{k-1}^*, c_k, 1/n, \ldots, 1/n) \geq 0$.

In the case $k = 1$, the desired inequality holds trivially since $c_k = 1 - (n - 1) \cdot (1/n) = 1/n$. Thus we assume $k > 1$ for the remainder of this proof. Let l denote the real number satisfying

$$\sum_{i=1}^{n} c_i^* = l\frac{k-1}{n}.$$

Observe that for $c_i = \frac{1}{n}$ the corresponding summands in Δ' vanish. Thus we have to show that

$$\sum_{i=1}^{k-1} \frac{1}{c_i^*}\left(1 - n^2 {c_i^*}^2\right) + \frac{1}{c_k^*}\left(1 - n^2 c_k^2\right) \geq 0.$$

Using the definition of l from above and inequality (4.8) gives $\sum_{i=1}^{k-1} \frac{1}{c_i^*} \geq (k-1)^2/(\sum_{i=1}^{k-1} c_i) \geq \frac{n(k-1)}{l}$. Thus, it suffices to show

$$n(k-1)\left(\frac{1}{l} - l\right) + \frac{1}{c_k^*}\left(1 - n^2 c_k^2\right) \geq 0. \tag{4.13}$$

Since the c_i add up to one, we can express the dependency between l and c_k by

$$c_k = \frac{(k-1)(1-l)+1}{n} \qquad \text{or by} \qquad l = \frac{k - nc_k}{k-1}. \tag{4.14}$$

Inserting this into (4.13) gives

$$\begin{aligned}
\Delta'(c_1, \ldots, c_n) &= \left(\tfrac{1}{l} - l\right) n(k-1) - \tfrac{1}{c_k^*}\left((1-l)^2(k-1)^2 + 2(1-l)(k-1)\right) \\
&= \tfrac{k-1}{l}\left[\left(1 - l^2\right) n - \tfrac{l}{c_k^*}\left((1-l)^2(k-1) + 2(1-l)\right)\right] \\
&= \tfrac{k-1}{l}\left[(1-l)\left((1+l)n - \tfrac{l}{c_k^*}((1-l)(k-1)+2)\right)\right].
\end{aligned}$$

Since the first factor is always positive, it suffices to show that the factor in the square brackets, denoted by $P(l)$, is positive for every l that can occur in our setting. We do this by a case distinction on the value of c_k^*.

Case 1: $c_k^* \leq 1/n$. Noting $c_k \in [c_k^*, \frac{1}{n}]$ and the dependency (4.14) between l and c_k, we have to show that $P(l) \geq 0$ for all $l \in [1; \frac{k-nc_k^*}{k-1}]$. We observe that P is a polynomial of third order with zero points of P given by $P(1) = 0$ and

$$r_\pm = \frac{k + 1 - nc_k^* \pm \sqrt{(k+1-nc_k^*)^2 - 4(k-1)c_k^* n}}{2(k-1)}$$

with r_+ denoting the larger of these two numbers. With some algebra it also follows that $P'(1) \geq 0$ if and only if $c_k* \leq 1/n$. From the functional form of

$P(l)$ – a polynomial of the third degree with negative leading coefficient – we can then infer that $l = 1$ must be the middle zero point of P. To prove that $P(l) \geq 0$ in the critical interval, it remains to show that for the rightmost zero point, we have $r_+ \geq \frac{k-nc_k^*}{k-1}$:

$$\begin{aligned}
&\frac{k-nc_k^*}{k-1} && \leq r_+ \\
\Leftrightarrow &\frac{2k-2nc_k^*}{2(k-1)} && \leq \frac{k+1-c_k^*n+\sqrt{(k+1-nc_k^*)^2-4(k-1)c_k^*n}}{2(k-1)} \\
\Leftrightarrow &k-1-nc_k^* && \leq \sqrt{(k+1-nc_k^*)^2-4(k-1)c_k^*n} \\
\Leftrightarrow &c_k^*n && \leq 1
\end{aligned}$$

completing the proof for the case $c_k* \leq 1/n$.

Case 2: $c_k* \geq 1/n$. In this case we are dealing with the interval $l \in [\frac{k-nc_k^*}{k-1}; 1]$. The same calculations as above yield

$$\frac{k-nc_k^*}{k-1} \geq r_+ \qquad \text{if and only if} \qquad c_k^*n \geq 1,$$

in particular $r_+ < 1$. Thus l always lies between the middle and the rightmost zero point of $P(l)$, and in particular, $P(l) \geq 0$ for all $l \in [\frac{k-nc_k^*}{k-1}; 1]$. □

Proof of Theorem 4.3. Let the X_i center around $B_i > 0$. Then $\mathcal{E}[X_i - B_i] = 0$, and we observe

$$\mathcal{E}\left[\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i\right)^2\right] = \mathcal{E}\left[\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n (X_i - B_i)\right)^2\right] + \left(\tfrac{1}{n}\textstyle\sum_{i=1}^n B_i\right)^2.$$

Analogously, we obtain

$$\mathcal{E}\left[\left(\textstyle\sum_{i=1}^n c_i X_i\right)^2\right] = \mathcal{E}\left[\left(\textstyle\sum_{i=1}^n c_i (X_i - B_i)\right)^2\right] + \left(\textstyle\sum_{i=1}^n c_i B_i\right)^2.$$

Like in Theorem 4.2, we define $\Delta(c_1, \ldots, c_n) := \text{MSE}(\bar{\mu}) - \text{MSE}(\hat{\mu})$ as the difference in MSE between both estimates and show that $\Delta(c_1, \ldots, c_n) \geq 0$ if equation (4.3) is satisfied.

$$\begin{aligned}
\Delta(c_1, \ldots, c_n) &:= \mathcal{E}\left[\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n (X_i - B_i)\right)^2\right] - \mathcal{E}\left[\left(\textstyle\sum_{i=1}^n c_i (X_i - B_i)\right)^2\right] \\
&+ \left(\tfrac{1}{n}\textstyle\sum_{i=1}^n B_i\right)^2 - \left(\textstyle\sum_{i=1}^n c_i B_i\right)^2.
\end{aligned}$$

By Theorem 4.1 and/or Theorem 4.2, the first line is greater or equal to zero, and by equation (4.3), the second line is also non-negative. Thus $\Delta(c_1, \ldots, c_n) \geq 0$, showing the superiority of differential weighting.

For the second part of the theorem, we just observe that

$$\mathcal{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - B_i)\right)^2\right] - \mathcal{E}\left[\left(\sum_{i=1}^{n} c_i(X_i - B_i)\right)^2\right] \geq \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2. \qquad (4.15)$$

□

Proof of Lemma 4.4. It is easy to see that the conditions of the corollary satisfy the requirements of part (a) of Theorem 4.3. This yields the desired result for the first part of the theorem. For the second part, let the X_i all center around $B \neq 0$. Then $X_i - B$ is unbiased, and we observe

$$\begin{aligned} \mathcal{E}\left[\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i\right)^2\right] &= \mathcal{E}\left[\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(X_i - B)\right)^2\right] + B^2 \\ \mathcal{E}\left[\left(\textstyle\sum_{i=1}^{n} c_i X_i\right)^2\right] &= \mathcal{E}\left[\left(\textstyle\sum_{i=1}^{n} c_i(X_i - B)\right)^2\right] + B^2. \end{aligned}$$

Therefore, under the conditions of the theorem,

$$\Delta(c_1, \ldots, c_n) = \quad \mathcal{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - B)\right)^2\right] - \mathcal{E}\left[\left(\sum_{i=1}^{n} c_i(X_i - B)\right)^2\right]$$

showing that Δ only depends on the centered estimates. □

Proof of Fact 4.5. First we deal with straight averaging:

$$\begin{aligned} &\mathcal{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right] - \mathcal{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathcal{E}\left[X_i X_j\right] - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathcal{E}\left[Y_i Y_j\right] \geq 0. \end{aligned}$$

The proof exploits that X_i and Y_i have the same variance, thus $\mathcal{E}\left[X_i^2\right] = \mathcal{E}\left[Y_i^2\right]$. The proof for differential weights is similar, making use of the fact that the c_i are the same for X_i and Y_i because they only depend on the variance of the random variable. □

Proof of Theorem 4.6, part (i). First, assume without loss of generality that $c_i \geq c_{i+1}$ for all $i < n$. Thus, our assumption on the $\mathcal{E}[X_i X_j]$ reduces to $\mathcal{E}[X_i X_k] \leq \mathcal{E}[X_j X_k]$ for $i \geq j \neq k$. First, we show the theorem under the

assumption that all $\mathcal{E}[X_iX_j]$ with $i \neq j$ are equal, say $\mathcal{E}[X_iX_j] = \gamma$. By Theorem 4.1 and/or 4.2, it suffices to show that

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathcal{E}\left[X_i, X_j\right] - \sum_{i=1}^{n}\sum_{j\neq i} c_i c_j\, \mathcal{E}\left[X_iX_j\right] \geq 0.$$

Inserting $\mathcal{E}[X_iX_j] = \gamma$ this reduces to

$$\gamma \cdot \left(\frac{n-1}{n} - \sum_{i=1}^{n}\sum_{j\neq i} c_i\, c_j\right) \geq 0. \tag{4.16}$$

The point $(1/n, \ldots, 1/n)$ is a global minimum of the function $f(\mathbf{x}) = \sum_i x_i^2$ under the constraints $x_1, \ldots, x_n \geq 0$ and $\sum_i x_i = 1$. Thus we have

$$\frac{1}{n} = f\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) \leq f\left(\mathbf{c}\right) = \sum_{i=1}^{n} {c_i}^2. \tag{4.17}$$

Observing $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i\, c_j = \left(\sum_{i=1}^{n} c_i\right)^2 = 1$ and combining this equality with (4.16) and (4.17), we obtain

$$\frac{n-1}{n} - \sum_{i=1}^{n}\sum_{j\neq i} c_i\, c_j = \frac{n-1}{n} - \sum_{i=1}^{n}\sum_{j=1}^{n} c_i\, c_j + \sum_{i=1}^{n} {c_i}^2 \geq 0$$

thus proving the statement in the case that all $\mathcal{E}[X_iX_j]$ are the same.

For the general case let us assume that not all c_i are the same (otherwise the theorem is trivially true). Thus we either have $c_1 > c_{n-1}$ or $c_2 > c_n$ since the c_i are ordered decreasingly. In the following, we assume $c_2 > c_n$, the other case works with a similar argument. First observe that

$$\sum_{i=1}^{n}\sum_{j\neq i} c_i c_j\, \mathcal{E}\left[X_iX_j\right] = 2\sum_{i=1}^{n}\sum_{j<i} c_i c_j\, \mathcal{E}\left[X_iX_j\right].$$

Thus, we can concentrate on $\{\mathcal{E}[X_iX_j] | i > j\}$. We fix a natural number c and let S_c be the set of all vectors $(\mathcal{E}[X_iX_j])_{(i>j)}$ fulfilling the conditions of our theorem and $\sum_{i>j}\mathcal{E}[X_iX_j] = c$. We then consider the functional

$$\begin{aligned}\tilde{\varphi}(e) &:= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j<i}\mathcal{E}\left[X_iX_j\right] - \sum_{i=1}^{n}\sum_{j<i} c_i c_j\, \mathcal{E}\left[X_iX_j\right] \\ &= \frac{1}{2}\left[\sum_{i=1}^{n}\sum_{j\neq i}\mathcal{E}\left[X_iX_j\right] - \sum_{i=1}^{n}\sum_{j\neq i} c_i c_j\, \mathcal{E}\left[X_iX_j\right]\right]\end{aligned}$$

on S_c. Observe that every S_c contains exactly one point e_{eq} where all $\mathcal{E}[X_iX_j]$ are equal. By the first part of this proof, $\tilde{\varphi}(e_{eq})$ is non-negative. Thus, it suffices

to show that e_{eq} is an absolute minimum of $\tilde{\varphi}$ on S_c. First, observe that the value of $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j<i}\mathcal{E}\left[X_i, X_j\right]$ is constantly $\frac{c}{n^2}$ on S_c, thus it suffices to show that

$$\varphi(e) := \sum_{i=1}^{n}\sum_{j<i} c_i c_j \,\mathcal{E}\left[X_i X_j\right] \tag{4.18}$$

attains its maximum on S_c in e_{eq}.

To do so, we show the following: For every $e \in S_c$ with $e \neq e_{eq}$ there is some $e' \in S_c$ with $\varphi(e') > \varphi(e)$. In particular, φ does not take its maximum on S_c in e. Thus assume that $e = (\mathcal{E}[X_i X_j])_{(i>j)} \in S_c$ is given. Since $e \neq e_{eq}$ there are some indices $s > t$ and $k > l$ such that $\mathcal{E}[X_s X_t] \neq \mathcal{E}[X_k X_l]$. Furthermore, we can assume that $t \geq l$. Without loss of generality (by potentially replacing one of the two entries with $\mathcal{E}[X_s X_l]$) we can assume that either $s = k$ or $t = l$. In the following we assume $s = k$, the other case works similarly. The idea of the following construction is: We show that moving towards a more equal distribution of the entries $\mathcal{E}[X_i X_j]$ increases $\varphi(e)$. In particular, we construct $e' = (\mathcal{E}'[X_i X_j])_{(i>j)} \in S_c$ as follows: In every row $r_i := \langle \mathcal{E}[X_i X_1] \ldots \mathcal{E}[X_i X_{i-1}] \rangle$ of e we replace all the entries of this row by their arithmetic mean. Formally, that is for all i and j (independent of j):

$$\mathcal{E}'[X_i X_j] = \frac{1}{i-1}\sum_{l<i}\mathcal{E}[X_i X_l].$$

Trivially this operation satisfies for all i:

$$\sum_{j<i}\mathcal{E}[X_i X_j] = \sum_{j<i}\frac{1}{i-1}\sum_{j<i}\mathcal{E}[X_i X_j] = \sum_{j=1}^{i-1}\mathcal{E}'[X_i X_j]$$

and thus also for the double sum:

$$\sum_{i=1}^{n}\sum_{j<i}\mathcal{E}[X_i X_j] = \sum_{i=1}^{n}\sum_{j<i}\mathcal{E}'[X_i X_j].$$

In particular e' is in S_c. Furthermore, we have assumed that the c_i are ordered decreasingly. Recall that $c_k > c_j$ implies $\mathcal{E}[X_i X_k] \leq \mathcal{E}[X_i X_j]$ by assumption, therefore the rows r_i were ordered increasingly, and thus the rows of $e' - e$:

$$\mathcal{E}'[X_i, X_1] - \mathcal{E}[X_i X_1]; \ldots; \mathcal{E}'[X_i, X_{i-1}] - \mathcal{E}[X_i X_{i-1}]$$

are ordered decreasingly (since the rows of e' are constant). In particular, we have for any i:

$$0 = \sum_{j<i}\mathcal{E}'[X_i X_j] - \mathcal{E}[X_i X_j] \leq \sum_{j<i} c_i c_j (\mathcal{E}'[X_i X_j] - \mathcal{E}[X_i X_j]) \tag{4.19}$$

where the $\leq$ comes from the fact that both c_j and $\mathcal{E}'[X_iX_j] - \mathcal{E}[X_iX_j]$ are decreasing in j. Summing that up over all i we get that

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \sum_{j<i} \mathcal{E}'[X_iX_j] - \mathcal{E}[X_iX_j] \\ &\leq \sum_{i=1}^{n} \sum_{j<i} c_ic_j \left(\mathcal{E}'[X_iX_j] - \mathcal{E}[X_iX_j]\right) = \varphi(e') - \varphi(e). \end{aligned}$$

Thus we have $\varphi(e') \geq \varphi(e)$ as desired. Now observe that (4.19) for $i = s$ is the following:

$$\begin{aligned} 0 &= \sum_{j<s} \mathcal{E}'[X_sX_j] - \mathcal{E}[X_sX_j] \\ &= \sum_{j<s, j\neq t,l} \left(\mathcal{E}'[X_sX_j] - \mathcal{E}[X_sX_j]\right) + \mathcal{E}'[X_sX_t] - \mathcal{E}[X_sX_t] + \mathcal{E}'[X_sX_l] - \mathcal{E}[X_sX_l] \end{aligned}$$

with both,

$$\sum_{j<s, j\neq t,l} \mathcal{E}'[X_sX_j] - \mathcal{E}[X_sX_j] \leq \sum_{j<s, j\neq t,l} c_sc_j \left(\mathcal{E}'[X_sX_j] - \mathcal{E}[X_sX_j]\right)$$

and

$$\begin{aligned} &\mathcal{E}'[X_sX_t] - \mathcal{E}[X_sX_t] + \mathcal{E}'[X_sX_l] - \mathcal{E}[X_sX_l] \\ &\leq c_sc_t(\mathcal{E}'[X_sX_t] - \mathcal{E}[X_sX_t]) + c_sc_l(\mathcal{E}'[X_sX_l] - \mathcal{E}[X_sX_l]). \end{aligned}$$

By construction we have $\mathcal{E}[X_sX_t] \neq \mathcal{E}[X_sX_l]$, thus we would have a strict inequality in the last summand (and thus in the entire sum) if we knew that $c_t \neq c_l$. Unfortunately, this is not always the case. However, we have put ourselves in a situation where applying the same construction again with $\mathcal{E}'[X_2X_1]$ and $\mathcal{E}'[X_nX_1]$ replacing $\mathcal{E}[X_sX_t]$ and $\mathcal{E}[X_sX_l]$ yields the desired (since we have assumed that $c_2 > c_n$). To see this, observe that

- $\mathcal{E}[X_2X_1] = \mathcal{E}'[X_2X_1]$ by construction
- $\mathcal{E}'[X_sX_1] > \mathcal{E}[X_sX_1]$ since $\mathcal{E}[X_sX_t] \neq \mathcal{E}[X_s, X_l]$ and $\mathcal{E}[X_sX_1]$ is the minimal element in the row r_s
- $\mathcal{E}[X_2X_1] \leq \mathcal{E}[X_sX_1]$ by assumption.

Thus we have

$$\mathcal{E}'[X_2X_1] = \mathcal{E}[X_2X_1] \leq \mathcal{E}[X_sX_1] < \mathcal{E}'[X_sX_1] \leq \mathcal{E}'[X_nX_1].$$

By assumption we have $c_2 > c_n$ and repeating the construction from above with columns replacing rows and $\mathcal{E}'[X_2, X_1], \mathcal{E}'[X_n, X_1]$ as the two reference points

yields the desired.

Proof of Theorem 4.6, part (ii): We have to show that the statement holds if all $\mathcal{E}[X_i X_j]$ with $i \neq j \in I$ are the same. The step from this case to the general statement works as in the proof above. As in the proof of i), it suffices to show that

$$\frac{1}{n^2} \sum_{i \in I} \sum_{j \neq i \in I} 1 \geq \sum_{i \in I} \sum_{j \neq i \in I} c_i c_j.$$

Let $\bar{c} = \frac{1}{|I|} \sum_{i \in I} c_i$. By equation (4.9) we have

$$\sum_{i \in I} c_i^2 \geq \frac{1}{|I|} \left(\sum_{i \in I} c_i \right)^2 = \frac{1}{|I|} |I|^2 \bar{c}^2 = |I| \bar{c}^2$$

thus

$$\sum_{i \in I} \sum_{j \neq i \in I} c_i c_j \leq (|I|^2 - |I|) \bar{c}^2 \leq |I|^2 - |I| = \frac{1}{n^2} \sum_{i \in I} \sum_{j \neq i \in I} 1$$

with the last inequality coming from our assumption that $\bar{c} < 1$.

Proof of Theorem 4.7: Let the benchmark agent have standard deviation $s > 0$, that is, variance s^2. We will show that $\Delta(s, \sigma_1, \ldots, \sigma_n)$—the MSE difference between the differentially weighted and the straight average—is strictly monotonically decreasing in the first argument. To this effect, we calculate

$$\Delta(s, \sigma_1, \ldots, \sigma_n) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 - \left(\frac{1}{\sum_k c_k} \right)^2 \sum_{i=1}^{n} c_i^2 \sigma_i^2.$$

Now we show that $\frac{\partial}{\partial s} \Delta(s, \sigma_1, \ldots, \sigma_n) \leq 0$, where c_i' denotes $(\partial/\partial) s c_i$:

$$\begin{aligned}
\frac{\partial}{\partial s} \Delta(s, \sigma_1, \ldots, \sigma_n) &= -\frac{\partial}{\partial s} \left(\sum_{i=1}^{n} \frac{c_i^2}{(\sum_k c_k)^2} \sigma_i^2 \right) \\
&= -\sum_{i=1}^{n} \sigma_i^2 \cdot 2 \cdot \left(\frac{c_i}{\sum_k c_k} \right) \frac{c_i' \sum_j c_j - c_i \sum_j c_j'}{(\sum_k c_k)^2} \\
&= -\frac{2}{(\sum_k c_k)^3} \sum_{i=1}^{n} \sigma_i^2 c_i \left(\sum_{j \neq i} c_i' c_j - c_i c_j' \right) \\
&= -\frac{2}{(\sum_k c_k)^3} \sum_{i=1}^{n} \sum_{j<i} \left(\sigma_i^2 c_i - \sigma_j^2 c_j \right) \left(c_i' c_j - c_i c_j' \right).
\end{aligned}$$

Since we are only interested in the sign of the first derivative and since $-\frac{2}{(\sum_k c_k)^3} < 0$, it suffices to show that

$$\left(\sigma_i^2 c_i - \sigma_j^2 c_j \right) \left(c_i' c_j - c_j' c_i \right) \geq 0. \tag{4.20}$$

We show that the terms in both brackets have the same sign.

For the first bracket we have:

$$\begin{aligned}\sigma_i^2 c_i - \sigma_j^2 c_j &= s^2 \tfrac{\sigma_i^2}{s^2+(n-1)\sigma_i^2} - s^2 \tfrac{\sigma_j^2}{s^2+(n-1)\sigma_j^2} \\ &= s^4 \tfrac{\sigma_i^2-\sigma_j^2}{(s^2+(n-1)\sigma_i^2)(s^2+(n-1)\sigma_j^2)}\end{aligned}$$

which is larger than or equal to 0 if and only if $\sigma_i^2 > \sigma_j^2$. Similarly, we observe for the second bracket that

$$c_i' = \frac{2(n-1)s\sigma_i^2}{(s^2+(n-1)\sigma_i^2)^2}.$$

which allows us to conclude

$$\begin{aligned}& c_i' c_j - c_j' c_i \\ = \;& \tfrac{2(n-1)s\sigma_i^2}{(s^2+(n-1)\sigma_i^2)^2} \cdot \tfrac{s^2}{s^2+(n-1)\sigma_j^2} - \tfrac{2(n-1)s\sigma_j^2}{(s^2+(n-1)\sigma_j^2)^2} \cdot \tfrac{s^2}{s^2+(n-1)\sigma_i^2} \\ = \;& 2(n-1)s^5 \; \tfrac{\sigma_i^2-\sigma_j^2}{(s^2+(n-1)\sigma_i^2)^2(s^2+(n-1)\sigma_j^2)^2}.\end{aligned}$$

Thus, both factors in (4.20) have the same sign, implying $\frac{\partial}{\partial s}\Delta(s,\sigma_1,\ldots,\sigma_n) \leq 0$ which is want we wanted to prove. □

Chapter 5

Expressive Voting: Modeling a Voter's Decision to Vote

5.1 Introduction

Elections are a central element of group decision making. It is voting that creates the link between a myriad of individual preferences, opinions and interests and the actions and decisions of the community. Given this importance of voting, it is not surprising that electoral patterns and behavior became a major topic of interest for political commentators and social scientists. Arguably, understanding and even predicting voting patterns is one of the central competences for navigating the political sphere. There is hardly any political action that won't be related to past or future elections by at least some commentators. Unsurprisingly, the most intense phase of such public attentions occurs right before and after any election day. When opening a newspaper the day after any major election, one of the main questions to be found will be "Why did people vote the way they did?" The question of why people vote the way they do has attracted the attention of many different fields, history, psychology, philosophy, political science and lately also computer science, using experimental, empirical and theoretical work.

Anthony Downs, in his seminal book [48], has focused on a slightly different question: Why do people vote at all? In a large election, so the argument of his *paradox of voting*, there is almost no chance that an individual action will have any effect on the outcome. Thus, if any cost is associated with the act of voting, the only *rational* choice[1] for a voter is not to participate in the

A version of this paper is currently under review. Parts of this chapter are based on joint work with E. Pacuit, see [92] for a related article

[1] In the sense that taking into account the cost of the act of voting, the expected utility of

election. However, there is a second approach, prominently put forward by Geoffrey Brennan and Loren Lomasky [32], that avoids the paradox of voting. In many cases, so they argue, we choose some option not because it maximizes our immediate expected outcomes, but because it is ethically correct, fair, polite, or coheres in some other way with some ideals we follow or want to follow. That is, we derive our utility not from the immediate outcomes of our actions, but from the fact that the actions or utterances themselves are in accord with certain principles we value.[2] Brennan and Lomasky argue that this reasoning particularly applies to voting considerations in political elections. Many decisions in the political realm reflect value judgments, for instance by depending upon a particular conception of fairness, siding with some particular camp or being sensitive to other ethical considerations. Consequentially, so Brennan and Lomasky, a typical voter derives her utility straight from the act of voting for a certain alternative, and thus the paradox of voting does not arise. Rather than considering the different possible outcomes, the only requirement of rationality for an expressive voter is to truthfully report her preferred alternative.

The main difference between these two approaches is the agents' ultimate goals, i.e., their source of utility. Downs holds that voters draw their utility from the *outcome* of an election, while Brennan and Lomasky argue that the agents derive their utility already from expressing their preferences. To introduce a bit of terminology, we refer to the first type of voting behavior, considering the outcome of an election, as *instrumental* voting, while calling the second type of voting behavior, attaching utilities to statements of preferences, *expressive* voting.

Both of these theories are primarily normative, informing a rational voter what she *should* ideally do. They do, however, have some descriptive backing. Brennan and Lomasky show [32, pp. 40-46] that actual voting behavior is best explained by a superposition of instrumental and expressive considerations, where the weights given to the two accounts depend upon various factors such as, for instance, the stakes involved or how close the election is expected to be. Crucially, the two accounts will, in general, give divergent advice, about whom to vote for, but also about whether or not to participate in the election at all. To give a prominent example, strategic voting, misrepresenting one's true preferences for tactical reasons, is a major topic in instrumental voting [138],

voting for a preferred candidate (or set of candidates) will be negative.

[2] Of course, all these options do maximize expected utility in a broader sense. However as Sen in [143] or Nozick's discussion of sunk cost [123] show, these actions can only be maximizing if we attach utility to the fact of adhering to certain norms and principles.

whereas an expressive analysis will always advice voters to represent their preferences *truthfully*. Given that voters will resort to both types of considerations, expressive as well as instrumental, a suitable approach for analyzing voting situations or comparing different electoral systems is to start by analyzing these from both standpoints, in order to later compare and combine these findings.

While instrumental accounts of voting have been thoroughly explored within the (formal) literature on voting, the expressive side has received considerably less attention yet. In this chapter, we will explore a formal model for expressive voting. Notably, there are two recent papers on expressive voting that we will refer to in our analysis. In both, the electoral decision is based on an agenda of topics the voters care about. This agenda can reflect anything of importance for the voters, ranging from a general liberal-conservative distinction to particular topics such as whether or not to bail out the car industry or to go to war in Syria. Candidates or parties merely serve as proxies for the different attitudes towards the agenda items they stand for. A candidate is associated with her position on each of the agenda items and the voters evaluate the candidates by these positions.

In the first paper, [6], Enriqueta Aragones, Itzhak Gilboa and Andrew Weiss discuss the question on whether to vote or not from an expressive standpoint. In their approach, abstentions are not caused by any factors external to the election. Rather, abstaining expresses the fact that, given all available alternatives, an empty ballot sheet best expresses the voter's preferences. Aragones *et al.* discuss different voting systems, in particular majority rule and approval voting, with respect to their ability to prevent an expressive voter from abstaining. That is, they study the potential of each system to offer some option that is more attractive than submitting an empty ballot sheet.

In the second paper, [46], Walter Dean and Rohit Parikh study the dynamics of expressive voting. In their account, candidates have not yet publicly committed to a unique position on the agenda. Rather, they are pondering when, how and whether to commit on the remaining topics, depending on the voters preference and their attitude towards uncertainty. The decision about whether or not to commit on any given topic will, in general, depend upon factors such as competing candidates, the distribution of voter preferences, but also how benevolent voters are in filling out informational gaps or how they revise their belief in light of newly incoming evidence. In particular the last topic relates the analysis of voting behavior to existing debates in belief revision and the appropriate rules thereof.

The goal of this chapter is to explore and discuss a formal framework of expressive voting, based on an agenda of topics. We will be mainly interested in two aspects, participation and dynamics. The first of these topics, participation, frequently arises in the comparison of different voting systems. The choice between different voting systems, such as plurality vote or approval voting, is guided by a variety of aspects, including their simplicity, their propensity to create stable outcomes,[3] their adequacy or their propensity to foster electoral participation. For this chapter, we will restrict our attention to the latter of these, that is, we are interested in analyzing the phenomenon of abstentions within expressive voting. Our second topic of interest, next to participation, is the dynamic patterns occurring within voting behavior. Voters' preferences and opinions are not carved in stone, but they gradually develop over time, taking into account political and economic developments, but also reacting to public debates or campaigning efforts of the different sides. Here, we are interested in how far such dynamic patterns can be represented within a formal framework for expressive voting.

To be a bit more precise, we will make three main contributions in this chapter, to be found in sections 5.3, 5.5 and 5.6 respectively. The latter two of these refer to the two aspects identified above, participation and dynamics. The first is an internal critique of the framework of Aragones *et al.* The core idea of our framework rests on their model presented in [6] that we are highly sympathetic towards. We hold, however, that it has a crucial conceptual shortcoming in its treatment of approval voting. Our first contribution will be to identify that shortcoming in section 5.3 and offer an alternative semantics of approval voting in the subsequent section 5.4 that does not fall to the same criticism. To strengthen our approach, we will show that it is compatible with three different choice rules the agents might entertain. Our second contribution then, to be found in section 5.5, is related to the participation in elections. Expanding on previous work by [6], we will study the occurrence of abstentions in three different voting systems, plurality vote, approval voting and range voting. Considering the extreme cases of completely strategically positioned parties on the one hand and randomly distributed parties on the other, we will show that the latter two voting systems, approval and range voting, are exponentially better than plurality voting in avoiding abstentions. Finally, our third contribution in section 5.6 is related to the dynamics of electoral campaigns. Beyond their positions on the individual topics, the decisions of individual voters crucially depend

[3] This criterion is especially relevant for parliamentary elections where the resulting parliament has to agree upon an ideally stable government.

upon which topics they *focus* on when casting their vote. If public debate is centered around the economy, many voters will think of the economically relevant agenda items when making their decision. Conversely, if foreign policy receives a lot of public attention, voters will heavily rely on their stances towards the European debt crisis or going to war in making their decision. Consequentially, different interest groups will aim to amend public focus in their respective interests. In section 5.6 we present a conceptual framework that allows to model focus and changes in focus brought about by public events. Further, we will explore some of the difficulties and subtleties arising when parties try to influence public focus in their own interest. Finally, we offer a conclusion and some directions of future work in section 5.7. All proofs and calculations are in the appendix.

5.2 The Model

In this section we present our basic model for elections. This model is borrowed from [6], hence we will refer to it as the AGW-model. The central object of study is the agenda of topics or issues of concern for the upcoming election, denoted by $I = \{1 \dots n\}$. We assume the agenda items to be propositions, such as *"There should be stricter gun control"* or *"European states should vouch for each other in the debt crisis"*, that the individual parties and candidates can either endorse or oppose to various degrees. Second, we denote the set of parties or candidates by $T = \{1 \dots m\}$. Each party $j \in T$ is characterized by its positions on the various issues of concern $I = \{1, \dots, n\}$. To this end, each party $j \in T$ is associated with a vector $\mathbf{p}^j \in [-1; 1]^n$ giving j's positions on each of the issues. The intended reading is that $p_i^j \in [-1; 1]$ is the degree to which candidate j supports issue i, where $+1$ stands for total support and -1 for total opposition to the topic in question. Just as in [6], we assume that parties have extreme positions on all topics, that is $p_i^j \in \{-1; 1\}$. Briefly, there are two different justifications for this assumption. First, [6] argue that political discourse moves parties to extreme positions. In an attempt to position themselves on the political scale and to stand out from their opponents they will finally have to commit to a clear position on each topic. Second, parties are identified with the policies they would enact, if elected. We assumed the individual agenda items to be propositional, thus these items can only be enacted or non-enacted. Of course, a party could breach some of their promises and act differently to what they claimed prior to the election day. Yet, any policy implemented will either enact some agenda item i or not, there is no space for a middle ground. Thus, there

is no space for graded judgments, but parties will eventually have to decide for or against implementing any particular item on the agenda. As for parties, also each voter is represented by a vector $\mathbf{v} \in [-1;1]^n$, representing her position on the various topics. Note that we do allow voters to have positions anywhere in $[-1;1]$ in order to allow for uncertainty about the right course of action or to display varying degrees of interest in the different topics.[4] The only case we exclude are universally disinterested voters, thus we assume $\mathbf{v} \neq \mathbf{0}$. For notational convenience, we will use $\mathbf{p}$ with decorations to denote parties or candidates, where $\mathbf{v}$ with all its variants denotes voters. Bold letters will always refer to vectors, while their entries are denoted in italics, for example $\mathbf{v} = (v_1 \dots v_n)$.

Last, we define a ballot, i.e., the possible votes a voter can cast. For our purposes, we represent each ballot by a vector $\mathbf{x} = (x_1, \dots, x_m) \in \mathbb{R}^m_+$, where the zero vector $\mathbf{0} \in \mathbb{R}^m_+$ denotes abstention. More precisely, each voting system consists of a set of admissible ballots $F \subset \mathbb{R}^m_+$ together with an aggregation rule for these feasible values. The two voting methods considered in AGW are:

> *Plurality rule:* Each voter selects a single candidate, and the candidate with the most votes is declared the winner. Thus, the feasible ballots are $F^M = \{\mathbf{0}\} \cup \{\mathbf{e}^j\}_{j \leq m}$, where $\mathbf{e}^j$ is the vector with 1 in the jth position and 0 everywhere else, denoting that the voter supports candidate p_j.
>
> *Approval Voting:* Voters select any subset J of the candidates they approve of. Again, the candidate receiving most approval wins the election. Thus, the set of ballots are
>
> $$F^A = \left\{ \mathbf{x}^J \in \mathbb{R}^m \mid J \subseteq \{1, \dots, m\} \text{ and } \mathbf{x}^J = \sum_{j \in I} e^j \right\}$$
>
> where the e^j are again the vectors having a 1 as jth component and a 0 everywhere else. With other words, $\mathbf{x}^J$ is the vector with a 1 for every party the voter approves of and a zero everywhere else.

Obviously there are more ballots available to voters under approval voting than plurality rule (i.e., $F^M \subseteq F^A$), thus within expressive voting abstentions should be less frequent than in the former framework.

[4] The rationale behind this definition is the following: Let's assume that a voter is given by *two* vectors, $\mathbf{u} \in [0;1]^n$ and $\mathbf{b} \in [0;1]^n$. The first of these, $\mathbf{u}$, denotes the importance the voter attaches to each topic; the second ,$\mathbf{b}$, her beliefs about the right action. Thus if $b_i = 1$ or $b_i = 0$ the agent is certain that the proposition representing agenda item i needs to be made true or wrong respectively, whereas $b_i = 0.5$ stands for maximal uncertainty. If we assume that each agent gets a payoff of u_i or $-u_i$ if proposition i later turns out to be the right policy or not, the expected value of implementing policy t is exactly $(2b_t - 1)u_t \in [-1;1]$. Thus, we could assume a voter's position $\mathbf{v} \in [-1;1]^n$ to reflect the vector $\langle (2b_1 - 1)u_1, \dots (2b_n - 1)u_n \rangle$, the product of her uncertainty and her attachment of relative importance to the topics.

Next, we, need to determine how an expressive voter *should* choose among the possible ballots. Rather than determining some particular utility-function $u : F \to \mathbb{R}$, we give a condition that every reasonable payoff function should satisfy and that is sufficient to determine the voter's choice. This principle is that a voter will prefer a party that is closer to her own position to a party that is further away. To make this precise, we measure distances in $[-1;1]^n$ in the euclidean distance: $dist(x,y) = \sqrt{\sum (x_i - y_i)^2}$, or equivalently in the 2-norm $|\cdot|_2$, which leads [6] to define the following choice rules:

> Under *plurality rule*, the voter $\mathbf{v}$ votes for the candidate that is closest to her or abstains if the empty ballot $\mathbf{0}$ is closer than any of the candidates. With other words, she chooses the ballot $\mathbf{x}_m \in F^M$ which is closest to her own standpoint in the euclidean distance. Formally speaking, this can be expressed in the following sum:
>
> $$\mathbf{x}_m = argmin_{\mathbf{x} \in F^M}\ dist(\mathbf{v}, \sum_{i \leq m} x_i\ \mathbf{p}^i).$$

Recall that $x_j = 1$ iff $\mathbf{v}$ votes for party $\mathbf{p}^i$. Thus the sum $\sum_{i \leq m} x_i\ \mathbf{p}^i$ is equal to $\mathbf{p}^j$ if $\mathbf{v}$ votes for $\mathbf{p}^j$ and $\mathbf{0}$ if $\mathbf{v}$ abstains. For approval voting, we need to extend the above rule to approval ballots. In particular, we need to specify how the voters compare different approval sets. To this end [6] represent every approval set $\mathbf{x}^J$ with its arithmetic mean[5] $\frac{1}{|J|} \sum_{j \in J} \mathbf{p}^j$, see table 5.1 for illustration. This leads to the following decision rule:

> Under *approval voting*, the voter $\mathbf{v}$ chooses the ballot $\mathbf{x}_a \in F^A$ which is closest to her own standpoint in the euclidean distance, or formally
>
> $$\mathbf{x}_a = argmin_{\mathbf{x}^J \in F^A}\ dist(\mathbf{v}, \frac{1}{|J|} \sum_{i \leq m} x_i\ \mathbf{p}^i).$$

We end this paragraph with a brief overview of the main results of [6]. Their analysis is centered around the question: *When do people vote?* Note that in our framework a voter only abstains if the corresponding position vector $\mathbf{0}$ is closer to her than any other possible ballot. That is, abstentions are not caused by external cost, as in Down's analysis, but by the fact that the voter fails to find any alternative that is more appealing. The central results of [6] compare the two voting systems with respect to their potential of generating a high degree of electoral involvement, measured by the number of abstentions. Since the possible ballots in approval voting form a strict superset of the ballots in plurality voting, we can expect an increased electoral participation within approval voting. This intuition is made precise in the following theorems:

[5] To facilitate our presentation, we set $\frac{1}{|\emptyset|} \sum_{j \in \emptyset} p^j := \mathbf{0}$, thus the empty approval set is represented by the zero-vector.

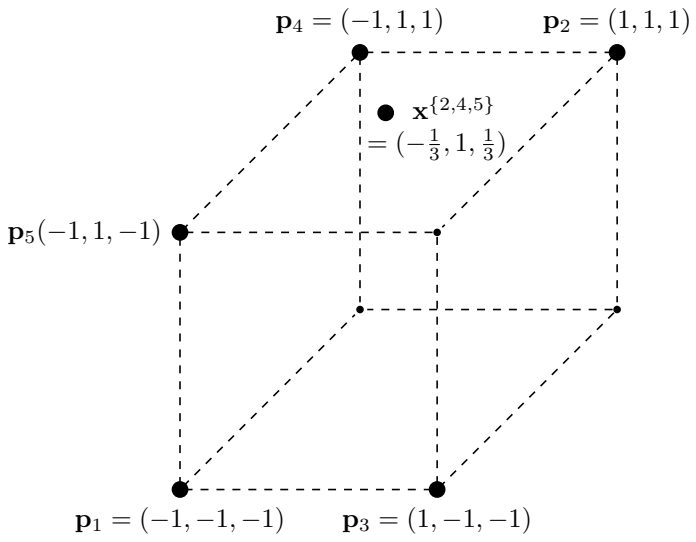


Table 5.1: In approval voting the approval set $\{2, 4, 5\}$ gets represented by the arithmetic mean of its members.

Theorem 5.1 (Theorem 1 of [6]). *i) Under approval voting 4 strategically positioned parties are enough to ensure that no voter abstains*
ii) Under plurality vote, the number of parties necessary to ensure that no voter abstains is exponential in the number n of agenda items.

Theorem 5.2 (Theorem 2 of [6]). *Assume the agenda exists of n topics and we randomly place n parties on this agenda (i.e., for every party we have a fair lottery of the 2^n possible positions). As $n \to \infty$, the probability that a voter abstains in this setting goes to 1 under plurality rule and to 0 under approval voting.*

5.3 Criticism of the AGW Approach

In this section, we identify two conceptual shortcomings of the AGW treatment of approval voting. These shortcomings will then motivate our alternative account, presented in the following section. First, we note that voting systems such as plurality vote or approval voting can be used in various settings. The different voting systems can be used in single winner elections, such as the French or American presidential elections or they could be applied to the French or German Parliamentary elections, that is, many winner elections in which the winners that make it to parliament subsequently have to enter coalitions if they

want to be part of government.[6] In their paper, AGW were silent about any intended interpretation. We will show two things here. First, that the general framework developed in [6] decisively rests on intuitions from single winner elections and, second, that the account of approval voting outlined in the last section additionally assumes certain features of multi-winner elections that are irreconcilable with these single-winner elements.

As a second shortcoming, we will identify an *internal* consistency requirement for expressive voting, violated by the AGW approach. This requirement refers to the very special case in which the entire electorate happens to share the same position about every item on the agenda. We show that even under such ideal conditions, the above rule for approval voting cannot guarantee that the voters are comfortable with the outcome. Even worse, we show that under these ideal conditions applied to a single winner election, the AGW approach can bring the *least* preferred candidate into office.

For our first criticism, we will show that the treatment of approval voting presented in the previous chapter rests on incoherent assumptions about the nature of the voting situation. On the one hand, we will demonstrate that some of the central assumptions made for the general framework implicitly need to assume the election to be based on a single-winner scenario. On the other hand, we will show that the semantics assumed for approval voting rests on some intuitions coming from multi-winner elections. We will deal with these points in reverse order, starting with the particular semantics for approval voting and then continuing with the assumptions underlying the general framework.

We start our discussion of approval voting by analyzing the AGW proposal in a bit more detail. Their decision rule suggests that a voter should vote for that subset J_0 of candidates for which the corresponding arithmetic mean $\mathbf{x}_0 = \frac{1}{|J_0|} \sum_{j \in J_0} e^j$ used to represent J_0 satisfies the condition

$$\mathbf{x}_0 = argmin_{\mathbf{x}^J \in F^A} dist(\mathbf{v}, \frac{1}{|J|} \sum_{i \leq m} x_i \ \mathbf{p}^i).$$

That is, in order to determine her approval of some nonempty set I of candidates, the voter calculates the straight average of the parties' stances. Or to put it differently: She composes a virtual coalition of the different parties in which every member receives the same weight. In order to consider virtual coalitions, the voter needs to reason about the election with a multi-winner mindset.

[6] Yet other systems, for instance the first past the post system used in British parliamentary elections, are situated between these two extremes. We hold that for the aspects relevant for our analysis, they are closer to multi-winner systems such as the German or French parliamentary elections.

The entire concept of forming coalitions with all related considerations such as neutralizing opposing opinions and interests and reaffirming joint interests is borrowed from a multi-winner setting. These considerations are alien to single winner elections and it would be at least strange to assume that voters would refer to them while reasoning about single winner settings. Thus, in order to justify the choice of decision rule for approval voting, we need to assume that the voting system is applied to a multi-winner election. This argument receives further backing by the example given at the end of this section, showing that the proposed analysis can lead to the worst possible results in a single winner setting.

On the other hand, we assumed, in line with AGW, that parties take extreme positions on every subject. This assumption is crucial for some of the central results in [6], in particular their proof of Theorem 2 cited above. We have offered two independent arguments for this assumption. To recall, our first argument was that, once in power, a party has to either implement a policy or its converse, thus there is no room for a graded representation of approval. Second, this is the main argument of [6], we argued that the campaigning and political discourse preceding an election will move the parties to extreme positions. The first argument obviously depends on a single-winner reading of elections, since only then the single winner, rather than some confederate or coalition partner, will have to decide on every single topic. So what about the second argument? Upon observing almost any coalition government of the last 50 years, we learn that the position of a coalition is *not* decided by applying one fixed averaging mechanism to every position. Rather, every coalition partner has some core interests that he is unwilling to compromise even to the slightest extent. In other areas, the parties are more willing to reconsider or even sacrifice their own stance, hoping to thereby gain support on other topics more important to them. In established multi-winner systems, all these considerations will be familiar to voters and candidates. Candidates will thus not move to extreme positions, but maintain some space for negotiations and indicate which topics are central to them and which are not. And voters, knowing this, will of course anticipate the differential stances of various coalitions. Thus, the assumption of fully opinionated parties is incompatible with multi-winner practices. In other words, the underlying model silently assumes that elections have a single winner, which is incompatible with the decision rule for approval voting that requires multi-winner election to have any traction. This finishes our first argument.

While the previous argument showed that the model rests on incoherent

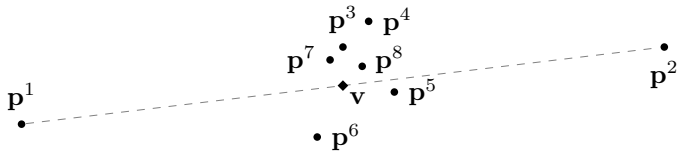


Table 5.2: Voter **v**'s position is (exactly) the arithmetic mean of the two most extreme parties

intuitions about the voting system, our second argument shows that approval voting might trigger paradoxical approval patterns for individual voters. Recall that expressive voting is completely blind towards any possible outcome of an election. Rather, expressive voters do obtain their utility straight from the act of submitting their ballot, independent of the resulting outcomes. However, if we do not wish to produce a formal model for hypocrites, we need to introduce some consistency requirements between expressed consent and electoral outcomes. We take the following as an uncontroversial desideratum.

> *In elections where all voters share exactly the same preferences, and thus submit the same ballots, any single voter would approve of the resulting outcome.*

We show that the AGW semantics for approval voting violates this criterion. To be more specific, we will show that in their setting, approval voting as defined above can bring some least desirable candidate into power. To capture the gist behind the following example 5.3, assume a situation where there is a set of moderate parties and two opposing extremist parties. It might so happen that the position of a moderate voter **v** is exactly the average between two extremist parties - even though every moderate party is closer to her than each of the extremists. See table 5.2 for an illustration. Under an AGW-style semantics of approval voting, **v** would have to approve of exactly the two extremist parties. Now if every voter had the same preference as **v**, all votes would go to the two extremist parties, and thus one of the two would get into office. By our assumptions this is the outcome **v** dislikes most among all possible outcomes.

Since we represent parties by their fully opinionated positions on particular topics, rather than by degrees of extremism, we cannot directly translate the above story into a formal counterexample. The following example mimics the main features of the above setting.

Example 5.3: Assume the agenda consists of 9 issues $t_1 \dots t_9$. The first four items concern the economy, taxes, environmental issues and the social system, four issues that **v** has some mild opinions about. She assigns position vectors

$\frac{1}{3}, -\frac{1}{3}, \frac{1}{3}, -\frac{1}{3}$ to $t_1 \dots t_4$. The other 5 topics concern difficult decisions in foreign policy where $\mathbf{v}$ finds it hard to choose sides, so she assigns them a weight of zero. The two extremist parties are $\mathbf{e}_+$ assigning 1 to every topic and $\mathbf{e}_-$ assigning -1 to every topic. Every other party $\mathbf{p}_i$ assigns weights $1, -1, 1, -1$ to the first four topics and 1 to all remaining topics. Then the setup is as claimed above, i.e., all moderate parties $\mathbf{p}_i$ are closer to $\mathbf{v}$ than both $\mathbf{e}_+$ and $\mathbf{e}_-$, but $\{\mathbf{e}_+, \mathbf{e}-\}$ is the approval set chosen by $\mathbf{v}$. We relegate the straightforward calculations to the appendix.

So far, we have identified two major problems of the AGW account of approval voting: inconsistency in the background assumptions and a violation of a coherency desideratum for voters. In the following chapter, we will present an alternative choice rule for approval voting that avoids both of these shortcomings. We will also show that our approach can naturally be extended to a third voting rule, range voting.

5.4 Our Model

In this section, we offer an alternative framework for approval voting. We will show that this framework naturally squares with the above decision rule for plurality vote, while avoiding the two pitfalls identified in the previous section. Crucially, approval voting and plurality vote invoke different choice strategies. Plurality vote requires the voter to optimize, that is identify the best among the parties and either vote for that party or abstain if that is more attractive. Approval voting, on the other hand, is built around the notion of satisficing. The central task for a voter in an approval election is to identify some minimal requirements she has towards the potential candidates. The voter will then approve of every party that meets or exceeds these requirements. Naturally, these minimal requirements will depend on the individual agenda items and the degrees of importance attached to these. In this section, we will identify three different ways in which voters could formulate their minimal requirements, expected payoff, geometrical proximity and a grading system, and show that these all lead to the same formalism.

To introduce our framework, recall that a voter's position on some topic t_i is given by a number $v_i \in [-1; 1]$, where -1 stands for total opposition and 1 for absolute consent. We can decompose the voter's attitude into

$$v_i = sign(v_i) \cdot |v_i|$$

where the sign[7] $sign(v_i)$ indicates whether $\mathbf{v}$ is inclined in favor or against t_i while the absolute value $|v_i|$ measures the degree of commitment[8] $\mathbf{v}$ attaches to topic i. For our model, we assume that this commitment $|v_i|$ is related to the payoff that $\mathbf{v}$ can obtain on this agenda item. More specifically we assume that, by voting for some party $\mathbf{p}$, our voter $\mathbf{v}$ gets a payoff $|v_i|$ on item i if $\mathbf{v}$ and $\mathbf{p}$ share the same inclination about topic i, that is both are in favor or both against. Else, the voter gets a payoff of $-|v_i|$ from voting for $\mathbf{p}$. In other words, the payoff is described by the following formula

$$|v_i| \text{ iff } v_i \cdot p_i \geq 0$$
$$-|v_i| \text{ iff } v_i \cdot p_i < 0.$$

Since we have assumed that $p_i \in \{-1; 1\}$, these two conditions can be combined into one rule: The payoff a voter $\mathbf{v}$ gets by voting for party $\mathbf{p}$ on topic i is $v_i \cdot p_i$. Thus, the total payoff $u(\mathbf{v}, \mathbf{p})$ a voter $\mathbf{v}$ gets by voting for party $\mathbf{p}$, that is the sum over all the individual payoffs on the different topics, is described by the standard scalar product as

$$u(\mathbf{v}, \mathbf{p}) = \mathbf{v} \cdot \mathbf{p} = \sum v_i p_i.$$

Thus, exploiting again that $p \in \{-1; 1\}$, the maximal payoff a voter could get from her optimal party is given by the 1-norm, i.e., by the following formula:

$$|\mathbf{v}|_1 := \sum_i |v_i|.$$

Finally, in order to state our decision rule for approval voting, we need the voter's approval threshold, stating how much a candidate can deviate from the optimum before loosing approval. To do so, we fix an *approval coefficient* $k \in [-1; 1]$. The lower this coefficient, the more tolerant a voter is towards deviations from her optimal position. We can thus formulate the following decision rule:

> *Approval Voting:* Let $\mathbf{v}$ be a voter with approval coefficient $k \in [-1; 1]$. Then $\mathbf{v}$ approves of all parties $\mathbf{p}$ that satisfy
>
> $$\mathbf{p} \cdot \mathbf{v} = \sum p_i v_i \geq k \cdot \sum |v_i|. \tag{5.1}$$

Before exploring the mathematical properties of this framework, we take a moment to analyze this definition a bit further and offer some remarks. First,

[7] As usual, $sign(x)$ is 1 if $x \geq 0$ and -1 else.

[8] Here, the commitment may again reflect the importance $\mathbf{v}$ attaches to that topic as well as her uncertainty about the right course of action. See footnote 4 for more details.

note the subtle dependency on the approval coefficient k. For the extreme value of $k = 1$, the candidate will only approve of an optimal party sharing her inclination on every topic. If there is no such party, the voter will submit an empty approval set, i.e., abstain. Conversely, a voter with an approval coefficient of -1 will indiscriminately approve of every party, no matter what that party claims, wants or does. Finally, a middle value of $k = 0$ corresponds to a fairly tolerant voter, approving of every party that agrees with her more often than it disagrees. For most of the following applications we will thus assume that $k \geq 0$.

Next, we examine two alternative intuitions of how a voter could choose which parties to approve of, the first resting on a grading system, the second on geometrical proximity. As it turns out, both of these alternatives are equivalent to our choice rule. We take this as an argument for the naturalness of our definition. The first alternative choice rule is given in terms of a grading system, resting on percentual agreement. An agent chooses a percentual threshold $t \in [0; 100]$ and approves of every party that agrees with her on at least t percent of the topics. Since the voter has different degrees of commitment to the various agenda items, this percentual agreement needs to be weighted with the agent's commitments $|v_i|$. Thus, the corresponding rule is:

> *Approval Voting 1^{st} Alternative:* Let $\mathbf{v}$ be a voter with percentual threshold $t \in [0; 100]$. Then $\mathbf{v}$ approves of all parties $\mathbf{p}$ that satisfy
>
> $$\frac{1}{\sum |v_i|} \sum_{p_i v_i > 0} |v_i| \geq \frac{t}{100}.$$

As the following, straightforward, lemma shows, this decision rule is equivalent to our original decision rule.

Lemma 5.4. *A voter $\mathbf{v}$ approves of some party $\mathbf{p}$ with approval coefficient $k \in [-1; 1]$ if and only if she approves of $\mathbf{p}$ in the alternative definition with percentual threshold $t = 100 \cdot \frac{1+k}{2}$.*

The second alternative we consider is of a geometric nature. Recall that we represent voters and parties through their positions on the agenda items, that is, as a vector in $\mathbb{R}^n$. So why not define the voter's approval decision through geometric proximity? Arguably, an adequate measure for proximity is the angle between two position vectors, showing how far these two diverge in their political opinions.[9] The maximal angle of 180° between a voter $\mathbf{v}$ and

[9] To elaborate a bit further on why we take the angle between two vectors and not, for instance, their length: Recall that a change in length of some vector $\mathbf{v}$, that is, replacing $\mathbf{v}$ by

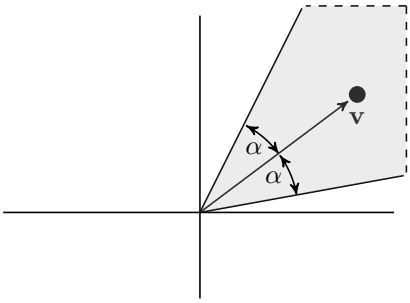


Table 5.3: The approval cone of voter $\mathbf{v}$ (shaded).

some party $\mathbf{p}$ means that $p_i \cdot v_i \leq 0$ for every i, that is $\mathbf{v}$ and $\mathbf{p}$ disagree about every single topic. Conversely a relatively small angle between a party and a voter corresponds to a high degree of agreement between the voter's inclination and the party position, see table 5.3. Again, we need to fix a threshold angle α for formulating the corresponding decision rule. For some given threshold angle α, let $\mathcal{C}(\mathbf{v}, \alpha)$ be the cone of all vectors $\mathbf{y}$ in $\mathbb{R}^n - \{0\}$ such that the angle between $\mathbf{v}$ and $\mathbf{y}$ is at most α.

Approval voting 2^{nd} alternative Let $\mathbf{v}$ be a voter with threshold angle $\alpha \in [0; 180]$. Then $\mathbf{v}$ approves of all parties $\mathbf{p}$ that satisfy

$$\mathbf{p} \in \mathcal{C}(\mathbf{v}, \alpha).$$

Again, this alternative is closely related to our original decision rule. This time, however, the exact relationship between the approval coefficient k and the threshold angle α depends upon the exact position of voter $\mathbf{v}$. Our correspondence is:

Lemma 5.5. *Let $\mathbf{v}$ be a voter and let k be as in the definition of approval voting. Then there is some angle α depending upon n, k and $\mathbf{v}$ such that for each party $\mathbf{p}$ holds*

$$\mathbf{p} \in C(\mathbf{v}, \alpha) \Leftrightarrow \mathbf{p} \cdot \mathbf{v} \geq k \cdot |\mathbf{v}|_1.$$

Furthermore, the angle α satisfies $\arccos(k) \leq \alpha \leq \arccos(\frac{k}{\sqrt{n}})$.

Thus, for any possible voter $\mathbf{v}$, the three different possible interpretations of approval coefficients are equivalent to each other. Before proceeding to some

$\lambda\mathbf{v}$ for some $\lambda > 0$, simply denotes a change in political commitment while leaving the general position intact. On the other hand, a non-zero angle between two voters $\mathbf{v}$ ad $\mathbf{v}'$ implies that the two disagree about the relative importance attributed to the various topics or even about the right course of action about some agenda item i.

general results, we revisit the worries about the AGW definition of approval voting raised in the last section. To recall, our first worry was that the AGW account of approval voting introduced intuitions from multi-winner elections into a general framework that rests upon a single winner reading. The second worry was that the AGW decision rule violates the minimal consistency condition that a voter should approve of the outcome of any election where all voters act exactly like her. For the present approach, we address both worries simultaneously. In our decision rule, the voter evaluates each party individually as to whether or not it is worthy of approval. In this evaluation, no reference to other parties is made. Hence the decision rule does not rely on any element, intuition or mechanism from multi-winner elections and our decision rule avoids the first criticism. Almost the same argument applies to our second worry. The approval set of an agent only contains parties the voter approves of *individually*. If every voter happened to submit the same approval set as $\mathbf{v}$, the winner would be some member of this approval set, and thus a party $\mathbf{v}$ approves of.

Finally, as a slight generalization of our framework, we introduce a third class of voting systems that we wish to include in our analysis. We will show that this class of voting systems can be treated with tools similar to the ones developed for approval voting. *Range voting* refers to an entire family of voting system. The underlying idea behind all of these is that voters are asked to grade candidates within some given scale of grades. The systems then differ in how they determine the winner of an election.[10] For our present purposes we can safely ignore this problem, since the payoff of a voter is already determined by the ballot she submits, independent of any outcome of the election. In order to define our account of range voting, we first fix a set of grades $g_1, \ldots, g_n$ with g_1 being the worst grade and g_n the best. Also, we need to fix a set of *grade requirements* $-1 = t_1 \leq \ldots \leq t_n \in [-1; 1]$ for these grades. Then we can define our definition rule for range voting as:

> *Range voting* Let $\mathbf{v}$ be a voter and let $-1 = t_1 \leq \ldots \leq t_n \in [-1; 1]$ be her grade requirements. Then the grade some candidate $\mathbf{p}$ receives is given by
>
> $$grade(\mathbf{v}, \mathbf{p}) := \max \{i | \mathbf{v} \cdot \mathbf{p} \geq t_i |\mathbf{v}|_1\}$$

Also here, some remarks are in place. First, note that approval voting is a special case of range voting with only two possible grades, *approval* and *disapproval*. Thus, the set of candidates some voter approves of under approval

[10] For instance if the grade scale is numerical, some systems rank candidates according to their median values, others use average grades instead.

voting is exactly the set that she grades with *approval*, the higher of the two possible grades. The following, straightforward lemma shows that this analogy is compatible with our formal decision rules for range voting and approval voting.

Lemma 5.6. *Assume there are two grades g_1 and g_2 and let* $\mathbf{v}$ *be a voter with grade requirements $-1 = t_1 < t_2$. Then* $\mathbf{v}$ *grades any candidate* $\mathbf{p}$ *with the maximum grade g_2 if and only if she approves of* $\mathbf{v}$ *with approval coefficient $k = t_2$.*

Finally, we note that the two alternative interpretations for approval voting we have presented, percentual agreement and geometrical proximity, can be extended to range voting, using analoga of Lemmas 5.4 and 5.5 respectively. In the first case, percentual approval, this gets us even closer to grading as known from school contexts. For each grade step, a certain percentage of agreement between a voter and a party is necessary. In other words, the party neeeds to score a certain number of points on the political agreement scale of that voter. More interestingly, the second alternative account, geometric proximity, translates grade requirements into geometrical objects. Instead of a single approval cone, each set of grade requirements $-1 = t_1 \leq \ldots \leq t_n$ translates into a sequence of ever narrower cones around the voter $\mathbf{v}$.

$$\mathcal{C}(\mathbf{v}, \alpha_1) \supseteq \ldots \supseteq \mathcal{C}(\mathbf{v}, \alpha_n).$$

Here, the indices of the cones stand for different grades, that is $\mathcal{C}(\mathbf{v}, \alpha_i)$ depicts the area a party needs to fall in for receiving grade *i or higher*. The actual grade some party receives is thus the index of the *narrowest* cone it is contained in or, equivalently, the number of different cones it falls in.

5.5 Results

In this section, we will explore the different voting systems presented in the previous section and their propensity to foster a high electoral turnout. We begin our discussion by clarifying the relationship between approval voting and plurality vote. As we have argued above, the two voting instruments appeal to two different reasoning strategies, maximizing and satisficing, that is identifying an optimal solution or setting an approval threshold and going with all solutions above that threshold. Despite this difference in reasoning modes, there are some coherence conditions connecting both accounts. It might happen that, under approval voting, some voter finds exactly one party to be acceptable. Thus,

the ballot she submits, putting forward only one party, could also occur under plurality vote. It seems reasonable to demand that under plurality vote she would vote for the same party. That is, *given that* $\mathbf{v}$ *submits a single party ballot*, this ballot should be the same under approval voting and plurality rule. Of course, the implicit statements contained in her ballot are different in the two voting systems: In one case $\mathbf{v}$ merely expresses that her candidate $\mathbf{p}$ is the best party, while in the approval case she additionally utters that all other parties are unacceptable to her. We merely demand that the party identified as best be the same in both cases. The following lemma shows that this indeed holds true, save for the possibility of abstentions.

Lemma 5.7. *Let* $\mathbf{v}$ *be a voter. Assume that under approval voting* $\mathbf{v}$ *approves of the set* $\{\tilde{\mathbf{p}}\}$ *while under plurality vote she votes for* $\mathbf{p}'$*. Then* $\mathbf{p} = \tilde{\mathbf{p}}$*.*

Having introduced our formalism and the different voting rules, it is time to come back to our initial question: "When do people vote?". Within the expressive framework we are working in, the first tentative answer to this question is easy. People vote if they find some possible ballot they prefer over abstaining. But when is that the case? Or more concretely: How likely is it that some voter finds a ballot she prefers over abstaining? Our results will fall in two groups, roughly corresponding to Theorems 1 and 2 of [6]. The first class of results asks about the minimal number of parties needed to avoid abstentions. Writing a party program is most often a strategic exercise. Parties have to balance various interest groups among their supporters, while also keeping an eye on the wishes and desires of the general electorate. Further, the chances of winning an election will also depend on the competing candidates. In some cases, it might be more promising to be the only party catering to a small fraction of the electorate rather than competing with four other parties about a larger group of voters. In the most extreme case, studied in [102] , the different parties reposition themselves constantly in order to maximize electoral support. For the present purpose, we are a bit more cautious, being only interested in a strategically chosen but static party position. For a new party, it might be a promising strategy to avoid competition and focus on those voters that are not yet attracted by any of the existing candidates. Thus, a reasonable question to ask is how many strategically positioned parties are needed to guarantee that every voter finds some candidate she prefers over abstaining. For plurality vote, this question has been answered by Theorem 1 of [6], cited above. The number of parties necessary to ensure that no voter abstains under plurality vote is exponential in the number n of agenda items.

In the case of approval voting, the answer to this question will depend upon the tolerance of the voters. Naturally, the more demanding voters are, that is, the higher their approval threshold, the more candidates are needed in order to ensure that every voter finds a suitable candidate to approve of.

Theorem 5.8. *i) If $k \leq 0$ two parties are enough to ensure that every voter approves of at least one party.*
ii) If $k > 0$ the number of parties needed to ensure that no (possible) voter abstains is exponential in the number of topics.
iii) Assume there are at least three topics on the agenda. If voters are infinitesimally more demanding than $k = 0$, approving only of those parties that, given the voters' weights, share strictly more than half of their position, i.e., $\mathbf{p} \cdot \mathbf{v} > 0$, then exactly $n + 1$ parties are needed to ensure that every voter approves of at least one party.

Next, we consider the opposite extreme. Consider for a moment a completely idealistic candidate that has only decided to run in the elections to see her own position represented. Such a candidate will obviously not refer to tactical considerations about her competitors or the distribution of voters. She will simply report her true opinions about all the agenda items. Such candidates may be healthy for the political system, however they are bad news for avoiding abstentions. No matter how many of these candidates are available, we cannot *guarantee* that every voter will find some candidate that is more attractive than abstaining. However, we can give probabilistic estimates how likely it is that a random voter is attracted by at least one of these candidates. As the number of topics on the agenda grows, it becomes more and more likely that some group of people is dissatisfied by all the existing candidates and hence decides to form their own party. We will thus assume for our analysis that an election with n different topics on the agenda attracts n such randomly distributed parties.[11] Under these conditions, let $P(n)$ denote the probability that some voter does not abstain in an election based on an n-topic agenda. Again, the case of plurality vote has been analyzed in [6] in Theorem 2, cited above: Under plurality vote we have $\lim_{n\to\infty} P(n) = 0$.

In the case of approval voting, the chance that some randomly chosen party is appealing to some generic voter $\mathbf{v}$ will depend upon the tolerance of $\mathbf{v}$. Naturally, a voter with a high approval coefficient k is more likely to abstain in such a situation than somebody with lower standards of approval. Thus, we extend

[11] That is, we draw each party's position from a uniform distribution over the 2^n possible positions.

our above definition of $P(n)$ to $P(n,k)$, denoting the probability that a random voter with approval coefficient k does not abstain in an n-topic situation. We have the following result:

Theorem 5.9. *For $k \leq 0$ we have $\lim_{n\to\infty} P(n,k) = 1$. On the other hand for $k \in (0;1]$ the converse holds: $\lim_{n\to\infty} P(n,k) = 0$*

For this result, a word of caution is in place. The skeptic result that a sufficiently demanding voter will almost surely abstain if the agenda is big enough, i.e., $\lim_{n\to\infty} P(n,k) = 0$, is only a worst case result, depending on the exact interest of the voter. Naturally, a universally interested voter, having opinions on most of the agenda items, is harder to accidentally satisfy than somebody who is only interested in a small selection of the agenda. In the extreme case, a voter is primarily focused on a single topic, that is $\mathbf{v} \approx \pm e_i$. Such a voter will almost surely find some party she approves, i.e., $\lim_{n\to\infty} P(n,k) = 1$ independent of her approval coefficient k.

Finally, we come back to range voting, the last voting mechanism we have introduced. In range voting, voters are asked to grade all parties within a given grade scale. Thus, there is nothing such as abstaining or submitting an empty ballot and we need to reformulate our original question. So let's assume a voter is motivated to engage in range voting just if she finds some relevant *differences* to be expressed in terms of grades. That is, we ask for the conditions under which our voter finds two parties that receive different grades. Just as in the case of approval voting, this will depend upon the exact grade requirements she has. For the case of strategically positioned parties we get:

Theorem 5.10. *Assume that every voter has some i with $t_i = 0$ and that there are at least three topics on the agenda. Then $n+1$ parties are enough to ensure that every (possible) voter finds two parties she grades differently. Conversely, if there is no such index i with $t_i = 0$, the number of parties needed to ensure that every possible voter finds two parties she wishes to grade differently grows exponentially in n.*

Second, we consider again the case of randomly distributed parties. We will use the same setting as for approval voting, that is we consider the case of n randomly distributed parties in an election ranging over an agenda with n topics. As above, our results will depend on the exact grade requirements used by the voter. For any set $\mathbf{t} = t_1 \dots t_n$ of grade requirements, let $P(n,\mathbf{t})$ denote the probability that, given n randomly distributed parties on an n-topic

agenda, a voter with grade requirements $\mathbf{t}$ will find two parties that she wishes to grade differently.

Theorem 5.11. *If there is some i with $t_i = 0$ then $\lim_{n\to\infty} P(n, \mathbf{t}) = 1$. On the other hand, if there is no such i then $\lim_{n\to\infty} P(n, \mathbf{t}) = 0$.*

With other words, the probability that $\mathbf{v}$ finds two random parties she wishes to grade differently crucially depends on whether she gives different grades to those parties coinciding with her opinion on at least half of the topics and those who do not. As a last remark, note that the same word of caution as for approval voting applies. Theorem 5.11 studies a worst case scenario that mainly applies to universally interested voters. Again, a less universal voter, primarily interested in one or two agenda items, will satisfy $\lim_{n\to\infty} P(n, \mathbf{t}) = 1$ independently of which grade requirements she uses.

5.6 Focus and Dynamics

The analysis of expressive voting we have presented so far is completely atemporal. In this section, we will add a dynamic element to our analysis by studying various changes in the electoral preference. Up and until now, we have focused on the single, decisive moment in which a voter submits her ballot. But this is, of course, not all there is to an election. The actual poll is preceded by a long and sometimes tedious period of campaigning during which each party tries to convince the electorate of their advantages. Within this period, voters form their opinions on whom to vote for. While gradually learning about the political landscape, the voters will sometimes update their beliefs about which candidate is best. We identify three different reasons that could trigger a change in some voter's preference order. First, a voter may obtain some new information about the standpoint of some candidate on a particular topic she cares about. Second, the voter might change her own position, either by changing her mind on some particular topic or by updating the relative importance she attributes to the individual topics. And third, during the several weeks of campaigning, some entirely new topic may appear on the agenda influencing the voter's preference orders. In their attempt to maximize electoral support, the different parties might appeal to all of these three mechanisms. They might, first, change their own position, or be intentionally opaque about some of their stances. Second, parties may, of course, try to convince voters of their stances towards some agenda items or, at least, try to direct the voters attention into a favorable direction. And third, candidates can try to place some entirely new topic on

the agenda. In this section, we explore several extensions of our model to pre-electoral dynamics and their strategic potential. We mainly concentrate on the first two of the above mechanisms, parties strategically relocating themselves on the agenda as well as their attempt to change the voters' interests. The first part of this section deals with strategizing parties, deliberating whether and how to determine or change their own positions in order to maximize electoral support. This case has been dealt with by Walter Dean and Rohit Parikh in [46], using a logical framework. Our main aim in the first part of this section is to unify the two models and demonstrate how their framework can be translated into our approach. In the second part of this section, we will study changes in the preference of voters. In many cases, the attention of voters is not distributed equally among all agenda items, but some topics are more in focus having a bigger impact on the voter's decision making. The main contribution of this section will be to outline a formal framework for focus and the evolution of focus over time. Finally, we will inquire into some topics and subtleties arising when different parties and candidates try to change the voters' focus strategically in order to maximize electoral support.

While inquiring into parties and voters changing their positions or preferences, we will always assume that the underlying landscape, the agenda of topics itself, remains constant. That is, we will not deal with new topics entering the agenda, the third reason for preference change identified above. However, we wish to emphasize that this would not pose a major problem to our account. As the following lemma shows, some voter's decision remains the same, whether a topic receives a weight of 0 or whether that topic is removed from the agenda altogether. Conversely, adding a topic to the agenda is technically equivalent to assuming that is has always been there with a weight of zero.

Lemma 5.12. *Under approval voting, let $K \subseteq \{1 \dots n\}$ be a subset of the agenda items. For any party $\mathbf{p}$ let $\mathbf{p}_K$ be the restriction of $\mathbf{p}$ to K and likewise for voters. Fix some $\mathbf{v}$ and assume that $v_i = 0$ for all $i \notin K$. Then*

$$\frac{\mathbf{v} \cdot \mathbf{p}}{|\mathbf{v}|_1} = \frac{\mathbf{v}_K \cdot \mathbf{p}_K}{|\mathbf{v}_K|_1}.$$

In the first part of this section, we focus on the strategic position and repositioning of different parties. Which possible utterance is best suited for fostering electoral success depends on a variety of features. Of course, the voting system will play a role, as will the positions of voters and competing candidates. Walter Dean and Rohit Parikh identify two additional aspects relevant for picking an optimal utterance. First, the choice depends on how voters will update

their information in light of newly learned evidence. This question relates to the theory of information change and related tools such as, for instance, AGM belief revision function [3]. The second important feature of a voter is how she evaluates her information about a candidate. Dean and Parikh consider three different types of voters. Optimistic voters interpret all remaining uncertainty in favor of a candidate, while pessimistic voters hold all uncertainty against the candidate. Finally, expected value voters represent a party by the expected value of the uncertainty set. To make this intuition precise, suppose that I is a set of information a voter has about some party $\mathbf{p}$. For $\mathbf{x} \in \{-1;1\}^n$ we write $\mathbf{x} \models I$ to denote the fact that all the information contained in I is consistent with the position of $\mathbf{p}$ being $\mathbf{x}$. Further, let $u_{\mathbf{v}}(\mathbf{x})$ be the degree of approval a voter $\mathbf{v}$ would have towards a party with position $\mathbf{x}$, thus $u_{\mathbf{v}}(\mathbf{x}) = \mathbf{v} \cdot \mathbf{x}$. Then,[12] Dean and Parikh define the following voter types:

$$\begin{array}{lrcl} \text{(optimistic voter)} & s^{opt}_{\mathbf{v}}(I) & = & \max\{u_{\mathbf{v}}(\mathbf{x}) \mid \mathbf{x} \models I\} \\ \text{(pessimistic voter)} & s^{pes}_{\mathbf{v}}(I) & = & \min\{u_{\mathbf{v}}(\mathbf{x}) \mid \mathbf{x} \models I\} \\ \text{(expected value voter)} & s^{ev}_{\mathbf{v}}(I) & = & \frac{\sum_{\mathbf{x} \models I} u_{\mathbf{v}}(\mathbf{x})}{|\{\mathbf{x} \mid \mathbf{x} \models I\}|} \end{array}$$

These voter types determine when a voter $\mathbf{v}$ will vote for some candidate in the different voting schemes. Under approval voting with some approval coefficient k, an optimistic voter with information I about some party $\mathbf{p}$ will thus approve of $\mathbf{p}$ iff $s^{opt}_{\mathbf{v}}(I) \geq k \cdot |\mathbf{v}|_1$, a pessimistic voter will approve of $\mathbf{p}$ iff $s^{pes}_{\mathbf{v}}(I) \geq k \cdot |\mathbf{v}|_1$ and, likewise, an expected value voter will approve of $\mathbf{p}$ if $s^{ev}_{\mathbf{v}}(I) \geq k \cdot |\mathbf{v}|_1$. Just as in the case of approval voting, it can be instructive to gain a geometric understanding of voter's uncertainty and its relation to the electoral decision. For this end, let $\mathbf{x}_1 \dots \mathbf{x}_n \in \{-1;1\}$ be the set of party positions that are consistent with the current knowledge some voter $\mathbf{v}$ has about $\mathbf{p}$, that is, $\{x_1 \dots x_n\} = \{\mathbf{x} \mid \mathbf{x} \models I\}$. Furthermore, we denote the set of all the intermediate positions between the (necessarily extreme) stands $\mathbf{x}_1 \dots \mathbf{x}_n$ by Δ, that is $\Delta \subseteq [-1;1]^n$ is the convex hull of $\{\mathbf{x}_1 \dots \mathbf{x}_n\}$. Then we get the following geometric equivalent of the above decision rule:

Lemma 5.13. *Let α be the approval angle corresponding to voter $\mathbf{v}$. Then:*
i) An optimistic voter approves of $\mathbf{p}$ iff $\Delta \cap \mathcal{C}(\mathbf{v}, \alpha) \neq \emptyset$
ii) A cautious voter approves of $\mathbf{p}$ iff $\Delta \subseteq \mathcal{C}(\mathbf{v}, \alpha)$
iii) An expected value voter approves of $\mathbf{p}$ iff $\mathbf{c} \in \mathcal{C}(\mathbf{v}, \alpha)$, where $\mathbf{c}$ is the gravitational center of Δ.

[12] The presentation is adapted to our framework. In the original paper, Dean and Parikh allow for any utility function based on the agenda items individually.

Thus, we have shown that the logical approach chosen by [46] can be incorporated into our framework and that their decision rule for filling uncertainty fits in naturally with all the three different approaches agents could have for choosing their approval threshold.

Having studied the strategic positioning of parties, we now leave the framework of [46] and return to the voters' side. During the period of electoral campaigning, voters' standpoints on the agenda are not carved in stone. Just to the contrary, individual voters constantly update their positions, reacting to all kind of incoming information. First of all, the voter may learn about some external events, recent developments in world politics or newly published economic data that impact her belief about what constitutes the best course of action. Second, she might also change her standpoint as a result of some deliberative process, discussions among her peers, following the news or simply pondering about her own values. Independent of the reasons for doing so, any change in the voters' preferences might naturally impact the electoral outcome. Hence a strategically acting party might try to influence voters' positions in their own interest. Of course, some voters may be hard to influence and the candidates will only have limited control over the various external events that could change the voters' positions. Yet, there are at least two different ways in which the candidates could go. First, a party can actively try to persuade voters of its position, hoping that some disagreeing voters gradually change standpoint on that topic. This process, however, is cumbersome, slow and rare in occurrence. There is a second strategy that looks more promising. In making their electoral decision, few voters will refer to the entire spectrum provided by the agenda. Most voters will rather *focus* on the two or three prominent topics they regard as most pressing for the near future. *Which* topic it is that voters focus on, however, may change several times during an election campaign. Public focus is not a very stable phenomenon, a fact that the different candidates might try to exploit. In general, the individual candidates will attempt to steer public attention towards topics where they have favorable opinions, trying to stay clear of areas where they could easily put off too many voters. In the following, we will study the influence of public *focus* and its dynamics. To start, consider the following example:

During the 2011 German state elections in Baden-Württemberg, it appeared that the governing Christian Conservatives would easily remain in power. The party's position on nuclear energy did not quite match the majority opinion, but most voters were focused on different issues. Then, on March 11, a tsunami

hit the Japanese province of Tohoku, causing a major nuclear incident at the Fukushima Daiichi Power Plant. Suddenly, nuclear energy was on everyone's mind. This had a drastic effect on the elections: After nearly 60 years of governing, the Christian Conservatives were swept out of office by a Green Left coalition (which strongly opposed nuclear energy).

We can model the German 2011 state election scenario using the framework discussed in the previous section. The set of issues is $I = \{i_1, i_2, i_3, i_4\}$ with[13]

i_1: "We must support the car industry."
i_2: "We should be conservative about public spending."
i_3: "We ought to continue nuclear energy."
i_4: "Do not increase funding for education."

In this framework, the Conservatives would be represented by a vector $\langle 1, 1, 1, 1\rangle$, while their two main opponents had a -1 on p_3 and also on some of the other items. Say that the Social Democrats are represented by $\langle 1, -1, -1, -1\rangle$ and the Greens by $\langle -1, 1, -1, -1\rangle$.

A typical voter from the Southwest emphasizes industry and/or education but displays only a relatively small concern about nuclear energy. For instance, the following two profiles represent typical voters in this area: $\mathbf{v}_1 = (0.8, 0.9, -0.3, 0.4)$ or $\mathbf{v}_2 = (0.4, 0.8, -0.3, 0.9)$. Under normal circumstances, this would lead to a crushing victory for the Christian Conservatives, using any of the voting methods discussed in the previous section. However, as stated above, the Fukushima Power Plant incident changed the voters' focus.

Arguably, a change in focus will, in general, not change the voter's attitude towards some agenda item i, i.e., the sign of the particular position. Focus change corresponds to a momentary redistribution of attention or importance attributed to the various topics, measured by the length $|v_i|$. A change in attitude, i.e., in $sign(v_i)$ on the other hand would require some significant engagement with that particular topic, brought about by different mechanisms than a mere change in focus that we do not want to discuss here. Consequentially, a focus change will only affect the length $|v_i|$ of some topic vector, while leaving $sign(v_i)$ intact. We think of a change in focus as a linear transformation of the space of positions for each voter. This suggests the following definition:

Definition 5.14 (Focus Matrix)**.** A **focus matrix** is a diagonal matrix $A \in [0, 1]^{n\times n}$ (i.e., for all $1 \leq i, j \leq n$, if $i \neq j$, then $A_{ij} = 0$). Voter $\mathbf{v}$'s position

[13]Some people claim that a significant number of voters originally based their decision on a fifth issue i_5: "This party has been in office for the last 60 years". We do not wish to comment on this claim here. We note, however, that our framework is rich enough to incorporate such considerations.

after a focus change with A, denoted $\mathbf{v}^A$, is calculated in the standard way using matrix multiplication. ◁

The following is a possible focus change matrix triggered by the Fukushima incident:

$$A_{Fuku} = \begin{pmatrix} 0.05 & & & 0 \\ & 0.05 & & \\ & & 1 & \\ 0 & & & 0.05 \end{pmatrix}.$$

Clearly, this will make nuclear energy the focus of attention for all voters. After applying this focus change to the two voters mentioned above, the resulting position vectors are $\mathbf{v}_1^{A_{Fuku}} = (0.04, 0.045, -0.3, 0.02)$ and $\mathbf{v}_2^{A_{Fuku}} = (0.02, 0.04, -0.3, 0.045)$. Such voters would end up supporting either the Social Democrats or the Green party.

The above example shows that redirecting the voters' focus is a powerful tool that can drastically change the outcome of an election. And indeed, as any political pundit will report, much of the rhetoric during an election is aimed at trying to focus the attention of voters on certain sets of issues. The genesis of public attention, however, is a complex matter. As our last example showed, focus is influenced by the general situation and external events the agents cannot influence. But it is also shaped by news coverage, the content of electoral campaigns and other factors that are at least partially under the control of parties or sympathizing groups. What makes the analysis of focus even more complex is that it does not suffice to study the different focusing attempts individually. The various messages from different interest groups interact in a complex way, potentially leading to unexpected and unintended results. We conclude this section with a number of examples to illustrate the subtleties involved in influencing the focus of a group of voters.

Example 5.15: Suppose that there are two parties $T = \{d, r\}$ competing in a two-topic election (i.e., $I = \{i_1, i_2\}$). The two parties have completely opposing views on both topics, say $\mathbf{p}^d = (1, 1)$ and $\mathbf{p}^r = (-1, -1)$. Suppose that almost half of the voters are clearly in favor of the second candidate, $\mathbf{p}^r$. The rest of the voters are relatively undecided, not feeling that either of the parties is particularly close to their views. This example shows that there is a way to focus the voters so that the first candidate, d, is the winner.

To make things more concrete, suppose that there are three voters, described by: $\mathbf{v}_1 = (-1, -0.8)$, $\mathbf{v}_2 = (-1, 0.7)$ and $\mathbf{v}_3 = (1, -0.62)$. Clearly, d will lose the election given these voters. However, d can win a plurality election by changing the voters' focus using the following matrix:

$$\begin{pmatrix} 0.65 & 0 \\ 0 & 1 \end{pmatrix}.$$

However, it does not suffice to direct voters attention to only one of the two issues. If voters focused on either of these topics alone, r would still win, having the support of $\mathbf{v}_1$ and, depending on the topic, either $\mathbf{v}_2$ or $\mathbf{v}_3$. Thus, a strategic focus campaign might not only need to identify the right focus set, but also balance the attention attributed to different topics.

Example 5.16: Suppose that there are three candidates $T = \{d, m, r\}$ and six issues $I = \{i_1, \ldots, i_6\}$. Assume that d is in favor of all the topics, $\mathbf{p}^d = (1, 1, 1, 1, 1, 1)$, and r opposes all the topics, $\mathbf{p}^r = (-1, -1, -1, -1, -1, -1)$. The candidates' campaign staffs have determined that d maximizes its share of votes if the voters focus on i_1, i_2 and i_3, while r receives the maximum support when the voters are focused on i_4, i_5 and i_6. In both cases, the maximum support among the voters is enough to win the election using plurality rule. In planning their campaigns, the candidates might try to guide public attention to the different focus sets $\{i_1, i_2, i_3\}$ and $\{i_4, i_5, i_6\}$ respectively. However, this may lead to a situation in which a third candidate m wins a plurality vote. Thus, the simultaneous attempt of two parties to influence focus might benefit a third party, even without that party engaging in any tactical endeavors at all. To give this example an extra twist: It is possible that, had party d foreseen r's focusing efforts, d could have won the election by omitting topic i_1 from her focus campaign. Hence, it is not sufficient to study focus changing events individually, but it is equally important to consider other competing efforts and their interaction.

To fill in the remaining details, suppose that m supports only issues i_3 and i_6 ($\mathbf{p}^m = (-1, -1, 1, -1, -1, 1)$. There are three voters with $\mathbf{v}_1 = \mathbf{v}_2 = (-0.25, 0.3, 1, -0.1, -0.1, -0.1)$ and $\mathbf{v}_3 = (1, -1, 0.9, 1, 1, 1)$. Now, it is not hard to see that this model satisfies all the properties claimed. That is,

1. In an election in which the voters are focused primarily on the sets i_1, i_2 and i_3, party d would win.
2. In an election in which the voters focus only on i_2 and i_3, party d would still win, but with *less* votes than if voters focused on all of i_1, i_2 and i_3. Thus, only the latter set maximizes support.
3. In an election in which the voters are focused primarily on i_4, i_5 and i_6, party r would win.

4. In an election in which the voters are evenly focused on all the issues $i_1, \ldots, i_6$, m would win. However, if none of the voters focuses on i_1, then d would win the election.

5.7 Discussion and Outlook

The theory of expressive voting is an attractive alternative to the instrumental analysis of voting behavior. Not least, because expressive accounts offer a solution towards problems left open by the instrumental account such as the question of why people vote. As the reader might rightly remark, neither alternative alone can provide a satisfactory account of much voting behavior we observe. Rather, any real voter will encompass instrumental and expressive considerations and potentially many other reasons simultaneously. Yet, we hold that a thorough analysis of expressive voting is an important step on the pathway towards an integrated understanding of voting behavior.

In this chapter, we have presented and explored a formal framework for expressive voting based upon an agenda of items. Both voters and parties position themselves on this agenda, and voters evaluate the different parties by their positions. In our analysis, we focus on two aspects of voting, participation and dynamics. The first of these aspects, participation, is related to Downs' paradox of voting, the questions why voters vote at all. Within our framework, we have analyzed three different voting systems, plurality vote, approval voting and range voting with respect to the likelihood of abstentions. The central result of section 5.5 is that the latter two systems, approval and range voting, are exponentially better in promoting a high degree of participation than the first system, plurality vote. For the second aspect, dynamics, we have studied changes in the voter's preference profile over time. More specifically, we have concentrated on focus as a major reason for preference change. In their electoral decision, many voters are heavily guided by two or three topics that appear especially prominent to them. If the topics in focus change, some voters' decisions on whom to vote may change accordingly. Our central result in section 5.6 was to outline a formal framework for focus change through focus changing matrices and to identify some of the subtleties and difficulties connected to strategically manipulating the voters' focus. For instance, it might not be enough to focus on individual topics and omit others, but a well-chosen strategic focusing action might need to address several topics at once while balancing the relative attention they receive. Or, to give a second example, strategic influences on focus by several parties may interact in unexpected ways and potentially lead

to completely unexpected and unwanted outcomes. Finally, a third contribution of this chapter is to integrate and expand two existing pieces of research on this field. While our basic formal framework is taken from the first of these, [6], we have identified a crucial shortcoming in their treatment of approval voting and subsequently presented a different treatment avoiding this criticism. The second paper, [46], is a logical model of parties strategizing their position in order to maximize support. There, we have shown that this model can be translated naturally to our account.

We conclude by briefly discussing two directions of future research. As we have shown in section 5.3, the model presented here rests on the assumption that elections produce a single winner. This covers many important cases such as various presidential elections. There is, however, a second class of elections, parliamentary elections, that produce an entire class of winners, the parliament that in turn will need to form a government. A first direction of future research would thus be to extend the present model to multi-winner elections. We conjecture that much of this work can be done by relaxing our conditions on the positioning of parties. However details have to be worked out on both, a conceptual and a formal level. In particular, it is not yet completely clear how relaxing our conditions would impact on the results presented in section 5.5. Our second direction of future research is related to the strategic potential of candidates prior to elections. So far, we have concentrated on exploring the strategic tools parties have to maximize their support, for instance by adapting their own position. The actual potential a party has for credibly altering her position or influencing public focus will depend upon its political or social capital, including factors such as visibility, reputation and track record. We thus hold that an integrated model of pre-voting dynamics should be sensitive to the individual parties' resources. A resource sensitive model may be closely related to yet another potential direction of future research, *subjective* models of the political realm. The attribution of political capital to the different parties may, in general, differ from voter to voter. Thus a resource sensitive model may require some of our tools to be agent relative. For instance, *subjective* focus change matrices could track the focus of individual agents, while different ways of interpreting the announcements of candidates may lead to subjective theories of what a party stands for.

5.8 Appendix: Proofs

We start by showing that example 5.3 satisfies all properties claims. In particular we have to show that *i)* $\text{dist}(\mathbf{v}, \mathbf{p}) < \text{dist}(\mathbf{v}, \mathbf{e}_\pm)$ and *ii)* that $\{\mathbf{e}_+, \mathbf{e}_-\}$ is the coalition approved by $\mathbf{v}$.

For *i)* observe that

$$\text{dist}(\mathbf{v},\, \mathbf{p}\,) = \sqrt{4 \cdot \left(\frac{2}{3}\right)^2 + 5} = \sqrt{\frac{16}{9} + 5} \quad \text{and}$$

$$\text{dist}(\mathbf{v}, \mathbf{e}_*) = \sqrt{2 \cdot \left(\frac{2}{3}\right)^1 + 2 \cdot \left(\frac{4}{3}\right)^2 + 5} = \sqrt{\frac{24}{9} + 5} \text{ for } * \in \{+, -\}$$

thus $\mathbf{e}_\pm$ are indeed the most extremist parties.

For *ii)* observe that

$$\text{dist}\left(\frac{1}{2}(\mathbf{e}_+ + \mathbf{e}_-), v\right) = \frac{2}{3}.$$

To see that this is the closest coalition we first show that any coalition C containing three or more members has a distance of at least $\frac{\sqrt{5}}{3}$ from $\mathbf{v}$. For any such coalition the last five entries of C are all at least $\frac{1}{3}$ (with the minimum reached if C consists of exactly three entries, one of them being $\mathbf{e}_-$). Thus $\text{dist}(C, \mathbf{v}) \geq \frac{\sqrt{5}}{3}$. A similar argument shows that for $C' = \{\mathbf{e}_+, \mathbf{p}\}$ holds $\text{dist}(C', \mathbf{v}) \geq \sqrt{5}$. Finally, for the coalition $\mathbf{C}''\{\mathbf{e}_-, \mathbf{p}\}$ we have

$$\text{dist}(C', \mathbf{v}) = \sqrt{2 \cdot \left(\frac{1}{3}\right)^2 + 2 \cdot \left(\frac{2}{3}\right)} = \frac{\sqrt{10}}{3}$$

thus finishing the proof.

Proof of Lemma 5.5**.** For $x, y \in \mathbb{R}^n$ the angle α between x and y is described by the following well-known equation

$$\frac{x \cdot y}{|x|_2 |y|_2} = \cos \alpha \tag{5.2}$$

where $|\mathbf{x}|_2 = \sqrt{\sum x_i^2}$ denotes the euclidean length. On the other hand inequality 5.1 can be transformed to

$$\frac{\mathbf{v} \cdot \mathbf{p}}{|\mathbf{v}|_1} \geq k$$

$$\Leftrightarrow \frac{\mathbf{v} \cdot \mathbf{p}}{|\mathbf{v}|_2 \sqrt{n}} \geq \frac{k}{\sqrt{n}} \frac{|\mathbf{v}_1|}{|\mathbf{v}_2|}.$$

Since $|\mathbf{p}|_2 = \sqrt{\sum_i 1} = \sqrt{n}$. This is exactly equation 5.2 for

$$\alpha = \arccos\left(\frac{k}{\sqrt{n}} \frac{|\mathbf{v}_1|}{|\mathbf{v}_2|}\right).$$

The last claim follows from the inequality

$$|x|_2 \leq |x|_1 \leq \sqrt{n}|x|_2$$

for all $x \in \mathbb{R}^n$. □

Proof of Lemma 5.7. Recall that under approval voting, $\mathbf{v}$ approves of $\mathbf{p}^*$ iff $\mathbf{v} \cdot \mathbf{p}' \geq k \cdot |\mathbf{v}|_1$, where k is $\mathbf{v}$'s approval coefficient. Since $\mathbf{p}'$ is the only party $\mathbf{v}$ approves of, we get[14]

$$\frac{\sum v_i p'_i}{\sum |p'_i|} = \max_{p \in T} \frac{\sum v_i p_i}{\sum |p_i|}.$$

On the other hand, the fact that $\tilde{\mathbf{p}}$ is the winner under plurality vote is expressed by the equation

$$\mathrm{dist}(\tilde{\mathbf{p}}, \mathbf{v}) = \min_{\mathbf{p} \in P} \mathrm{dist}(\mathbf{p}, \mathbf{v}).$$

Thus, we have to show that the following conditions holds for every party $\mathbf{p}^*$

$$\mathrm{dist}(\mathbf{p}^*, \mathbf{v}) = \min_{\mathbf{p} \in P} \mathrm{dist}(\mathbf{p}, \mathbf{v}) \Leftrightarrow \frac{\sum v_i p^*_i}{\sum |p^*_i|} = \max_{p \in P} \frac{\sum v_i p_i}{\sum |p_i|}.$$

Recall that $p_i \in \{-1; 1\}$ for each topic $i \in N$. Fix a voter $\mathbf{v}$. For any party $\mathbf{p}$ let $U_{\mathbf{p}} \subseteq \{1 \ldots N\}$ be defined by:

$$i \in U_{\mathbf{p}} \Leftrightarrow v_i \cdot p_i < 0.$$

Thus $U_{\mathbf{p}}$ is the set of indices where the signs of $\mathbf{v}$ and $\mathbf{p}$ disagree. Now we have

$$\begin{aligned}
\mathrm{dist}(\mathbf{v}, \mathbf{p}) &= \sqrt{\sum_i (v_i - p_i)^2} \\
&= \sqrt{n + \sum_i v_i^2 - 2 \sum_i v_i p_i} \\
&= \sqrt{n + \sum_i v_i^2 - 2 \sum_i |v_i| + 4 \sum_{i \in U_{\mathbf{p}}} |v_i|}.
\end{aligned}$$

Observe that only the last term depends on $\mathbf{p}$. Thus we have for any $\mathbf{p}, \mathbf{p}' \in P$:

$$\mathrm{dist}(\mathbf{p}, \mathbf{v}) \leq \mathrm{dist}(\mathbf{p}', \mathbf{v}) \Leftrightarrow \sum_{i \in U_{\mathbf{p}}} |v_i| \leq \sum_{i \in U_{\mathbf{p}'}} |v_i|.$$

[14] Recall that T is the set of parties.

On the other hand we have:

$$\sum_i v_i p_i = \sum_i |v_i| - 2 \sum_{i \in U_{\mathbf{p}}} |v_i|.$$

Thus also

$$\frac{\sum v_i p_i}{\sum |p_i|} \geq \frac{\sum v_i p'_i}{\sum |p'_i|} \Leftrightarrow \sum_{i \in U_{\mathbf{p}}} |v_i| \leq \sum_{i \in U'_{\mathbf{p}}} |v_i|.$$

□

Before we can prove theorems 5.8 and 5.9 we need the following lemma:

Lemma 5.17. *Let $m \in \mathbb{N}$. Then we have for any natural number n*

$$\frac{\sum_{k=\lceil n(\frac{1}{2}+\frac{1}{2m})\rceil}^{n} \binom{n}{k}}{2^n} \leq 2\left(\left(1+\frac{1}{2m}\right)^{-1}\right)^n. \tag{5.3}$$

Proof. For notational convenience we assume n to be even. First we show that for any natural number $i \in [0, \frac{n}{2m}]$ we have that

$$\binom{n}{\frac{n}{2}+i} \geq \left(1+\frac{1}{2m}\right)^n \binom{n}{\frac{n}{2}+\lceil \frac{n}{2m} \rceil + i}. \tag{5.4}$$

To this end observe that

$$\begin{aligned}
&\frac{\binom{n}{\frac{n}{2}+i}}{\binom{n}{\frac{n}{2}+\lceil \frac{n}{2m} \rceil + i}} \\
=&\frac{(\frac{n}{2}+\lceil \frac{n}{2m} \rceil + i)!(\frac{n}{2}-\lceil \frac{n}{2m} \rceil - i)!}{(\frac{n}{2}-i)!(\frac{n}{2}+i)!} \\
=&\frac{(\frac{n}{2}+i+1)\cdot(\frac{n}{2}+i+2)\cdot \ldots \cdot (\frac{n}{2}+\lceil \frac{n}{2m} \rceil + i)}{(\frac{n}{2}-\lceil \frac{n}{2m} \rceil - i+1)\cdot(\frac{n}{2}-\lceil \frac{n}{2m} \rceil - i+2)\cdot \ldots \cdot (\frac{n}{2}-i)} \\
=&\frac{\frac{n}{2}+1+i}{(\frac{n}{2}-\lceil \frac{n}{2m} \rceil - i)} \cdot \ldots \cdot \frac{\frac{n}{2}+\lceil \frac{n}{2m} \rceil + i}{\frac{n}{2}-i}.
\end{aligned}$$

Now it is easy to see that each of the quotients in the last formula is larger than $1+\frac{1}{2m}$, thus the entire product is larger than $(1+\frac{1}{2m})^n$ and 5.4 holds. In the following, let $\alpha := \left((1+\frac{1}{2m})^{-1}\right)^n$.

Repeatedly applying 5.4 gives us for all natural numbers j with $0 \leq j < \lceil \frac{n}{2m} \rceil$

$$\sum_{i=1}^{m} \binom{n}{\frac{n}{2}+j+i\lceil \frac{n}{2m} \rceil} \leq \sum_{i=1}^{m} \alpha^i \binom{n}{\frac{n}{2}+j} \leq \frac{\alpha}{1-\alpha}\binom{n}{\frac{n}{2}+j} \leq 2\alpha\binom{n}{\frac{n}{2}+j}$$

where the last inequality holds since $\alpha < \frac{1}{2}$. In particular we have

$$\sum_{k=\lceil n(\frac{1}{2}+\frac{1}{2m})\rceil}^{n} \binom{n}{k} = \sum_{i=1}^{m} \sum_{j=0}^{\lceil \frac{n}{2m}\rceil -1} \binom{n}{\frac{n}{2}+j+i\lceil \frac{n}{2m}\rceil}$$
$$\leq 2\alpha \sum_{j=0}^{\lceil \frac{n}{2m}\rceil -1} \binom{n}{\frac{n}{2}+j} < 2\alpha \sum_{j=0}^{n} \binom{n}{j}.$$

Resubstituting $\alpha = (1+\frac{1}{2m})^{-n}$ gives us

$$\frac{\sum_{k=\lceil n(\frac{1}{2}+\frac{1}{2m})\rceil}^{n} \binom{n}{k}}{2^n} \leq 2\left(1+\frac{1}{2m}\right)^{-n}.$$

□

Proof of Theorem 5.8**.** For *i*) observe that $\mathbf{p}_1 := (1,1,\dots 1)$ and $\mathbf{p}_2 := -\mathbf{p}_1$ have the property that for any voter $\mathbf{v}$ at least one of the two statements $\mathbf{p}_1 \cdot \mathbf{v} \geq 0$ and $\mathbf{p}_2 \cdot \mathbf{v} \geq 0$ holds. Thus each voter approves of at least one of these two parties.

ii). Let $\mathbb{V} := \{-1;1\}^n$ be the set of voters who have extreme positions on every single topic. We will show that the number of parties needed to ensure that every member of $\mathbb{V}$ votes is exponential in n. Fix some natural number $\frac{1}{m} \leq k$. Since the number of parties some voter $\mathbf{v}$ approves of is decreasing in k it suffices to show the theorem with $k = \frac{1}{m}$. Observe that for any party $\mathbf{p}$ and any voter $\mathbf{v} \in \mathbb{V}$ holds:

$$\mathbf{v} \cdot \mathbf{p} \geq \frac{1}{m}|\mathbf{v}|_1 \Leftrightarrow |\{i|v_i = p_i\}| \geq \frac{n}{2} + \frac{n}{2m}.$$

Since for any party $\mathbf{p}$ and any $l \in \mathbb{N}$

$$|\{\mathbf{v} \in \mathbb{V} : |\{i : v_i = p_i\}| = l\}| = \binom{n}{l}$$

this implies that each party $\mathbf{p}$ can be be approved by at most $\sum_{k=\lceil n(\frac{1}{2}+\frac{1}{2m})\rceil}^{n} \binom{n}{k}$ many members of $\mathbb{V}$. Since $|\mathbb{V}| = 2^n$ this implies that the number of parties needed to make sure that no member of $\mathbb{V}$ abstains is at least

$$\frac{2^n}{\sum_{k=\lceil n(\frac{1}{2}+\frac{1}{2m})\rceil}^{n} \binom{n}{k}}.$$

By lemma 5.17 this quotient is at least as large as $\frac{1}{2}\left(1+\frac{1}{2m}\right)^n$,in particular it is exponential in n. Since 2^n parties are enough to ensure that everybody votes, the number of parties needed cannot be worse than exponential.

The proof of *iii*) consists of two parts. First, we show that *at least* $n+1$ parties are needed in order to ensure that every voter finds a party she approves of. Assume to the contrary that $\mathbf{p}_1 \dots \mathbf{p}_n$ are enough to attract every possible voter. Recall that, by the voting rule used for *iii*), a voter $\mathbf{v}$ approves of a party $\mathbf{p}$ iff $\mathbf{v} \cdot \mathbf{p} > 0$. For $i < n$ define X_i to be the $n-1$ dimensional hypersurface defined by

$$X_i = \{\mathbf{x} \in [-1;1]^n | \mathbf{x} \cdot \mathbf{p}_i = 0\}.$$

Thus $X := X_1 \cap \ldots \cap X_{n-1}$ is a vector space of dimension at least 1 and therefore $Y = X \cap \{\mathbf{x} \in [-1;1]^n | \mathbf{x} \cdot \mathbf{p}_n \leq 0\} \neq \{\mathbf{0}\}$. Pick some non-zero $\mathbf{v} \in Y$. Then $\mathbf{v} \cdot \mathbf{p_i} = 0$ for $i < n$ and $\mathbf{v} \cdot \mathbf{p}_n \leq 0$, thus the voter $\mathbf{v}$ would abstain in an election with candidates $\mathbf{p}_1 \dots \mathbf{p}_n$, contradicting our assumption.
Next, we show that $n+1$ parties are sufficient to attract all voters if there are at least $n \geq 3$ topics. We start by showing the following general claim:
Claim: Let $\mathbf{v}_1 \dots \mathbf{v}_{n+1} \in \mathbb{R}^n$ be a set of vectors such that any n of these vectors are linearly independent. Then, after potentially replacing some $\mathbf{v}_i$ with $-\mathbf{v}_i$ we have that

$$X = \bigcap_{i \leq n+1} \{\mathbf{y} \in \mathbb{R}^n | \mathbf{y} \cdot \mathbf{x}_i \leq 0\} = \mathbf{0}.$$

To prove this claim let $\mathbf{v}_{n+1} = \sum_{i \leq n} \lambda_i \mathbf{v}_i$ be the unique representation of $\mathbf{v}_{n+1}$ in the basis $\mathbf{v}_1 \dots \mathbf{v}_n$. By our assumption all λ_i are non-zero and by replacing some $\mathbf{v}_i$ by $-\mathbf{v}_i$ we can assume that all λ_i are negative. Next we consider the space $X_i := \{\mathbf{y} \in \mathbb{R}^n | \mathbf{y} \cdot \mathbf{v}_i = 0\}$ for each $i \leq n$. Since $\mathbf{v}_1 \dots \mathbf{v}_n$ form a basis of $\mathbb{R}^n$, it follows for every $j \leq n$ that the space $\bigcap_{i \leq n}^{i \neq j} X_i$ has dimension 1. Using the assumption about the $\mathbf{v}_i$ again, we can pick for every $j \leq n$ some $\mathbf{x}_j \in \bigcap_{i \leq n}^{i \neq j} X_i$ with $\mathbf{x}_j \cdot \mathbf{v}_j > 0$. By our assumption about the $\mathbf{v}_i$, the $\mathbf{x}_1 \dots \mathbf{x}_n$ form again a basis of $\mathbb{R}^n$. Now, let $\mathbf{z} \in \mathbb{R}^n$ be any non-zero vector and let $\mathbf{z} = \sum \beta_i \mathbf{x}_i$ be its representation in the new basis. Then, if any β_j is larger than 0, we have

$$\mathbf{z} \cdot \mathbf{v}_j = \left(\sum \beta_i \mathbf{x}_i\right) \cdot \mathbf{v}_j = \sum \beta_i \mathbf{v}_i \cdot \mathbf{x}_j = \beta_j \ \mathbf{v}_j \cdot \mathbf{x}_j > 0.$$

If all β_j are smaller or equal than 0, we have

$$\mathbf{z} \cdot \mathbf{v}_{n+1} = \left(\sum \beta_i \mathbf{x}_i\right) \cdot \left(\sum \lambda_i \mathbf{v}_i\right) = \sum \beta_i \lambda_i \ \mathbf{v}_i \cdot \mathbf{x}_i > 0$$

where the last equation holds because all λ_i are smaller than 0. In any case we have $\mathbf{z} \notin X$ which finishes the proof of our claim. To finish our proof of *iii*), note that the set V consisting of the vector $(1 \dots 1)$ and the n vectors having exactly one -1 and all other entries 1 satisfy the conditions of the above claim.

Since each $\mathbf{x} \in V$ as well as every $-\mathbf{x}$ for $\mathbf{x}$ in V is an admissible party position, this completes our proof. □

Proof of Theorem 5.9. Fix a voter $\mathbf{v}$. Observe that for $k = 0$ and any party $\mathbf{p}$ at least one of the following two holds: $\mathbf{v} \cdot \mathbf{p} \leq 0 \cdot |\mathbf{v}|_1$ or $\mathbf{v} \cdot \mathbf{p} \geq 0 \cdot |\mathbf{v}|_1$. Let $\mathcal{P} = \{-1; 1\}^n$ be the set of all possible parties

$$\frac{|\{\mathbf{p} \in \mathcal{P} | \mathbf{p} \cdot \mathbf{v} \geq 0\}|}{|\mathcal{K}|} \geq \frac{1}{2}.$$

Since picking a random party is the same as randomly drawing a party from $\mathcal{P}$, the chance that a random party $\mathbf{p}$ satisfies $\mathbf{p} \cdot \mathbf{x} \geq 0$ is at least one half. Thus the chance that $\mathbf{v}$ approves of none of n random parties is at most $\frac{1}{2}^n$, thus $P(n, n, 0) \to 1$. Obviously, this implies $P(n, k) \to 1$ for any $k \leq 0$.
Since $P(n, k)$ is monotonous in k, it suffices to show that $P(n, n, \frac{1}{m}) \to 0$ for any natural number m. Let $\mathbf{v} = (1, 1, \ldots)$ be a voter who fully approves of all topics and let $m \in \mathbb{N}$. Observe that for any party $\mathbf{p}$ holds:

$$\mathbf{v} \cdot \mathbf{p} \geq \frac{1}{m}|\mathbf{v}|_1 \Leftrightarrow |\{i | p_i = 1\}| \geq \frac{n}{2} + \frac{n}{2m}.$$

Thus for the uniform distribution $\mathbb{P}$ over $\mathcal{P}$ we have

$$\mathbb{P}(\mathbf{v} \cdot \mathbf{p} \geq \frac{1}{m}|\mathbf{v}|_1) = \frac{\sum_{k=\lceil n(\frac{1}{2}+\frac{1}{2m})\rceil}^{n} \binom{n}{k}}{2^n} .$$

As above lemma 5.17 yields that

$$\mathbb{P}(\mathbf{v} \cdot \mathbf{p} \geq \frac{1}{m}|\mathbf{v}|_1) \leq 2 \left(\left(1 + \frac{1}{2m} \right)^{-1} \right)^n .$$

Thus $P(n, n, \frac{1}{m}) \leq 1 - (1 - 2\left(1 + \frac{1}{2m}\right)^n)^n$. It is a general fact that $(1 - kx^n)^n \to 1$ for any $x \in (0, 1)$ and $k \in \mathbb{R}$, thus $P(n, n, \frac{1}{m}) \to 0$ as claimed.

□

Proof of Theorem 5.10. Fix a voter $\mathbf{v}$ and let i such that $t_i = 0$. The third part of theorem 5.8 applied to voter $-\mathbf{v}$ shows that $n + 1$ parties are enough to guarantee that some party gets graded at most $i - 1$. Equally, the same theorem applied to $\mathbf{v}$ herself shows that the same $n + 1$ parties also guarantee that some candidate gets grade i or higher. Finally assume that there is no i with $t_i = 0$ and let j be maximal such that $t_{j-1} < 0$. Then the second part of Theorem 5.8 applied to $\mathbf{v}$ and $-\mathbf{v}$ shows that exponentially many parties are needed in order to ensure that some party gets a grade unequal to $j - 1$. □

Proof of Theorem 5.11. First assume that there is some i such that $t_i = 0$. Then, by theorem 5.9, the probability that at least one out of n random parties gets grade at least i goes to 1. Applying 5.9 to $-\mathbf{v}$ we see that also the probability that a party gets grade at most $i-1$ goes to 1. In particular, the probability for two parties receiving different grade assignments goes to 1, this proves the first part. For the second part assume that there is no such i. Let i_0 be such that $t_i < 0$ for all $i \leq i_0$ and $t_i > 0$ for all $i > i_0$. Then applying 5.9 with $k = t_{i_0}$ (if defined) yields that the probability that no party gets grade larger than i_0 goes towards 0. Applying 5.9 to $-\mathbf{v}$ yields that also the probability for parties getting a grade below i_0 goes to zero, thus the probability of all parties getting the same grade i_0 goes towards 1. □

Proof of Lemma 5.12. Observe that in 5.1 all summands indexed by some i with $v_i = 0$ vanish on both sides of the equation. Thus

$$\frac{\sum_{i\leq n} p_i v_i}{\sum_{i\leq n} |v_i|} = \frac{\sum_{i\in K} p_i v_i}{\sum_{i\in K} |v_i|}.$$

□

Proof of lemma 5.13. We prove *i)* and *ii)* simultaneously. Observe that the function $u_{\mathbf{p}}$ from $[-1;1]^n$ to $\mathbb{R}$, sending $\mathbf{x}$ to $\mathbf{x}\cdot\mathbf{p}$, is a linear function. By a well known theorem (see for instance [98, Theorem nn]), linear functions on polytopes assume their extrema on some vertex of that polytope. Since $\mathcal{C}(\mathbf{v},\alpha)$ is, by lemma 5.5, exactly the set of all $\mathbf{x}$ with $\mathbf{x}\cdot\mathbf{v} \geq |\mathbf{v}|_1$, we have

$$\begin{aligned} \Delta \cap \mathcal{C}(\mathbf{v},\alpha) \neq \emptyset \quad &\Leftrightarrow \quad \max\{u_{\mathbf{v}}(\mathbf{x}) \mid \mathbf{x} \models I\} = s^{opt}_{\mathbf{v}}(I) \geq k\cdot|\mathbf{v}|_1 \\ \Delta \subseteq \mathcal{C}(\mathbf{v},\alpha) \quad &\Leftrightarrow \quad \min\{u_{\mathbf{v}}(\mathbf{x}) \mid \mathbf{x} \models I\} = s^{pes}_{\mathbf{v}}(I) \geq k\cdot|\mathbf{v}|_1. \end{aligned}$$

For *iii)* note that an expected value voter approves of $\mathbf{p}$ iff $s^{ev}_{\mathbf{v}}(I) \geq k\cdot|\mathbf{v}|_1$. Further recall that the gravitational center of Δ is $\frac{1}{n}\sum_{i\leq n}\mathbf{x}_i$. We then have

$$\begin{aligned} s^{ev}_{\mathbf{v}}(I) \geq k\ |\mathbf{v}|_1 &\Leftrightarrow \frac{\sum_{\mathbf{v}_i} x_i\cdot\mathbf{v}}{|\{x_1\ldots x_n\}|} \geq k\ |\mathbf{v}|_1 \\ &\Leftrightarrow \mathbf{v}\cdot\frac{1}{|\{x_1\ldots x_n\}|}\sum_{x_i} x_i\cdot\mathbf{v} \geq k\ |\mathbf{v}|_1 \\ &\Leftrightarrow \mathbf{c}\in\mathcal{C}(\mathbf{v},\alpha). \end{aligned}$$

□

Chapter 6

The Dynamics of Generalized Trust

6.1 Introduction

Social capital is a leading concept in the research on overcoming collective action problems. Generalized trust is the touchstone of this concept as it captures the mechanism on the micro-level that drives people to cooperate with each other. A recent series of publications has identified a variety of social and individual benefits that arise from a high level of trust. These range from the performance of political institutions [135, 136] to economic capabilities of states [94, 95] to individual benefits such as health and a better quality of life [85, 88]. Motivated by these positive effects, we find a growing interest in the determinants of trust during the past years. Recent research has identified a variety of factors relevant for the emergence of trust including institutional factors, but also various cultural, societal and individual variables [84, 97]. We would like to add the idea that trust is not a stable phenomenon, but the result of an ongoing complex dynamic process. For understanding this process, it is not only important *which* factors contribute towards the emergence or destruction of trust, but also *how* they do so. In particular, the process underlying the emergence of trust is self-reinforcing, that is trust creates trust.

A multitude of empirical studies have shown the impact of factors such as the level of mobility [63, 134], the cultural background [62, 63, 70] and network structures [33, 132] on the emergence of trust. Instead of conducting yet another case study and adding new empirical evidence to these findings, we have taken these empirical results as a starting point for developing a multi-agent NetLogo

This chapter is based on joint work with J. Marx. See [90] for a related article.

simulation. In this chapter, we present the results obtained from our simulation, based on some of the above mentioned factors and mechanisms. On a more conceptual level, we see two major advantages in applying computer simulation to the emergence and dynamics of trust:

- First, a computer simulation helps to handle methodological problems inherent in our research. What might seem tautological - trust creating trust, that is: one factor reenforcing itself - and what might cause methodological problems in an empirical study can easily be cracked with the help of computer simulations. In order to deal with the tautology, we have developed our multi-agent NetLogo simulation in such a way as to endogenize trust. Computer simulations, unlike empirical studies, allow to determine at what point in time an individual has acquired trust proper, and when her behavior merely reflects second or third level trust, copying successful strategies of others. Building on a rational choice framework, we try to capture the causal mechanisms underlying the dynamics of trust.

- Second, we are interested in the quality of our theoretical knowledge of trust. We seek to understand the mechanisms that lead to a lower or higher level of trust in societies. Therefore we build our simulations on well established theories of the determinants of trust. If our simulation reacts to these variables in the way rational choice theory leads us to expect, we could take this as a validation of the underlying theories of trust. Vice versa, we could interpret unexpected results as a signal that the underlying theories on the emergence of trust are incomplete or misleading. In this sense simulations can help to evaluate the coherence and completeness of complex social theories.

In our simulation, we focus on the mechanisms responsible for the development of generalized trust. We are particularly interested in the interrelationship between agents and their immediate environment, how that environment reacts to the agents' willingness to place trust and vice versa. In particular, we will argue that the local neighborhood surrounding the individual agents is an underestimated factor in the creation and dynamics of trust within the population as a whole.

This chapter is structured as follows. We first clarify our concept of trust, based on a rational choice perspective. We then develop our basic dynamic model of trust used for our subsequent simulations. The model is built upon a population of agents playing trust games under the premise of uncertainty.

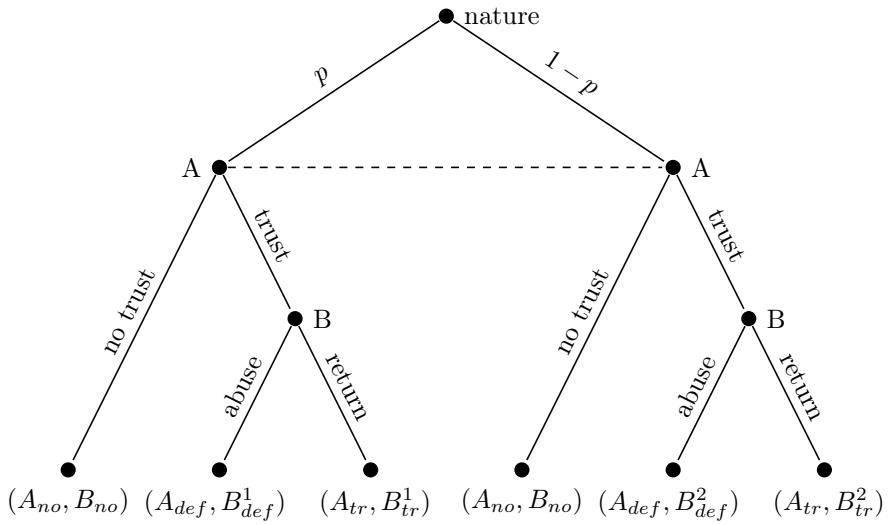


Table 6.1: Trust Game: The utilities satisfy $A_{tr} > A_{no} > A_{def}$ for the trustor A and $B^1_{tr} < B^1_{def}$ but $B^2_{tr} > B^2_{def}$ for the trustee. Trustees in the left branch will exploit trust while trustees in the right branch have a dominant strategy of cooperation.

In doing so, the agents have to decide whether their interaction partners are trustworthy or not. More precisely, they have to learn about the expected payoff of placing trust. In a second step, we elaborate our model, adding several context conditions that are believed to affect the general trust level: Mobility, network structures, and spatial inhomogeneities. We show that some claimed effects of these variables, well supported by empirical data, cannot be completely reproduced in our simulation. On the contrary, some of our findings are in straight opposition to predictions from the literature. For instance, we find a negative impact of low mobility on the individual level of trust where the literature would predict a positive impact. At the same time, we identify a strong effect of the direct neighborhood the agents are embedded in, a factor yet underrepresented in current theories. By randomizing the neighborhood of the agent and keeping all other variables constant, we can demonstrate the high impact of this factor on the variance of trust. We finally conclude by discussing the relevance of our results.

6.2 Trust as an Expectation

Trust is a multifaceted concept. In this paper, we focus on a narrow definition of trust as an expectation that can be held by rational agents.[1] More concretely, trust is considered to be an agent's expectation that her counterpart will act cooperatively in certain strategic situations. A high level of trust is essential for solving the cooperation problems that arise in situations of strategic interdependence. Such situations can be characterized by the following conditions: A resource is shifted from actor A, the trustor, to actor B, the trustee. The trustor's reason for this shift of resources is the expectation to gain from that interaction. However, in shifting resources, actor A makes herself vulnerable. Her utility diminishes if the trustee does not repay her initial investment. Trust games involve a crucial temporal asymmetry. The trustor pays prior to learning about the trustee's response and the latter, in turn, does not need to decide on her behavior until the trustor has moved. If the trustee proves trustworthy, both parties receive a positive payoff. If the trustee turns out to be untrustworthy, only the trustee benefits while the trustor ends up with the worst possible payoff. Thus, engaging in trust games is a conscious decision under risk. The trustor invests voluntarily and without guaranteed success. Of course, a rational trustor will only engage in a trust game if she expects the opponent to be trustworthy. But, given the temporal structure of the game, there is no guarantee that the trustee acts as expected. The trustor might assess her opponent's trustworthiness incorrectly, trusting an defector or refusing to play with a trustee that would have cooperated. To accommodate this risk of incorrect assessments, we will define trust as a graded variable (cf. [41, p.91-116]), describing *how likely* the trustor judges her counterpart to be trustworthy. Within the framework of bounded rational choice theory, this situation can be represented as a trust game in extensive form with incomplete information [34]. Here, the trustor's uncertainty about the trustee's trustworthiness is represented as uncertainty about her payoff structure, see Table 6.1.[2] In this model, the uncertainty about the motivations of the trustee is expressed as a draw by nature: With a certain probability p, A's counterpart is not trustworthy, i.e., A will interact with a partner with a dominant strategy of defecting. $1-p$ is the corresponding likelihood of playing with a trustworthy player, having cooperation as her dominant

[1] See for example [151, 152] for other conceptions.

[2] The payoffs in table 6.1 reflect the *all-out* utilities governing the agents' choice. The *material* structure of a trust game (cf. [18]) is usually assumed to be as in the left side of table 6.1 having *(no-trust,abuse)* as unique Nash equilibrium.

strategy.[3]

The central task for the trustor is thus to estimate the likelihood of being paired with a trustworthy trustee. She will agree to place trust in her counterpart if and only if she experts her to be cooperative. In this chapter, we are interested in the determinants of trust: Under which conditions will our agents develop the expectation that other agents are trustworthy? When will they expect others to defect?

Current literature on social capital [136, 153] distinguishes a thick and a thin notion of trust. The thick notion of trust, on the one hand, refers to personalized attitudes and expectations towards well known, individual others. This thick notion of trust is grounded in a well established social relation between the actors, based on acquaintance, joint past experience, institutional frames or expectations of future interaction. For instance, [122] presents a related simulation on the emergence of thick trust. The thin notion of trust, on the other hand, refers to the general attitude towards strangers, anonymous and hitherto unknown members of society that we might not expect to ever see again. Faced with situations of thin trust, agents base their behavior on prevailing social norms, past experience in similar situations, demeanor, appearance or, more general, membership in certain social groups [21]. It is this second notion of thin or *general* trust that reflects the aggregated social capital of a society we are interested in. In our simulation, we focus on several factors relevant for the evolution and emergence of *thin* trust.

Previous research has revealed several determinants of trust. First, the stability of the social context should allow agents to learn the utility of trust by iterated interactions [41, p.91-116]. Second, the level of trust initially existing within a society should, of course, have an impact on the long term stability and emergence of trust. This initial trust level is classically understood as a form of cultural heritage, passed on to next generation by means of socialization. Differences in such cultural heritage could, of course, explain why some populations display a high level of trust and others do not [136]. Finally, networks and, more generally, social cohesion are perceived as a source of trust [110]. Accordingly, we expect mobility to have a negative impact on the trust-level whereas isolating spatial structures and networks should have a positive effect. We would further expect a close relation between the percentage of defecting

[3]Cooperative preferences of the trustee can be motivated by several factors as for example social norms, sanctions, reputation or anticipation of future interactions. For our current purposes, the reasons underlying the trustee's behavior are irrelevant. The only thing we need for this simulation is that there are two types of trustees with different dominant strategies.

agents and the overall trust level at the end of our simulations. Naturally, that average trust level at the end of a simulation should be closely connected to the real percentage of defecting agents in the population.

6.3 The Model

We take these bounded rationality considerations as a starting point for our simulation. Our model is based on a population of agents dynamically moving within a larger society, constantly faced with the decision on whether or not to trust a stranger. These agents thus need to determine whether it pays off to trust some anonymous member of society that one has never seen before. They do so by gradually learning the expected payoff of offering trust to others, by continuously engaging in trust games. Crucially, all agents will assume both roles sometimes acting as trustors and sometimes as trustees. We need to determine the agents' behavior in either role and their corresponding choice mechanisms. We start with the easier of these cases, the agents' behavior as trustees. We take such behavior to be guided by some deeper mechanism such as a social norm [19, 20] some concern for reputation, or the fear of legal prosecution. Crucially, all these factors evolve on a much larger time scale than the beliefs relevant for the rational decision whether to trust, thus we can safely assume them to be constant over time. On the other hand, trustors are guided by the rational choice approach described above. A trustor engages in trust games as long as she expects them to be advantageous for her, on average. Thus, she seeks to learn about the expected payoff of offering trust, or to put it differently: She sets out to inquire about the level of trustworthiness present in the given society. We want to emphasize here that there is no interaction between the two roles, an agent's behavior as trustor is completely detached from her actions as a trustee.

Formally, we represent every member A of a society as a twofold agent, encoding both her behavior as a trustor and a trustee. Since we take the trustee's behavior to be invariable over time, we can represent it by a simple binary parameter *trustee*, with the values 1 (A always cooperates as a trustee) and 0 (A always defects). On the other hand, A's behavior as trustor is completely determined by her expectation of whether trustees cooperate or not. For this purpose, A is equipped with a variable *trust memory* tracking her expectation of a generic trustee being trustworthy or not. As the name suggests, that expectation is primarily based on the agent's past experience. The value of trust memory is updated with every new piece of information A gathers, that is, with

every trust game she is involved in, be it as a trustor or a trustee. We model this update with a weighted average between the old expectation and the newly acquired information. Our mechanism is in line with the paradigm of Bayesian Sensor Integration, see for instance [87, p.10] for details and a discussion of alternative updating rules. To be more precise, we let the variable *trust memory* range from 0 to 10, where 0 is the expectation that the trustees will defect for sure while 10 is the expectation that every trustee will cooperate, no matter what. For updating the trust memory, we need to fix the weight $\beta \in [0;1]$, called *increment*, the agent attributes to newly acquired information. Updating on a newly incoming piece of information E is then defined as a weighted average between the old trust memory and the newly incoming information E, where the latter receives weight *increment*:

$$\text{trust memory}_{new} = (1-\beta) \cdot \text{trust memory}_{old} + \beta \cdot E. \qquad (*)$$

We allow for two different ways in which actor A can make new observations E. First, when acting as a trustor, she has direct access to a new piece of evidence about the behavior of trustees: If the current trustee cooperates, the trustor receives a positive feedback ($E = 10$). A defecting trustee, on the other hand, triggers a negative feedback ($E = 0$). However, there is also a second order way of obtaining information about the behavior of trustees: When assuming the role of a trustee, an agent can observe whether the corresponding trustor places trust in her or not. Taking that trustor to be a rational agent, playing her best strategy, this conveys some indirect or *social* clue about the expected behavior of other trustees. In the current model, we treat this indirect way of learning on par with the direct information collected as a trustor. Thus the possible observations are $E = 10$ if the trustor is willing to place trust and $E = 0$ if she refuses to do so. Later, we will inquire into the relationship between direct and indirect information by varying the relative weights attributed to the different types of information.

Next, we need to describe the trustor's choice rule, based on her collected information. Following our rational choice approach, agents place trust in others if the expected return of doing so outweighs the expected return of not doing so. The Harsanyi transformation of the trust game (see Table 6.1) reduces this expected utility calculation to the simple question of how likely it is that the opponent is trustworthy. In our model, the expectation on trustworthiness is guided by the agent's past observations, recorded in the *trust memory* register. Our agents are equipped with a simple decision rule, determining their trustor behavior as a function of their trust memory. The decision rule we use in this

paper is the threshold rule:

Play trust if trust memory ≥ 5, *else do not play.*

The exact threshold in the above rule reflects a variety of different considerations. Castelfranchi and Falcone [36, 37, 56] describe trust as a multifaceted attitude, resting on a variety of parameters such as the actual monetary stakes, the trustor's aversion towards misplacing trust, her assessment of the trustee's capacity to perform relevant tasks and, of course, her actual trustworthiness, her willingness to perform the task attributed to her. All of these factors impact, in some way, on the trustor's utility assignments. For the sake of simplicity, we assume the first three of these to be constant over time, making the trustor's trust memory, her expectation about the trustee's trustworthiness, the only *variable* parameter in her decision whether to place trust. Thus, an agent will agree to place trust if her trust memory exceeds a certain threshold, else she will refuse to accept the trustor's role. The exact value of that threshold reflects, of course, the exact values of the various other, fixed parameters: the monetary stakes, the trustor's aversion towards having her trust exploited and so forth.

The main focus of our model is not to form expectations about the behavior of some particular individuals, but to learn about the expected payoff of placing trust in some generic, hitherto unknown individuals. Therefore, our basic model does not contain any individualized learning mechanisms. It is only in a second step, a mild extension of our model, that we investigate the effect of a minor learning mechanism about the trustworthiness of individual others. In this extended model, every agent can keep track of individuals that have proven to be exceptionally trustworthy. In situations where the trustor can choose between various trustees, she will then preferably interact with these trustworthy agents, if any of them are available.

In order to prevent small-world effects, that is agents repeatedly interacting with the same partners, we assume a relatively large grid of 51×51 fields, populated with 1500 randomly distributed agents. The various parameters of the agents and the initial distribution are described below. To increase the homogeneity of the model, agents crossing the right edge of the grid reappear at the left edge and vice versa – the same holds for the top and bottom edge. Each round of simulation consists of a first phase in which the agents interact with each other, followed by a second phase, in which they move around. In the interaction phase, the agents randomly pick some individual that is not yet engaged in any trust game in their immediate vicinity, their van-Neumann

neighborhood. If no such partner is available, the agent stays unpartnered and does not engage in a trust game in that round. Thus, every agent can only be part of at most one pair, acting there as either trustor or trustee. After all pairs played a trust game as described above, all agents, including the non-partnered, move n steps in a random direction, where the moving speed n is controlled by the input parameter *mobility*. Each spot can only be occupied by one agent at a time. If an agent steps on an already occupied field, she will repeat the moving routine until she finds a free spot. This procedure is iterated 1000 times before the final measures, the average trust memory and the final percentage of trusting agents are extracted.[4]

Before describing the relevant parameters of the model, we should point out a crucial asymmetry of the mechanism laid out above. While trusting agents, i.e., agents with a trust memory larger than 5, happily collect information round after round, the pessimistic agents are in a less promising state of mind. By refusing to accept the trustor's role in a trust game they deprive themselves of the possibility to collect further direct evidence about the expected utility of trust. The only chance these agents have to adjust their trust memories and to eventually leave their pessimistic states is to assume the role of a trustee and to therein obtain new information through the corresponding trustor. Pessimists thus have to encounter a *trusting* agent, accepting the role of a trustor and thereby conveying a new, positive piece of evidence, in order to become trusting again.

We are interested in the influence of four particular parameters on the emergence and maintenance of general trust. These are, first, the actual percentage of trust-abusers or defectors among all agents, instantiated by the variable **percentage of defectors**. These trust abusing agents are randomly chosen among all available agents at the beginning of each simulation run. We expect this variable to have a high influence on the emergence of trust, since the percentage of defectors reflects the actual value of the parameter the agents intend to learn about. However, given that our interactions have a local character, density fluctuations in the distribution of defectors might have a significant effect on the learning process.

The second variable we are interested in is a global parameter, the **starting trust degree**, encoding the agents' prior beliefs about the cooperativity of others. The starting trust degree is given by a number in the interval $[0; 10]$.

[4]We have identified certain stable configurations that our simulations do not leave again once they are reached. In case such a stable situation was reached we stopped the simulation early and took the final measures.

The initial trust memory of each agent is then picked from a normal distribution with standard variation of 2 around that starting trust degree. As agents interact in the model, these priors will be gradually updated and overwritten by actual experience of the agents. We therefore expect only a moderate influence of the starting trust degree on the final state. However, if the starting trust degree is too low, agents will refrain from accepting the trustor role in the first place, thereby never being able to learn about the *actual* state of the world. Furthermore, observing other agents also not accepting the trustor role will instantiate an informational cascade. Seeing these other agents refuse to engage in trust games is taken as additional evidence against trusting others, thus reinforcing one's own negative priors. We therefore expect the game to converge towards a state of total distrust, once the starting trust degree falls below a certain threshold.

The third variable of interest is the agents' **mobility**, a positive number encoding the distance agents move each round. While the partner for the trust game is always chosen from a local neighborhood of the agent, the speed with which agents move between games can vary. This variance is given by the factor mobility, describing the velocity of the diffusion process in society. Arguably there is a crucial difference between a mobility of zero and any dynamic case, characterized by a positive mobility. The first case mirrors immobile agents learning about the static, small vicinity surrounding them. The attitudes of such agents may evolve dynamically, but their surrounding neighborhood remains constant over time. In the present model, we focus on the second, dynamic case, that is, we require agents to move a positive distance each round. Our expectation is that little mobility will at least temporarily give rise to local differences in the level of trust, since agents are moving slowly enough to allow for localized trustful or distrustful regions. Higher levels mobility, on the other hand, should prevent these regional patterns from arising. Classic predictions from the literature [134] postulate a *negative* relationship between mobility and average trust memory and expect a close connection between low mobility and a high level of trust.

Finally, the fourth value of interest, is the weight β governing the agent's updating mechanisms. We encode this weight by the factor **increment**, in the interval $[0; 1]$. This increment represents the amount of epistemic weight agents attribute to newly acquired information, compared to their previously held believes. Arguably, the role of this weight factor is most intricate. An all too minimal level of β results in agents remaining in their initial state of belief,

hardly reacting to their environment at all. On the other hand, a too high level of β might indicate that agents distrust their own past information, therefore attributing a high weight to recent and incoming evidence. A high level of increment will thus make the agents' beliefs unstable by putting much weight on the stream of incoming information with all its noise and local fluctuation. To put it differently, a high level of increment makes the agents all too vulnerable to momentary frustration: Two or three consecutive negative experiences might already suffice to convert even the most optimistic agent with an initial trust memory close to 10, into a pessimist with a trust memory below 5, a state hard to escape as we have argued above. Thus, best results are to be expected for intermediate values of β. Empirical studies have shown the values attributed to newly incoming information to lie anywhere between 3% and 10%, (see [4, p.154] p.154, or [17]). For most of our results we will thus average over the different values of increment between three and ten percent to wage against phenomena resting on any particular value.

Finally, we will present several extensions of the model, accommodating various realistic assumptions. First, we introduce a very limited memory capacity: While most of our daily interactions slip our attention quickly, certain important experiences leave a more permanent impression, especially those that are reinforced by several encounters within a short time span. To model this, we introduce the following iterative learning scheme: If an agent A encounters the same partner twice within a short lapse of time, at the most 9 other positive interactions in between, and this partner turns out to be trustworthy, A learns about the nature of this particular partner and adds the latter to her personal list of trustworthy agents. This learning is done by building up a unilateral network tie to that partner. Now, whenever A is to pick a trustee, she preferably picks a partner from that list, if available. We expect the introduction of this ability to increase the chance of being paired up with a trustworthy agent, and thus to foster the emergence of general trust.

Second, we introduce certain geometric inhomogeneities: impenetrable walls or enclosed areas with bottleneck access. These spatial elements of the grid create closed local units that are almost uncoupled from the rest of the population, allowing them to develop their own, local traditions and expectations. Local inhomogeneities can, in principle, represent any type of mobility confinements to individual agents, be it class distinctions, economic differences or actual geographical obstacles.

Third and finally, we are interested in the impacts of external shocks on the

base model and its various extensions: Events such as rumors, misperceptions or particularly prominent showcase examples, propagated and reiterated through the media, can influence the general trust expectation above and beyond the agents' first hand experiences. Here, we are interested in the question of whether such small disruptions are evened out in the long run, or whether they might have a lasting effect on the emergence and destruction of trust. Further, we are interested in which factors and conditions make a society vulnerable or resilient to such short-term shocks.

6.4 Results

In setting up our simulation, we are primarily interested in two types of results, understanding the local dynamics of trust over a limited amount of rounds and understanding the limit behavior of the system and how that depends on the various input parameters. We will primarily use two measures to track the emergence of trust within a society. The first of these is the share of trusting agents at the end of each simulation. Following our decision, these are exactly those agents with a trust memory of at least 5, thus we are interested in the measure *percent trusting*, defined by

$$\textit{percent trusting} = \frac{\text{Number of agents with trust memory} \geq 5}{\text{Number of agents}}.$$

Our first result is that, in the long run, our model always converges towards the extreme values 0 or 1 of *share trusting*. Prima facie, this is not implausible. All agents are interested in the *same* question, whether the actual share of trustworthy agents is high enough to justify trusting unknown others, thus it is not surprising that they all eventually converge towards the same result. However, we attribute the high degree of uniformity in this convergence partially to the social part of our learning mechanism, guided by second order information: The trustee learns about the trustor's behavior and infers from there to the trustor's informational state. Once the value of *percent trusting* is sufficiently close to either 0 or 1, *i.e.* almost all trustors share the same trust attitude, the second order information received is so uniformly negative (resp. positive) that it drives the state of society further towards that respective extreme. Thus, the two states of universal trust and universal distrust could be understood as stable behavioral equilibria in the iterated trust games performed within our society. Once a simulation has reached one of these stable equilibria, we will call it *trusting* or *distrusting* respectively. We are, however, not only interested

in *how* our simulation converges to these equilibria, but also *which* equilibrium the simulation approaches in the first place. That is, we are interested in how often the different simulations converge towards a state of universal trust and, more generally, in the exact dependency of this equilibrium selection on the various input parameters of the model. Thus, the second output measure we are interested in is **share trusting**, the share of simulations that converge towards the trusting equilibrium. Given a set of simulations S, this output measure is defined by

$$\textit{Share trusting} = \frac{\text{Number of } \textit{Trusting} \text{ Simulations in } S}{\text{Number of Simulations in } S}.$$

It is exactly this study of equilibrium selection and equilibrium convergence that computational models are most helpful in. While classic game theoretic solution concepts have little to say about the way an equilibrium emerges, our simulation does not only offer a reproduction of the game theoretical results, but it also gives a plausible reconstruction of the way these equilibria develop.

6.4.1 The Basic Model

We start our examination with two basic results that serve to validate our model. Naturally, we expect the starting trust degree, the cultural heritage determining the initial trust expectation of agents, to be positively correlated with the emergence of trust, while the share of trust abusing agents should have a negative influence. Our model does validate these two predictions. All other variables held constant, we find a positive correlation between the value of *starting-trust* and the likelihood of a simulation converging towards the trusting equilibrium described above. As can be seen in table 6.2, a higher level of initial trust degree makes it more likely that all agents will expect others to be trustworthy at the end of the simulation. In this basic form, our simulation displays the structural behavior predicted by rational choice theory. We take this as a first validation that our simulation is an adequate implementation of the game theoretic model.

Equally, we find a negative correlation between the percentage of defectors and the probability of a simulation converging towards the trusting equilibrium (see table 6.3). Again, this is in line with the predictions of game theory and we take it as a validation of our model. Both these effects are stable towards changes in the remaining parameters and easy to reproduce. We thus hold that our basic simulation is an adequate implementation of the game theoretic model on trust.

Table 6.2: Impact of *starting trust degree*

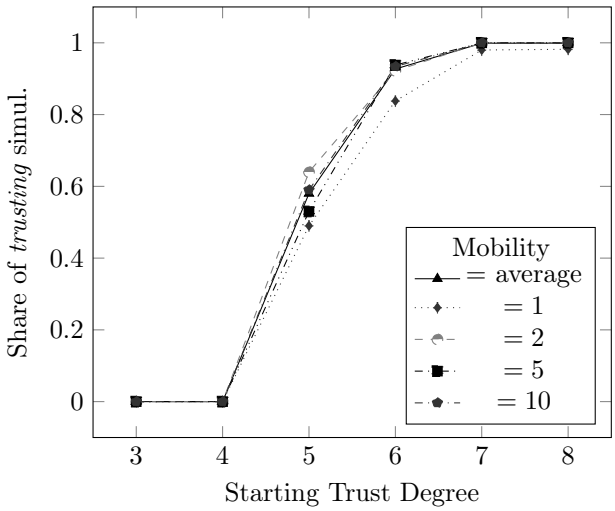
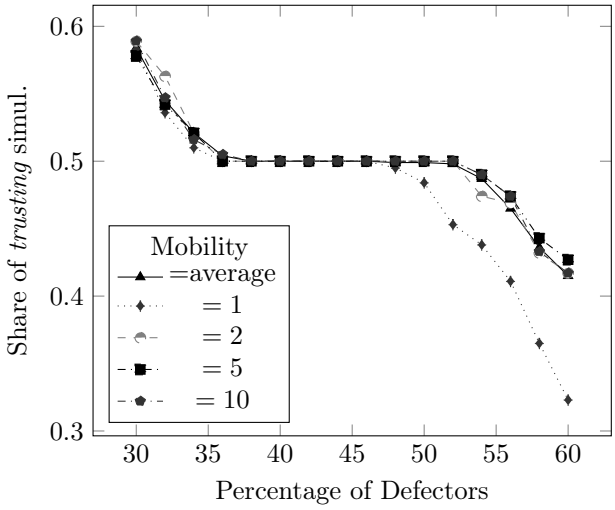


Table 6.3: Impact of *percentage of defectors*



For subsequent experiments, we identified a region where the results of the simulations are not yet determined by any of these two variables alone. In the following, we will study the influence of four different parameters on the emergence of trust. In particular, we will use our simulation to examine the validity of the corresponding predictions about these four parameters. The first step is to examine the effect of mobility on the average trust degree. Next, we vary the weight attributed to newly incoming information. In a third step we allow agents to construct networks as a very limited way to keep track of trustworthy partners. We then proceed as a fourth and last step by introducing spatial restrictions to our grid that limit the mobility of our agents. Finally, as an extension, we examine the vulnerability of our model to external shocks on the trust memory.

6.4.2 The Enhanced Model

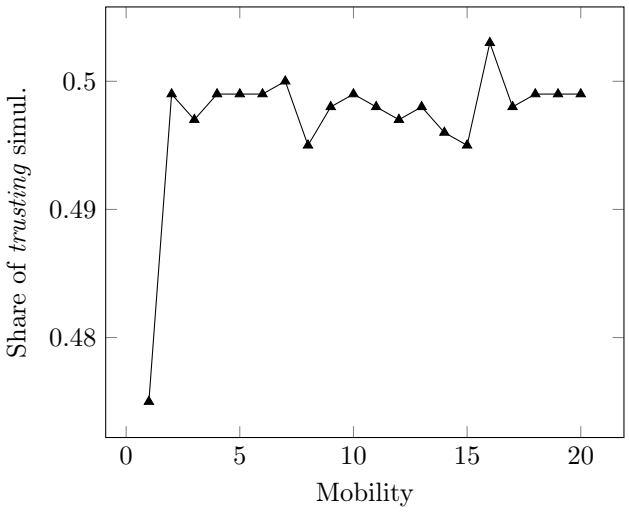
The following analysis is based on a total sample space of 61440 single simulations with different settings, two for every combination of parameters within the range studied. Half of the simulations included networks. The parameters were chosen such that the experiment is not determined by either the starting trust degree or the percentage of defectors alone, but by an interplay of these two together with the other parameters of the model. Unless noted otherwise, the results presented in subsequent sections are based on the set of basic simulations without networks or spatial restrictions, averaging over different values of initial trust level, percentage of defectors and increment.

Mobility

In recent theories on social capital, the factor mobility is believed to have a *negative* impact on the general level of social trust. The current state of research would argue that social norms and general trust tend to be stronger in smaller contexts (Putnam 1995). Mobility is sometimes even identified as the central characteristics of modern society responsible for the decline of social trust. All the more surprising are our findings: In the present model, the factor mobility correlates *positively* with the emergence of general trust. As can be seen in table 6.4, a mobility of 1 is detrimental to the emergence of trust, while higher levels of mobility do not have any traceable impact.[5]

[5]Within larger grids, similar but weaker effects can also be generated for a mobility level of 2. Thus, we take this phenomenon to be caused by the interplay of field size and mobility rather than the second factor alone.

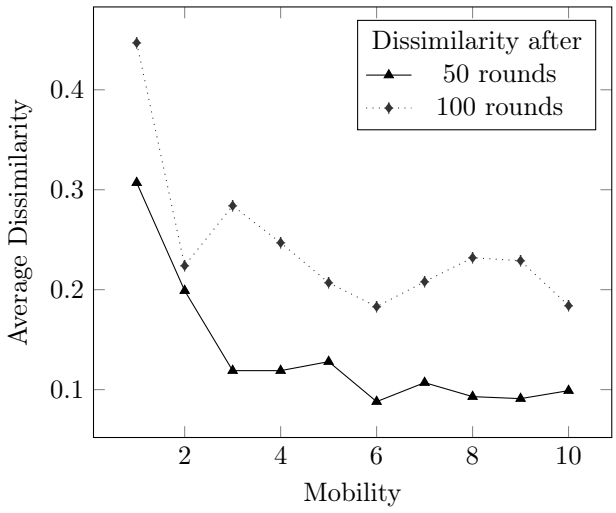
Table 6.4: Effects of Mobility



Our general data, presented in tables 6.2 and 6.3, allows for a more fine grained analysis of this phenomenon and under which conditions it appears. As it turns out, the influence of mobility is categorically weaker than the first two parameters, starting trust degree and the percentage of defectors. Said negative influence of a low mobility only appears when the simulation is not already determined by any of the other two parameters alone. The major impact of a low mobility is at moderate starting trust degrees of 5 or 6. For higher levels of initial trust, the model is already too heavily bent towards an all-trusting state to allow for a significant impact of mobility. A similar analysis holds true for the impact of percent-defectors, as can be seen in table 6.3. Here, a low and medium percentage of defectors already loads the dice too heavily towards an all-trusting state, whereas the highest impact of mobility appears at a sufficiently high level of defectors.

We claim that this difference between a mobility of 1 and higher levels of mobility can be traced back to a local clustering effect. At a mobility of one, local clusters of trust and distrust occur, as illustrated in figure 6.1, while higher levels of mobility contribute towards a more isotropic distribution of trust and distrust. To continue our argument, we proceed as follows: We first show that the emergence of trust is inversely correlated to the existence of local clusters before arguing *how* a local clustering can contribute to the emergence of distrust. To show that clusters primarily occur at a mobility of 1 and, much less so, at a mobility of 2, we calculate the index of dissimilarity between trusting

Table 6.5: Effects of Mobility on Dissimilarity



and distrusting agents, measuring how unevenly these two types are distributed across the entire field. For the current purpose, we measure dissimilarity with the modified Bray-Curtis Index of Similarity,[6] based upon a subdivision of the field in 3×3 square districts of equal size. Table 6.5 shows the index of dissimilarity taken after 50 and 100 simulation rounds for different values of mobility. Our findings show a rise of the clustering index at the beginning of our simulation. This observation is remarkable since we did not implement any mechanism directly supporting the clustering of the agents.[7] This clustering is instead an endogenous effect of our model that is not reducible to any special variable.

Finally, we argue *how* a local clustering could favor the emergence of distrust. More precisely, we show that clusters of distrust spread out, gradually infecting the trusting regions around them as illustrated in figure 6.1. The reason for

[6]Let M be a map divided into a set I of different sectors and let p and q be two populations on that map. For each sector $i \in I$ let p_i and q_i be the number of p and q agents respectively living in that sector. Then the modified Bray Curtis index of similarity between p and q is given by

$$\frac{1}{2} \sum_{i \in I} \left| \frac{p_i}{\sum_{j \in I} p_j} - \frac{q_i}{\sum_{j \in I} q_j} \right|.$$

Thus, the modified Bray Curtis index measures the classical Bray Curtis dissimilarity [31] between the local density functions $\frac{p_i}{\sum_{j \in I} p_j}$ and $\frac{q_i}{\sum_{j \in I} q_j}$.

[7]Such mechanisms are quite common in agent-based-simulations, cf for example the segregation model of Thomas Schelling [139]. In this model agents remain at their place of living as long as a certain criterion (having a certain amount of neighbors that are similar to themselves) is met and move elsewhere if not. Even under moderate input parameters, the simulation produces a highly segregated output model.

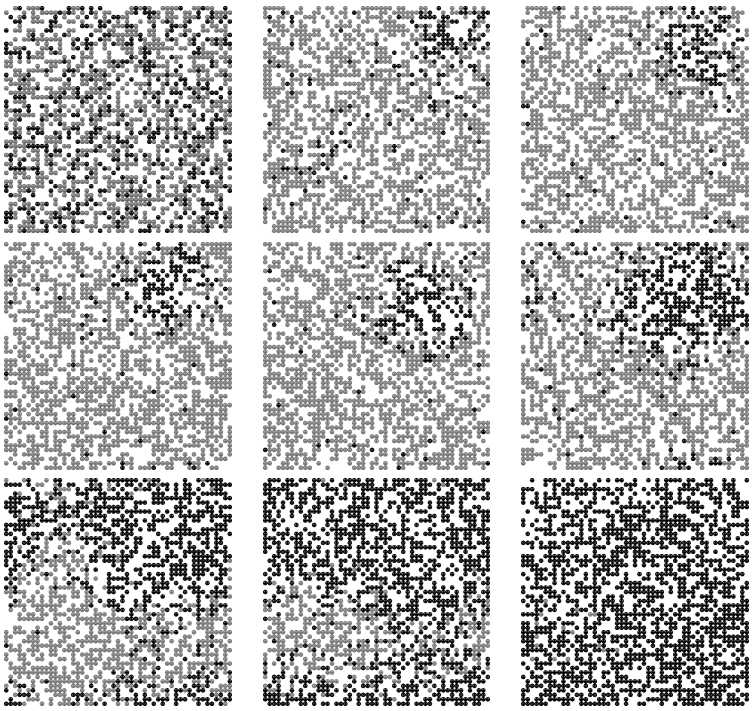


Figure 6.1: Expanding Cluster of Distrust (black) under *mobility* = 1

this, briefly, is that distrust is more stable than trust. We identify two related explanations for this stability of distrust. First, we argue that the *strength* of the respective beliefs will be much higher in a distrusting cluster than in the corresponding trusting cluster. Thus, once trusting and distrusting clusters clash, the distrusting agents will be more resilient in their beliefs, making it more likely for them to convert others than vice versa. To see this, we note that once a group of agents converged to a general state of distrust, agents refuse to engage in any further trust situations. Thus, the only *new* information available to such agents is the uniformly negative second order information trustees extract from the fact that no other trustor is willing to place any trust. Therefore, the general trust memory in such a cluster will gradually decline towards 0, the absolute minimum. Within a cluster of *trusting* agents, on the other hand, the individual actors continue to collect new first order information about the average level of trustworthiness. The second order information inside such trusting clusters will be as uniformly positive as it is negative in distrusting clusters. However, this second order information is mixed with the same agents' first order experience, which is sometimes positive and sometimes negative, depending on the part-

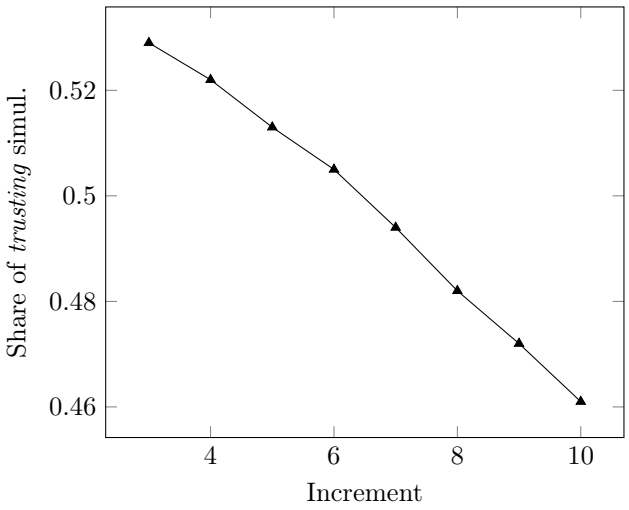
ner's trustee type. In particular, receiving both types of information, agents within the trusting cluster will never converge to the maximal trust level of 10. In particular we expect the *strength* of beliefs to be weaker inside the trusting cluster than in a distrusting cluster. As argued above, this asymmetry affects the interplay of trusting and distrusting agents occurring at the border areas between the respective clusters. Being less extreme in their beliefs makes trusting agents also less resilient towards changing their attitude towards trust. It is more likely for a trusting agent to become distrusting than vice versa, resulting in a gradual growth of the distrusting cluster.

The second reason we give for the stability of distrust is related to the agents' learning speed. We will show that trusting agents update their beliefs more often than distrusting agents. Thus, since every change in trusting behavior is triggered by some informational change, it is also more likely that a trusting agent changes her trusting behavior than a distrusting agent. To go a bit more into detail, the newly collected information of some agents will be usually mixed, containing pieces of positive and negative evidence. This phenomenon is especially prominent at the border areas between trusting and distrusting clusters, where both the first and second order information can be positive or negative depending on whether the corresponding partner is trustworthy (respectively trusting) or not. The volatile stream of incoming data has a certain chance of containing short segments of uniformly positive or negative information, the main reason for agents to switch their types. However, note that trusting agents collect double as much information[8] as distrusting agents in the same time interval. The former collect new information in the roles of trustors and trustees, while the latter only use their trustee role to update their informational state. Thus,[9] the chance of a trusting agent switching her state in any given time interval is higher than the chance of a distrusting agent doing so, simply because the trusting agent collects more information and has thus a higher chance of being subjected to a short stream of uniformly negative information necessary for a change in trusting behavior. Hence, there is a higher chance of some trusting agent becoming distrusting than vice versa, thus causing the distrusting population to grow gradually.

[8]Recall that all agents have an equal chance of being the first or second party in a trust game. It is only *inside* such a trust game that distrusting agents prefer to play *no trust* (see table 6.1), preventing them from learning about the trustee's behavior. Every agent can only be engaged in at most one trust game at a time, thus being distrusting does not increase the chance of being picked as a trustee.

[9]Of course, the chance of a trusting agent switching her state will depend upon the exact number of trustworthy agents present. If this number exceeds a certain threshold, trust might actually become more stable than distrust.

Table 6.6: The Effect of *increment* on the Emergence of Trust



Increment

Next, we turn our attention towards the factor increment, the relative importance agents attribute to their more recent information. Our agents estimate the expected payoff of trust through a continued adaptive learning process. In this process, each piece of incoming information is incorporated into the agents' trust expectation through a weighted average as depicted in formula $(*)$. The factor increment describes the weight attached to the newly incoming information, the higher the factor increment the more evidential weight an agent puts on her most recent encounters. This parameter is of a different nature than that of the other parameters examined, since it is not directly or indirectly determined by social conditions alone, but hard wired into the human learning mechanism. Empirical estimates of *increment* in the literature lie anywhere between 3 and 10%, see[17] for an overview.

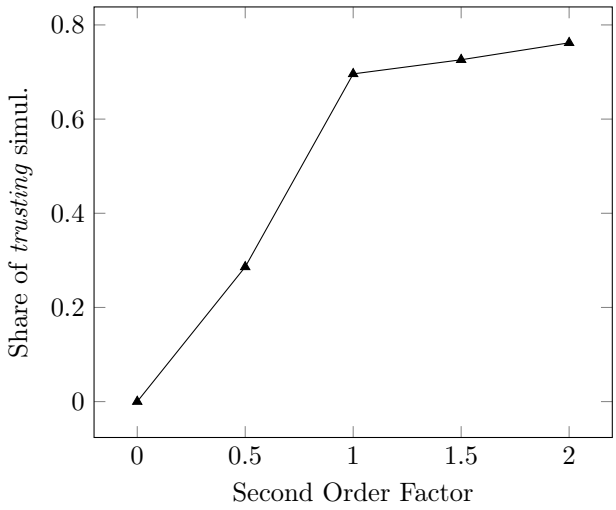
As it turns out, the parameter increment *does* have a relevant impact on the amount of simulations converging towards universal trust. Within the range of increment studied, there is a *negative* relation between increment and the probability of a simulation running towards the trusting equilibrium. Lower increments make agents less dependent on their more recent information and focus more on the bigger picture of their collected information. Arguably, this focus on the bigger picture fosters the creation and stability of trust, as long as there is, in fact, a sufficient amount of trustworthy agents.

We explain this again by the asymmetry between trusting and distrusting agents described above. Even in a society with fairly high degrees of trustworthiness, an unfortunate sequence of trust-abusing encounters can move an originally trusting agent into distrust, a state hard to escape from. The higher the value of increment, that is the more weight an agent puts on her most recent encounters, the less consecutive negative information is needed to thwart some agent's trusting behavior. To illustrate this with some numbers, a highly trusting agent with a trust memory of 9.4 and an increment of 5 requires 13 consecutive negative experiences before loosing her willingness to place trust. The same agent, but with an increment of 10 would already be frustrated by a short stream of 6 negative encounters. Notably, the *qualitative* outcomes of our simulations do not hinge on the precise values of increment. While the different values of that parameter do generate different quantitative outcomes, the qualitative relationship between the remaining parameters remains invariant. We take this as a further validation of our simulation, showing that our qualitative results are robust towards smaller changes in the parameter increment. For the rest of this chapter, we hedge against the impact of any particular values of increment by averaging over the integer values of increment in the interval between 3 and 10%.

Second Order Learning

In our basic model we treat the first and second order learning mechanisms on par. That is, our agents attribute equal weights to the direct, first order experience about other trustees and to their second order experience, gained as trustees. Of course, this need not be the case. For instance, some trustor that is skeptic towards the motivations or competence of others might wish to attribute less importance to her second order information. Or, conversely, some agent newly arriving at alien surroundings might assume locals to be better informed than herself and thus primarily try to learn by mimicking their behavior. Such an agent would, presumably, attribute a higher weight to her second order information, the observed behavior of others, than to the gradually incoming stream of her own first hand experiences. Similarly, some agent who does not want to stick her head out, but behave in line with the majority, would attribute a higher weight to her second order experiences, tracking the majority behavior. Here, we are generally interested in the impact second order information has on the emergence of trust. In particular, we want to show again that the qualitative outcomes of our simulation are robust towards minor

Table 6.7: Impact of Second Order Learning



changes in the relationship between first and second order information. To this end we introduce the parameter *second order factor* into our model, measuring how much more or less importance the agents attribute to their second order information, relative to their first order evidence. Treating first and second order learning on par, as we do in our initial model, corresponds to a second order factor of $\gamma = 1$. On the one extreme a second order factor of $\gamma = 0$ indicates that no second order learning is taking place at all, while a second order factor of 2 corresponds to agents attributing double as much weight to social clues than to their own first hand experience. Formally this results in new updating rules for how the agents incorporate new information. For the case of first order learning, the information an agent collects as trustor, we maintain the update rule displayed in equation $(*)$. Second order information, collected as a trustee, is incorporated by the formula

$$\text{trust memory}_{new} = (1 - \beta \cdot \gamma) \cdot \text{trust memory}_{old} + \beta \cdot \gamma \cdot E$$

where β denotes the usual increment while γ stands for the second order factor. The exact value of the second order factor may reflect epistemic considerations, taking others to be more or less knowledgeable, as well as non-epistemic considerations such as a desire for uniformity. Notably, any epistemic conclusion to be drawn from second order information rests on two implicit assumptions about the source of information. The second order information does not give direct access to the trustor's past experience of trustworthiness, but merely to

her willingness to place trust in others. Obviously, this information looses its value if that trustor herself had little or no information to base her beliefs on. Thus, the first assumption we need to make is that the source of second order information is herself knowledgeable. But more is true: Obviously, a trustor that *always* places trust in others, no matter what, or that uses some altogether different decision rule is of not much worth as a source of information. Thus the second implicit assumption underlying social learning is that the observed trustee applies some reasonable decision rule, similar enough to the one of the learner. In situations where it is unclear whether these two assumptions are satisfied, for instance in highly diverse or inhomogeneous societies, agents might be cautions towards their second order information, resulting in a reduced second order factor.

As argued above, the pure first-order mechanism has a bias towards distrust, since it can convert trusting agents into distrusting ones but not vice versa. The second order mechanism, on the other hand, moves the overall trust expectation towards the current majority opinion. If there are more trusting than distrusting agents, the majority of trustees will make the positive experience of being trusted and thus collect a positive feedback. Conversely, if there are less trusting than distrusting agents, the second order experience will be negative on average. All our simulations start with an excess of trusting agents, thus we expect a positive impact of the second order factor on the emergence of trust. As can be seen in table 6.7, this is indeed true, a higher second order factor contributes towards the emergence of trust. However, just as with increment, it is more important to us that the different values of second order learning do not affect the *structural* interplay between the remaining parameters. This holds true as long as the second order factor is not overly low, thus validating our choice of $\gamma = 1$ as a good representative value in the initial model.

Networks

In our basic model, agents are not embedded in any social context, nor do they have any sort of personalized learning mechanism. This unrealistic assumption is now mitigated by the implementation of networks. We allow agents to build up social ties with other agents after making positive experiences. If a trustor encounters the same trustworthy partner twice within a short period of time, that is with at most 9 positive interactions in between, the trustor creates a network tie to this trustee. We see this mechanism not as the emergence of some sort of social groups, but as a more detailed way of learning about

the surrounding environment. In particular, networks are unilateral, thus the trustee will not learn that she has entered any network. Trustors will resort to their trust network[10] when asked to pick between different trustees. If agents have to choose between different trustees, they will preferably opt for agents in their trust network. Thus, increasing the chance of being paired up with trustworthy partners, we expect trust networks to have a strongly positive effect on the emergence of trust. However, we could not produce any such effect in our simulations. Adding networks alone did not have any significant effect on the emergence of trust. The results presented in the next sections, however, will reveal some effects of networks in combination with other factors such as geographical inhomogeneities or external shocks.

Spatial Structure

As a last adaptation of our basic trust simulation we implemented spatial restrictions into our simulation. The baseline simulations have been conducted on a total isotropic space, depicting a homogeneous society without any restrictions on the agents movements. To relax this restriction, we incorporate some spatial structure into the grid, creating protected pockets and confined areas with bottleneck access, as illustrated in figure 6.2. Importantly, we will still assume the underlying society to be connected,[11] thus any square can, in principle, be reached by any agent.

Crucially, spatial restrictions have a major impact on the convergence behavior of our simulations. It is no longer true that all simulations converge towards a state of universal trust or distrust. Rather, spatial restrictions facilitate the emergence of local equilibria, that is local clusters of trust or distrust that emerge independently from the rest of society.

Overall, we find a negative impact of spatial restrictions on the emergence of trust. A further analysis, combining spatial restrictions with networks, shows a slightly positive effect of the latter on the general trust level. However, all of the effects described are rather miniscule compared to the overall number of simulations. A closer analysis reveals spatial structures to have a significant effect only under low values of mobility.

[10] The list of agents as network members can contain at most 20 members.

[11] Recall that agents falling of at the right edge of the grid reappear at the left edge and similarly for top and bottom.

Figure 6.2: Example of a Spatial Restrictions (black)

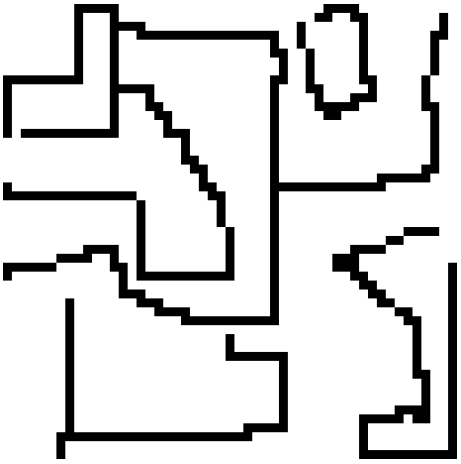
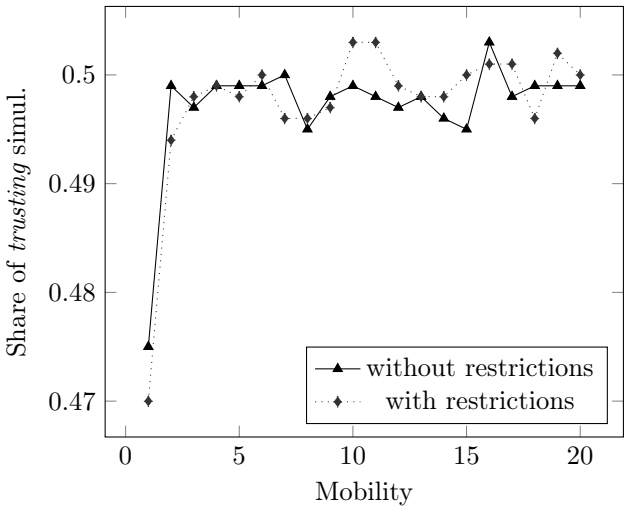


Table 6.8: Spatial Structures and Mobility

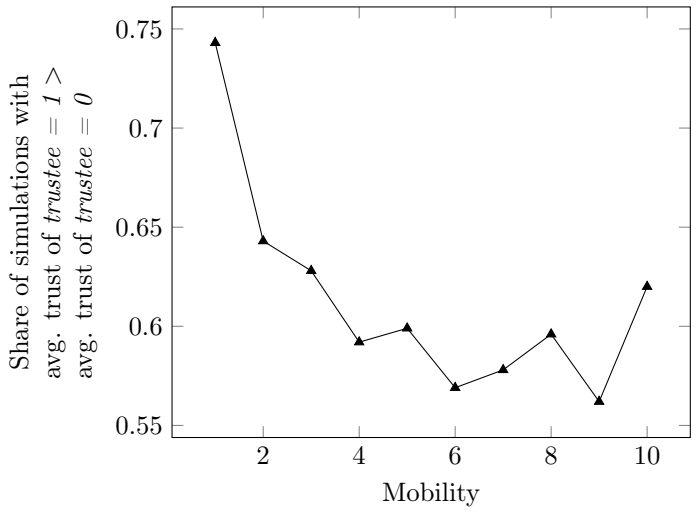


Trust and Trustworthiness, a Puzzle

Finally, we report a puzzling observation about our simulation. Many theoretical frameworks postulate a correlation between an agent's behavior as trustor and trustee. Models such as [57] claim that an increased trust in others fosters one's own propensity to be trustworthy and, vice versa, being trustworthy is correlated with higher expectations towards others. Crucially, no such mechanism was implemented in our model. Just to the contrary, each agent's trustee type is fixed throughout the entire simulation while the behavior as trustor is guided by past experience exclusively. Nor did we model any retaliation mechanism that other trustors could use against trust abusing agents. Our basic simulation did not include any learning mechanism, thus all trustors will be treated equally regardless of their behavior as trustees. Yet, our simulation shows a significant correlation between the agents' trustee types and their expectation about the trustworthiness of others. Given this complete independence between the two roles, we would expect the agents' trust memories, tracking their past experience, to be independent of their trustee type. That is, we would assume it to be as likely as not that the average trust memory of all trustworthy agents is higher than the average trust memory of untrustworthy agents. As can be seen in table 6.9, this does not hold true, but trustworthy agents will have a higher trust expectation than their untrustworthy peers in far over half of the simulations. The effect is strongest at low mobility levels, with trustworthy agents outperforming trust abusers in up to 75% of all simulations. Admittedly, we do not have a convincing explanation for this phenomenon yet. We conjecture that being trustworthy and thus producing positive feedback as a trustee increases the chance of being surrounded by trusting agents and thus indirectly increases the likelihood of receiving positive second order information when acting as a trustee. This conjecture is, however, far from being well tested and we regard this phenomenon as an open puzzle.

Taken together, these results are disillusioning: A multitude of factors that were identified in the literature as being beneficial for the emergence of trust either have no influence or even an effect contrary to what is predicted by the literature, as is the case with spatial restrictions, mobility and networks. In our simulation trust develops best in a society without any spatial or social restrictions and with a high level of mobility.

Table 6.9: Correlation between Trust Memory and Trustee Type

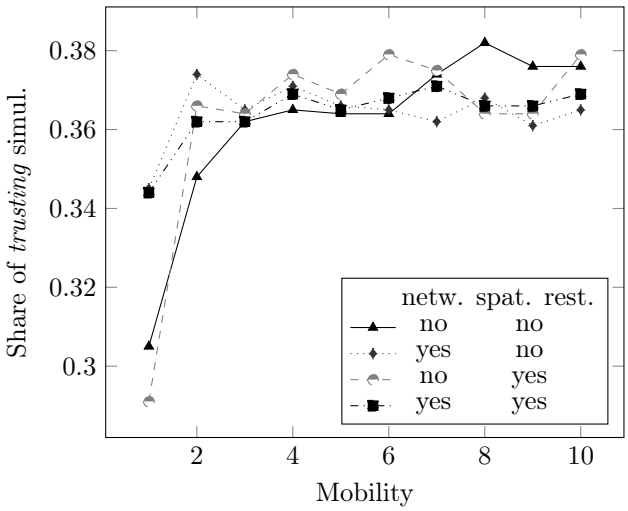


External Shocks

Sometimes external events impact on the agents perception of society. Some prominent event, picked up and repeated by the media, might have a stronger effect on some agent's perception of trustworthiness than even an entire series of private encounters. But also spreading rumors or cleverly designed media campaigns could impact on the agents' propensity to trust others. We will refer to all these events as *shocks*, singular events impacting the agents' propensity to trust that are not based on any direct experience. In principle, such shocks could go in either direction, raising or lowering the general willingness to engage in trust. In the present simulation, we focus on negative shocks only, short term events that thwart the agents' expectation of trustworthiness. After 200 rounds of simulation, we introduce an external shock that diminishes each agent's trust memory by a random amount between 0 and 5 points.

We are not so much interested in the short term impact of such shocks as in their long term effects. Our main topic of interest is when and how such shocks can have lasting effects on the long term behavior of a society. That is, we want to know when such negative shocks can convert some simulation that is underway towards the trusting equilibrium into a state of universal distrust. And, of course, we are also interested in which factors could make societies vulnerable or resilient to long term effects of informational shocks. Here, we are primarily interested in the role of social networks as a tool that could help a

Table 6.10: Networks and Shock Vulnerability



society to recover from an informational shock and to return to its status quo ante.

As it turns out, the relationship between networks and vulnerability to external shocks is intricate: At low levels of mobility, the existence of networks reduces the vulnerability towards external shocks significantly. For instance at a low mobility rate of 1, only 30.5% of all simulations without networks converge towards the trusting equilibrium. Allowing agents to form network ties increases this share to 34.5% of all simulations. However, this pattern reverses as the agents become more mobile. For medium levels of mobility, the existence of networks *increases* the vulnerability of a society to shocks, see table 6.10 for details. Similar results hold in the presence of spatial restrictions. There too, networks in combination with a low level of mobility have a strongly positive effect on the resilience to external shocks, while the effect of spatial restrictions alone is weaker and ambiguous.

6.5 Conclusion and Outlook

Generalized trust is a driving factor for the economic and political success of societies. The factors and determinants relevant for a high level of trust have, by now, been the subject of a vast body of empirical and theoretical research. However, social theories usually have problems in capturing the procedural character of social life. Computer simulations can help to fill this blind spot by reproduc-

ing and exploring the dynamical processes underlying various social phenomena. In this chapter, we have presented a computer simulation on the dynamics of generalized trust. Our agent based model rests on two pillars. The first pillar is a rational choice framework, describing the decision to trust others as a rational decision under uncertainty. The second pillar is a body of empirical and theoretical work on trust, identifying various parameters and mechanisms we incorporated in our simulation.

The primary aim of this simulation was to find conditions and parameters that foster the creation of trust. The research literature identifies four variables as important for the emergence or disappearance of trust. These are the mobility of the agents, spatial structures in a given society that might or might not foster social bonds, social networks and the agents' prior expectations gained through previous interactions or acquired in their socialization process.

Our simulation allowed us to disentangle these factors and study each factor individually with respect to its role in the emergence of trust. In order to set up our simulation, we had to identify further relevant factors about society and the mental makeup of the agents often overlooked in the literature. These factors are the actual share of trust abusing agents within a society, but also the learning mechanism employed by the individual agents as well as their perception of other agents' goals and competences.

In our model, these input factors served as driving forces behind the dynamics of trust. As a first result, we showed that our baseline model reacts to the two central factors, the agents' initial trust expectation as well as the actual share of trustworthy agents, in the way rational choice theory leads us to expect. We take this as a validation of our general model. In the extended model, we obtained surprising results concerning the roles of mobility, networks and geographical heterogeneity for social trust. We interpret these unexpected results as evidence that our theoretical knowledge on the determinants on trust in a dynamic perspective is incomplete and misleading. In particular for the factor mobility, our findings are almost opposite to what is predicted in the literature. These results suggest that the corresponding empirical results might hinge on some hidden variables that are yet to be revealed, such as differences in the actual level of trustworthiness or some crossover between thick and thin notions of trust. On a more conceptual level, our results show that the use of simulations adds a helpful dimension in understanding trust and other social phenomena. In future extensions, we wish to complete this picture by incorporating further relevant influence factors and mechanisms. In particular, we

aim to endogenize and dynamify the behavioral rules guiding the actions of trustees. There, we aim to explore and evaluate different proposed explanations for and mechanisms of the trustees' behavior, ranging from social norms to a game theoretic perspective on legal prosecution.

Chapter 7

Conclusion

To conclude this thesis, I want to offer some general remarks on the *how*, *what*, and *why* of using formal tools in philosophy. Being a concluding chapter, my main aim here is not to offer novel insights into these questions, but to equip the reader with some framing remarks that should help to assess the content presented in the previous chapters.[1] Most of the things said here will be old news to people working in the respective fields, yet they are rarely made explicit. I feel that addressing these topics explicitly can help the reader to see how the work presented in the previous chapters could relate to other debates in the respective areas.

To begin with, let us fix some notation. I will use the words "formal" or "formalization" for any application of formal methods to the replication, representation, discussion or treatment of a given target system. This target system can, in principle, be about anything, a philosophical argument, some piece of human behavior or even a formal or semi-formal theory itself. To be a bit more precise, the term formalization has two readings, a narrow and a wide one. The wide scope refers to any activities related to the replication, representation, discussion or treatment of a target system with formal tools. These activities involve the actual translation into a formal framework, but also choosing and developing an appropriate formal system, preparing the target system for formalization or discussing and justifying the various steps involved. In contrast, the narrow scope reading of formalization refers only to the first of these steps, the actual representation of a target system within a formal framework. Since we are interested in the entire process surrounding formalizations, we will primarily use formalization in the first, wide scope reading, unless specified otherwise.

[1] For a more general introduction on the different roles that formal or scientific methods can play in philosophy and their potential benefits see [105].

Before continuing, we should remark de that the debate about formalization in philosophy shares many features with classic debates about models in science, see for instance [65, 117, 164]. While we will come back to this topic in section 7.3, I should emphasize that these debates depart in at least two relevant aspects from what I am interested in here. First, not every application of formal tools in philosphy, some would say, is, or is related to, a model, at least not in the sense of [163, 164]. And it goes without saying that also the converse holds true: Not every model in philosophy is formal in nature. Second, unlike some of the literature on models, I am not only interested in the outcome of formalizations, be it a model or not, but also in the process leading there including e.g., planning a model, formulating the modeling goals or choosing a formal framework. Chapter 3, for instance, is not directly aimed at constructing a model, but at exploring and comparing different logical frameworks one could use for models of informational states.

There is no unique way of going formal. Just to the contrary, the literature on formal philosophy contains an entire zoo of formal frameworks such as game theory, epistemic logic, Bayesian models or computer simulations to name but a few. Each of these frameworks has its special strength and weaknesses making it fit for some particular applications but not for others. And of course, several of the frameworks may be used to address the same target system. In some cases, interactive belief dynamics for instance, different frameworks may produce competing accounts of the same target system. In other cases, several frameworks need to join forces to address some target system adequately, with, for instance, logical models incorporating some probabilistic insights [39].

For this conclusion, I will concentrate on three particularly prominent formal frameworks from our previous case studies: logical and probabilistic tools and computer simulations. We take these to be good representatives of the wide range of formal methods in the literature. Notably, these different frameworks relate to different target systems, they represent different levels of interest, different modeling intuitions and different ways to their target systems. We will present the strengths and weaknesses of these frameworks individually, but also their potential interaction.

The structure of this concluding chapter is as follows: We will start by discussing a variety of functions that formal models could satisfy before introducing the individual framework families. After discussing the different techniques individually with respect to their *what*, *how* and *why*, we will address the potential interaction of these techniques. We finally present a spotlight section on the

different roles of dynamics patterns in formal models before concluding.

7.1 Why formalization?

Before diving into the *what* and *how* of formalization, we should spend a brief moment on the *why*. What drives the increasing popularity of formal tools? What are the promises, hopes or benefits attached to using formal models in philosophy and related fields? In the following, I present three major types of benefits motivating formal programs, clarification, verification and exploration. In the following we present a list of possible benefits from formalizations. Not all of them will we present in every application of formal methods. At times, several of these goals could or will be pursued in parallel, at others, they will conflict with each other.

In the first case, formalizations aim at *clarifying* various aspects of the target system. This aspect, clarification, is pursued in different ways along the various steps of the formal process. I will present three distinct ways in which formal models can increase clarity about the target system, explication, highlighting and revealing. For the first of these, note that most formal frameworks are phrased in a well defined and highly precise manner. Thus, formalization can serve, or help, to substitute some notions and concepts with highly precise formal concepts, thereby reducing ambiguity about the target system. This first aspect is closely related to *explication* in the carnapian sense [35, 105], that is, the replacing of some inexact, pre-theoretical concept with a more exact one. Second, formalization can help to clarify a target system by *highlighting* or focusing on some relevant structural patterns. Through highlighting certain parts while relegating or omitting others, formalizations allow the modeler to study selected aspects of a target system in isolation and greater detail. Ideally, a well chosen formal model represents some situation in a fashion that facilitates recognizing and understanding patterns and structural relationship within the target system. This second way is loosely connected to idealizations in modeling which we will come back to later in this chapter. Let me give a prominent example: Translating some social situations such as the prisoners dilemma to its underlying game matrix not only removes many aspects of the situation such as the social relationship between the players or their informational and emotional states. Game matrices also make it particularly easy to grasp the strategic structure of the situation, for example by identifying dominant moves or strategies. But, this is the third aspect, the clarifying function of formalization may even start before the actual translation into a formal system. Already

the attempt to formalize some target system can be informative about some of its features. One aspect of a formalization is to represent the target system in a usually highly precise formal framework. Composing such a representation forces the modeler to fill in any gaps or ambiguities present in, for instance, the informally given social theory she is interested in. Sometimes, especially in the social sciences, the corresponding gaps might not even be known prior to the formal endeavor. It is only in preparing the target system for formalization that these gaps are revealed.

The second purpose formalizations can have, besides clarifying, is to be *corrective* or *controlling*. A typical philosophic argument is presented in some informal, everyday language. But is it really conclusive, or does it rest on some hidden assumption, so familiar that we wouldn't notice? Social scientists explain macroscopic behavioral patterns through underlying motivations of the individuals. Are the mechanisms and motivations identified really sufficient to explain the explanandum? These questions can be addressed with the help of formal representations. There are at least two ways in which formalizations can help in controlling some informal theoretical model. First, formal models are highly explicit about their underlying assumptions and preconditions. Thus, a formal model allows to easily keep track of the assumptions a target systems makes or needs to make for deriving a certain conclusion. In the same vein, formal models can help at identifying minimal sets of assumptions necessary for deriving a certain conclusion. Second, formalizations can help in assessing the validity of various inferences in or about a target system. Formal tools allow to study general patterns of valid or invalid inference, as is done in propositional logic. But formal models can also be used to replicate individual inferential or dynamical processes. In particular in the social sciences, many target systems deal with social patterns gradually arising within some temporal system. Computer simulations allow to replicate the underlying dynamic processes of such systems, allowing to test whether some assumed causal or probabilistic relationship between input mechanisms and output phenomena does hold.

The third function of formalizations is *explorative* or *creative*. The use of formal tools is not restricted to a mere representation of some target system. Just to the contrary, many formal frameworks are equipped with a number of methods and techniques inviting to explore a given model even further. These methods or techniques come in many different flavors. Some frameworks, such as game theory, merely represent the target system in a manner easily accessible for further analysis. Other formalizations, in particular those using mathematical

tools, are accompanied by a large fundus of structural insights in the form of lemmas and theorems or inference tricks. Yet other systems, in particular those using computers, explore the target system by solving the underlying equations, but also by offering complex graphical representations. All these techniques can, in many cases, provide additional insights into the target system that would not have been reached without them. Some authors, especially in the literature on economic models [67, 149], compare this explorative use to experiments in the natural sciences. In setting up a formal model, just as in setting up an experiment, a scientists isolates some factors and conditions she wants to know more about [64]. The explorative use of formal models then provides new information about the interplay of these factors and conditions just as conducting the experiment generates a new set of data about a system.

Here is an example to illustrate this point. In 1950, Kenneth Arrow [8] published his famous impossibility theorem about preference aggregation that would later be used by others to cast serious doubt on the possibility of eliciting a general will [137]. Arrow showed that a certain set of innocent looking, highly desirable properties about voting rules is jointly inconsistent. This result would have been close to impossible to discover without the help of formal methods – both for its surprising and counterintuitive character as well as for the complexity of reasoning necessary for deriving it.

Note that the understanding, confirmation or exploration gained from some formal model, once established, persists independently of this formalization. That is, once a particular argument has been established as valid, it can be safely applied by anybody, independently of any knowledge or understanding of the underlying formal tool. To give a concrete example, the Condorcet Jury Theorem, stating that the majority opinion of a group is extremely likely to be accurate if the group is only large enough, has become folklore in social epistemology and parts of political philosophy. Yet, most people actively applying this result would be unable to produce a formal proof. On a related note, also our model in chapter 4 is interested in judgment aggregation. We aim to identify a certain set of conditions under which our differential aggregation rule outperforms straight averaging. In order to apply this mechanism, all that a decision maker needs to know are rough estimates of the different experts' competences. No other knowledge of the formal framework and no understanding of the proofs is required to apply the method successfully. But, of course, there is the usual caveat for any type of tools. Not knowing how they function precisely, in this case not understanding the assumptions or the intuitions behind some proofs,

will increase the risk of misapplying or misusing them.

Throughout her career, each scientist or philosopher develops a toolbox of useful methods and experiences, containing, inter alia, conceptual frameworks, formal techniques, argumentational patterns, computer programs, theorems and many other things. When faced with some new situation or field of inquiry, the scientist can resort to her toolbox, freely applying and combining any techniques she finds useful for solving this problem. Formalizations, as we have just shown, can influence such a toolbox in various ways. First and foremost, formal methods can simply constitute some of the instruments contained in such a toolbox. But even beyond this, formal methods can impact a scientist's toolbox in various, indirect ways. Formalizations can increase the trust towards some of the tools, clarify the exact conditions of application for others, yet newly create third ones.

7.2 Formal paradigms

The list of desiderata for a formal model is long. It should represent all the relevant aspects and mechanisms of the target system while being as clear and simple as possible. And it should ideally be accompanied by some method for exploring the target system and gaining additional insights. This method can be a calculus, some human or automated reasoning system or additional visualization tools. Of course, the exact choice of framework depends upon the target system, but also on the interest of the modeler. In this conclusion we focus on three main families of frameworks widely used in contemporary formal philosophy and the social sciences: Logic and probabilistic tools and computer simulations. We will further offer some remarks on game theory as a particularly widespread modeling framework in philosophy and the behavioral sciences.

The first two are sometimes also referred to as qualitative vs. quantitative models. All these frameworks have been used in the five case studies earlier in this thesis. We start by presenting the three frameworks to be targeted here.

The first framework, *logic*, is the hardest to define. Even within the logic community, there is little agreement of what logic exactly is. Some restrict logic to the study of correct reasoning patterns, while others would even propose a much wider reading than the one presented here. In the following, I outline a rough characterization of logic as a modeling tool. Logic focuses on broad structural regularities within some target system. Typically, logical models concentrate on one particular aspect of a target system, for instance the informational states of agents in interaction or the truth and falsity of individual propositions. Even further, logical models choose one or more degrees of

abstraction in which they want to represent that aspect. A logical model about belief, for instance, may only be occupied with *what* the agents believe. It is not interested in *why* the agents hold their beliefs nor *how strongly* they do so.

The prime vehicle of a logical model is a *logic*, consisting of a formal language for representing the properties of interest together with some account of their structural patterns. So let us have a look at these two defining parts of a logic and how they can be represented.

The first of these two, the language of a logic, is a set of formal expressions such as $a \to b$ or $\Box\varphi$ to describe the target system. A particularly prominent type of language, especially for modeling social interaction, are modal languages as used in chapters 2 and 3. Modal languages enrich the classical propositional language with a set of modal operators, each representing some selected aspect of the target system. Epistemic logic, to give but one example, amend classical logic with monadic operators K_i, where $K_i x$ expresses that agent i knows that x.

Next to its language, a logic also consists of a structural component describing the relationship between the individual formulas. It is this structural information that forms the core of a logic, encoding the properties of the individual terms and operators as well as their interplay. These structural properties can be given in various ways, implicitly or explicitly, using syntactic or semantic characterizations. The former of these, the syntactic approach, presents a logic by identifying certain relationships between the individual formulas of the logic. This can be, for instance, done through axioms or axiom schemes, marking some formulas as always true, but also through a calculus governing the derivation of formulas from each other. Semantic characterizations on the other hand, define the properties of the logic, in terms of their intended applications. A semantic characterization gives some representative set of models or situations for the logic. This set, together with an interpretation function relating the logical language to the individual models, then *defines* the logic. Its structural relationships are precisely those present in all of the designated models.

Depending upon the target system, logical formalizations can take various shapes. In case of a complex theory, say about the aggregation of preferences, an appropriate formalization may be an entire logic representing that theory. If the target systems is something more concrete, a social situation or a particular philosophic argument, a formalization can consist of a particular model (in the sense of model theory) of some appropriate logic, that is some formal object interpreting the respective logical language. Most logical frameworks

used for the analysis of human interaction specify a standard type of model, giving a blueprint for representing target situation, but also allowing to reason semantically rather than syntactically about the logic in question.

For the present purpose, I will use the term logical modeling broadly, referring to any use of formal techniques aimed at, motivated by, employing or working within a logical system. Logical formalization thus comprises the construction and application of a particular logic, the choice and justification of its axioms, but also the representation of some target situation as a model of some logic.

The second framework, *probabilistic* models, refers to the representation of a target situation with the means of probability theory. Probabilistic tools are used for a wide range of applications including the strength of beliefs, reasoning and perception in noisy environments or uncertainty about the behavior or group membership of some individual other. Within our case studies, we have applied probabilistic tools to represent degrees of expertise (chapter 4), but also subjective expectations about interaction partners (chapter 6). Probabilistic models represent certain features of the target system with a probabilistic vocabulary, using terms such as simple and conditional probabilities, mean values and variances. While implicitly assuming that the target system is or can be described by some probability distribution, many statistical models decide to remain silent about some of the details. They rather single out the parameters relevant for subsequent analysis, such as, say, the probabilistic dependency between individual variables or the variance of the distribution. The choice of relevant parameters depends upon the target system as well as on the aims to be pursued in the formalization. To illustrate this, consider the case study on expert judgment presented in chapter 4. While we assume that every expert can, in principle, be represented by some particular probability distribution, we remain silent about the exact distribution. We do not even specify a particular shape of distribution such as for, instance, a normal or a beta distribution. All that is relevant for our model are the first and second moments, the mean and variances of the different distributions and these are the only parameters ever mentioned in our model.

A particular strong point of mathematical models is their inferential capability. Modern day mathematics provides a highly efficient reasoning framework through both, the set of mathematical theorems available as well as widespread familiarity with the probabilistic framework. Representing the target system in a probabilistic language gives access to this reasoning framework for uncovering

structural properties of the target system, reasoning about a particular model or a set of models and predicting their future behavior.

Computational models, the third framework type, refer to any type of formal framework using computation to represent, predict and reason about the target system. I will use the term computational model in a broad sense, referring not only to the actual computational component, but the entire process related, including constructing or choosing a framework in question and justifying the choices made (see [61]). Computational models add towards the understanding of the "what" and the "how" of the target system, solving formal problems inaccessible to analytical methods, but also providing numerical or graphical representations of the target system. An important class of computational models to be found in the humanities (and also the social or natural sciences) are simulations [60, 61], tracking the temporal evolution of some target system. Here computational models help to determine, understand, depict and even manipulate the temporal evolution of some target system (see chapter 6). Other computational models do not immediately track the evolution of a dynamic system, but provide solutions to relevant mathematical problem that could not be obtained otherwise, for instance by calculating the Nash equilibria of certain games [99].

Before advancing to the next session let us spent some words on a further, particularly prominent modeling tool, game theory. In this conclusion I will slightly deviate from the standard use of the term game theory by treating it as a predominantly conceptual rather than formal framework.

7.2.1 Game Theory

From its origins as a study of board games, game theory has developed into a major conceptual framework for representing almost any type of goal directed, interactive behavior be it in coordination, cooperation or competition. Game theory has, over the last years, turned out to be an extremely versatile tool. It has been successfully applied within a large variety of fields, philosophy, economics, computer science, linguistics, addressing a broad range of topics from the analysis of global political summits [133] to the evolutionary emergence of bird cries [166].

Game theory is foremost a conceptual framework. It proposes a particular vocabulary for treating strategic interaction, involving terms such as "players", "moves", "strategies" and "payoffs". Various formalizations and representations are based on this vocabulary, some using logical, other probabilistic tools. It

is only through formal representations that game theory became as powerful a framework as it is. Tree representations and normal forms facilitate the analysis of the strategic structure of different situations, allowing to identify certain moves as rational or dominant, but also to identify similarities between social situations. If two situations, no matter how different in their subject matter, have the same game representations, they are, in a certain sense, structurally similar. The use of formal tools has helped to formulate, but also to discover various concepts and strategies such as dominance, equilibria or backward induction that heavily impacted our current understanding of social interaction. Individual concepts such as Pareto optimality have even escaped the technical surrounding in which they were originally defined in and long since made it into ordinary language use.

The strategic structure is, of course, only one aspect of a game situation. There are other relevant aspects, such as the social status of the players, their risk attitudes or their informational states. To illustrate this, imagine a striker shooting a penalty in a football game. For choosing between the right and left corner of the goal he will not so much refer to the strategic structure, but to his expectations about which corner the goalie will cover or how the latter would reason about him. For a more inclusive analysis of such a situation and the resulting actions, game theoretic models need to be complemented with formalizations of other relevant aspects. This is the starting point of *epistemic game theory*, a relatively young program aimed at formalizing the interaction between the game structure and the players' epistemic states. Epistemic game theory, just as classic game theory, can be represented with probabilistic [131] as well as logical [125] means. For instance, our first case study in chapter 2 presents a logical model of epistemic game theory, aiming to additionally incorporate the dynamics of agents' beliefs.

7.3 Formalizations, Models and Validity

The term formalization does not refer to a single activity, but to an entire set of processes, choosing a framework, preparing the target system for representation, creating some particular representation, reasoning in or about it and validating the model obtained. All of these steps may involve substantial choices and fierce debate. The various choices may, in some cases, reflect different perspectives or interests of the modeler, but they may also be relevant for later steps such as the justification or validation of a model. In the following, we will inquire further into two central steps in the formal processes, choosing an adequate

formal framework and validating a model.

7.3.1 Choosing a framework

A central step in the formal endeavor is the selection of an adequate formal framework. A sculptor setting out to carve a figure will have to decide between different materials, say clay, stone or marble. Similarly, a formal modeler trying to represent some situation will have to choose between different frameworks. This problem of choosing a framework was, for instance, the underlying motivation behind chapter 3. There, we compared different logical frameworks with respect to some properties a modeler might be interested in.

Picking a framework shares many features with the selection of a scientific theory. A formal framework, just as a scientific theory, makes some choice about which features are treated as primitive and which as derived. But the choice of framework, again just like a scientific theory, also imports certain structural assumptions through its underlying axioms and definitions. For instance, by choosing to represent some given epistemic situation within an S5 epistemic logic, we automatically buy into full positive and negative introspection and common knowledge of the situation for all agents, see also the discussion in chapter 3. In line with Kuhn's findings about theory choice, also the selection of formal framework is guided by many, potentially opposing considerations. We will come back to this point in discussing logical models, later in this section.

A particularly prominent use of formalizations is to create and explore representational formal *models*, goal directed representations of some target system. Consequentially, the philosophic debate about scientific models has much to offer towards discussing and understanding formalizations [60, 65, 117, 164]. An important example, relevant for the discussion of adequacy or validity, is the distinction between abstractions and idealizations [116, 147, 162], sometimes also described as Galilean resp. minimal idealizations. A formal model, as most other types of models, strips of certain parts and features of the target system, concentrating on others. The way in which this stripping of happens varies substantially from case to case. Some formalizations aim at what is dubbed an abstraction or Galilean idealization, ideally only omitting aspects irrelevant to the pattern or mechanism in question. Others are (minimal model) idealizations, intentionally misrepresenting or omitting some relevant factors or mechanisms. For understanding complex situations, it can be easiest to study the mechanisms involved separately, even if, in nature, they always appear conjointly. Similarly, it can be instructive to study some highly idealized or distorted, yet concep-

tually or mathematically tractable setting in order to identify or explore some possible mechanisms, see [149]. To give a concrete example, in chapter 6 we presented a computational model on the emergence of trust in societies. In this model, agents learn about the value of trust only through their own first-hand experience, discarding other factors, such as communication networks, gossip or observing the behavior of others, that might be equally relevant for the emergence of trust. The aim of this model is not to get a complete or realistic picture on the emergence of trust, but to understand one of the mechanisms involved in detail. Later, this understanding is to be combined with results on other mechanisms such as belief dynamics through personal interaction or the influence of media towards an integrated understanding of the dynamics of trust. In the following, I will discuss the formal paradigms studied, logical, probabilistic and computational tools, closer with respect to the individual steps involved in the formal process and the relevant choices therein.

Among the most pressing problem in logical modeling, our first framework, is the choice of a suitable logic. Even if restricting oneself to already existing frameworks, one might easily find many different logics for the same target system. In a recent study, Herzig [79] has identified no less than nine different logical frameworks for the strategic interaction of several agents. Comparing different logics, it turns out that some frameworks are close to each other. A particular logic may, for instance, be a mere refinement of another or two logics may be interdefinable, that is expressing the same properties in different languages, taking different aspects as primitive. But it can also happen that various logics focus on distinct parts of the same target system, rendering them mutually incomparable.

The choice for a particular logic is guided by a variety of criteria. To name just a few, these criteria can be simplicity, expressive power, succinctness, scope, computational complexity, decidability and, of course, the faithfulness or fit. As in the case of Kuhnian theory choice, there is no designated way of balancing these criteria. The ultimate decision is left to the modeler. Some of these criteria are naturally compatible with each other, while others, such as expressive power vs. complexity or realizability, almost inevitably pull in different directions, as we have shown in chapter 3. The relative importance attributed to the individual factors will depend upon the target system or the modeler's taste, but also upon the intended applications. When picking a logic for automated reasoning or software verification, computational complexity will feature high on the list, while the creation of formal models for philosophic clarification asks

for succinctness and naturality in the operators used.

While a discussion of these criteria transgresses the scope of this article, one final remark is in place about the dimension of fit. An explanatory use of modeling does not always strive for a faithful representation. Some would even argue explanatory successful models will never be faithful [59, 64, 148], but they need to abstract away from various aspects of the situation and focus only on some particular relevant properties. In this sense, the dimension of fit is, strictly speaking, not to be interpreted as a fit to the entire data or phenomenon, but as an adequate representation of these designated properties. To give an example, simple belief logics only distinguish those propositions an agents considers possible from those she doesn't. The fact that such a framework fails to distinguish propositions the agents judges barely possible from almost certain truth does not constitute a failure of fit. The model is simply placed at a degree of abstraction that ignores this difference. A failure of fit, in contrast, would occur if the target system significantly violated some axiom of belief logic such as positive introspection.

Probabilistic formalizations, our second framework, represent uncertainty in or about the target system, graded beliefs, or actions with uncertain outcome, within the language of probability theory. One of the difficulties with such mathematical models is that they can be, somehow, too precise. While logical frameworks allow to stay at a safe degree of abstraction, probabilistic models require the modeler, at least in principle, to be very precise about various aspects of the target system. It is, however, behaviorally or conceptually, difficult to distinguish between an agent holding some proposition to be 63% probable and holding the same proposition 63,7% probable. Nor should most arguments, one could argue, hinge on these precise numbers rather than, say, the rough structural relations between the different beliefs.

Let me illustrate this further with an example. Consider the notorious muddy children puzzle. An epistemic logician will address this situation with a *single* model and argue that it covers the entire situation, see [118]. A Bayesian, on the other hand, will not be done with one particular Bayesian model, specifying priors for all agents involved. Rather, the Bayesian will have to show that her solution of the puzzle holds for *all* initial beliefs the children could reasonably entertain. To put it more generally, the modeler has to show that the solution put forward holds for a set of worlds, big enough to credibly contain the actual world [149]. To give a different example, consider the framework we presented in chapter 4. We start with the assumption that the individual degree

of expertise is well defined, yet extremely difficult to elicit. Thus, we are not interested in identifying the ideal weights for every agents. Rather, we aim to find a set of conditions on the individual weights that still guarantee differential judgments to outperform straight average. These conditions should be weak enough that some decision maker with limited information about the individual experts could still be positive that his weight assignments satisfy them.

Our third framework, computational formalizations, represents, depicts or explores a given target system with the use of a computer. Computational models are particularly widespread in the study of social interaction as an important tool to understand, test and explore various theoretical accounts. Often, an informally given theory omits various factors loosely related to the target system. For instance, a theory about the emergence of trust presented in chapter 6 might be silent about the exact learning algorithms agents use. For a non-formalized theory, this counts as a advantage, a theory only becomes stronger if it does not rely on a particular background framework. However, for a computational model this poses a problem. In order to implement the theory, we need to fix some learning rule. Consequentially, the step of preparing a target system for formalization is particularly involved for computational models. Each relevant aspect of the situation needs to be specified in such a way that it can be implemented in a computer program. Crucially the mechanisms used to fill the gaps of the original model are in need of justification themselves. They may, in case, be borrowed from other theoretical accounts about, say, learning rules or communication networks. Consequentially, many simulations cannot test the particular target system in isolation, but only in conjunction with the chosen background theories used for filling these gaps.

The individual input mechanisms of a computational simulation need to be formulated in a fashion accessible to a computer, using for instance, the vocabulary of logic, probability theory or graph theory. Thus, computational simulations, in a certain sense, supervene on some prior formalizations. They do not, however, require homogeneity in the input frameworks. Within the same simulation there might be a probabilistic mechanism for the individual's beliefs, a logical mechanism for choosing an interaction partner and a graph theoretic mechanism for social influence.

Finally, we want to offer some remarks on the use of game theory. As stated above, we understand game theory as a conceptual framework for addressing interactive strategic situations. For setting up a particular model, this conceptual framework has to be combined with a formal framework, logical, probabilistic

or other. The exact choice of framework is guided by various aspects of the situation [69]: The complexity of the problem, risk attitudes and reasoning types of the agents or the quantity and quality of available information. Representing a greedy investor in, say, a stock market, who aims at maximizing her expected gains requires a different formalization than a conservative investor occupied with reaching and maintaining a certain status. In some situations the mere possibility of some outcome, say bankruptcy or a car accident, is sufficient to disqualify certain actions of the agents. These cases require a different representation than others where more fine grained information is needed for making a choice. Some of these cases call for a probabilistic treatment, others are best represented with logical methods.

Probabilistic models require uncertainty about the game to be quantifiable. This uncertainty can refer to necessary background information, but also the distribution of player types or expected behavior of other agents. It can even be beneficial for the playing agent to create uncertainty about her own behavior, using randomized strategies or bluffing in order to be less predictable and maximize long term expected gains. Probabilistic studies of games are primarily related to optimizing behavior, modeling agents as maximizing their expected utility, but also characterizing societal outcomes produced by the interaction of such agents.

Logical or qualitative models do not require the underlying uncertainty to be quantifiable. These models apply to situations of little information about the underlying situation, but also to the analysis of cautious agents, reasoning about a particular goal state to maintain rather than maximizing their expected gains. Logic or qualitative models are also used to track complex reasoning chains with several inference steps about various agents. These could be the iterated removal of dominated strategies, gradually discarding all those moves that are guaranteed to fare worse than some other options the agents have. Another case of iterated reasoning is backward induction, evaluating a strategic situation by gradually inferring back from the set of possible outcomes to the optimal moves and strategies to play.

7.3.2 Formalizations and Validity

Formalizations should, so the hope, inform us about the properties and behavior of their target system. Hence, we need to infer back from some formal representation to its target system. We need to ensure thus, in order to validate this inference, that the formalizations is similar enough to the target system,

see [68, 149]. There is a variety of approaches for assessing whether a formal framework is good enough in that sense. First, some axioms and assumptions underlying the construction can receive *direct* support. The underlying axioms of a logical framework, for instance, may be supported by some philosophical arguments, while some mechanisms of computational models may be borrowed from a well established theoretical framework. Second, a formal framework can be tested for its *external* consistency with observations or other theories [59, 68]. Probabilistic and computational models make predictions about target systems, that can, in some cases, be evaluated against outside data. Take for example Monty Hall's paradox, a fiercely debated puzzle [142] about the optimal behavior in a game show. This historically existing game show could, in principle be repeated often enough to produce a set of data that every theoretical prediction has to stand up against. Third and finally, formal frameworks may also turn out to be *internally* inconsistent. That is, they might produce puzzling, unwanted or incoherent patterns such as the liar's paradox. Another example of this type is our criticism of the approach by [6], presented in chapter 5, pointing out an internal inconsistency in their treatment of approval voting.

Closely related to the validity of a formal tool is its scope. Not every model is intended for every possible situation it could be applied to. Just to the contrary, formal frameworks come with a set of explicit or implicit preconditions restricting the domain of applicability [164]. Newtonian physics has a high degree of predictive, manipulatory and explanatory success for many applications, but only as long as the parameter values are within a classical range. Similarly, qualitative models of belief are a successful tool in modeling the informational dynamics of a police agent slowly acquiring evidence about a criminal case. They completely fail in describing the same agent figuring out whether a certain coin she finds at the crime scene is fair or not by tossing it over and over. Qualitative models of belief are neither fit for, nor aimed at representing the set of mental states relevant for representing the testing of scientific hypotheses. Similarly, the agent based model presented in chapter 6 is specifically tailored at the phenomenon of thin trust. It is neither aimed at, nor suitable for an adequate representation of thick trust.

But what if some model turns out to be inadequate or invalid? Let's assume for a moment that we discover some problems with a formalization, either some internal inconsistencies or a discrepancy with some data or expectations. In this case, there are at least three possible reactions. First, we might take this discrepancy to be informative about the target system, thereby challenging

some prior beliefs or theories. Second, we might restrict the scope of our formal framework to exclude the paradoxical application and third and finally, we may decide to alter or discard the formal framework altogether. Let me illustrate each of these cases with a concrete example. For the first case, learning about the target system, consider Schelling's famous segregation model. Within this model, Schelling showed that even a weak degree of preference for uniformity among the individual agents can create a highly segregated society. Albeit highly idealized, this model was not rejected as unrealistic, but changed our understanding of geographical segregation drastically [149]. Similarly, the results on the influence of mobility in chapter 6 inform us about the influence of this factor on the underlying mechanism, thereby revealing some shortcoming in current theories on the determinants of trust. For the second case, restrictions of scope, we refer to Sorites' paradox. Soritean reasoning derives highly undesired conclusion by applying classical logic to vague terms such as being a heap or being tall. Sorites is, in general, not taken to undermine the fact that classical logic tracks the preservation of truth and falsity of statements. Rather, Sorites is understood as revealing certain caveats in applying classical logic to a combination of vague terms and high reasoning depths. Finally, examples abound for the third case, counterexamples leading to discard certain formalizations. The development of deontic logic, for instance, has progressed along a chain of challenges and counterexamples, each rejecting one or more proposed logical frameworks. To pick a concrete example, Forrester's Paradox of gentle murder revealed the inadequacy of Standard Deontic Logic as a framework for deontic reasoning in general, and for conditional obligations in particular. The paradox contributed to the emergence and development of alternative competing frameworks for deontic reasoning, such as non-normal deontic logics and logics that do not validate modus ponens (see [5, 72]).

7.4 Interplay between the Paradigms.

In the last chapters, we have introduced three types of modeling frameworks, logical and probabilistic tools and computational approaches. By now, we have addressed these frameworks in isolation, focusing on their individual properties and particularities. But how do the paradigms relate to each other? Of course, the particular strengths and weaknesses of the individual frameworks often gear them towards different target situations or modeling interests. There is, however, a plethora of situations that attract the attention of more than one formal paradigm. In this case, there are, at least, three different ways in which the

different frameworks could relate to each other. First, different models may complement each other, giving different perspectives of the same target system. Second, they could compete against each other, providing contradictory analyses or merely quarreling about the status of being the standard framework for certain applications. And third and last, different frameworks may build on each other by importing building blocks from other formalizations. Let us study these three possibilities a bit further.

In the first case, the different models act in parallel, each looking at the target situation from a different angle. In some cases, the various formalizations merely differ in their coarsity. Different degrees of abstraction, so the hope, reveal different structural properties relevant for understanding the target system. To give a concrete example, there are logical and mathematical formulations of non-locality in quantum information [1] or of the first and higher order beliefs structures in games [52, 73]. In both cases the two formalizations are completely compatible with each other. The logical formulation simply presents a coarser perspective of the information captured in the probabilistic model, thereby emphasizing different aspects of the system. The transition from the finer to the coarser perspective is easily done by simply omitting some of the information. Similarly, logical and probabilistic methods study games in different degrees of coarsity. Also there, the resulting analyses are, at some times compatible. For instance, those moves maximizing the expected utility of an agent are always a subset of the strategies surviving any iterated removal of dominated strategies

In other cases, the different formalizations assume genuinely different perspectives towards the target system. In this case, there is, in general, no direct relationship between the different formalizations, even though they study the same phenomenon. In the times of facebook, for instance, there is an increased interest in the spreading of information on communication networks. Computational models, for once, simulate the entire dynamic process to get a grasp of the belief dynamics and its limits [119]. On the other hand, mathematical and logical models [39] identify conditions of the communication graph under which the same system converges to a stable equilibrium.

In our second case, different formal frameworks produce competing formalizations of the same target system. This competition may sometimes be purely formal, say about being the *standard* paradigm for some application. But the conflict might also mirror some *substantial* disagreement, for instance about some modeling choices or background theories. To give a concrete example, the Bayesian program models beliefs as subjective probabilities, while a logical

analysis represents beliefs through truth-valuations on sentences or sets of sentences. The relationship between these two has given rise to deep and lasting philosophical debates, for instance about the connection between probabilistic belief and logical concepts such as conditionals [107] or, more recently, the relationship between probabilistic and logical models of static belief [104] and the dynamic correspondence between logical and probabilistic belief revision rules [109]. Note that the debate displayed here is not so much a conflict between logical and mathematical tools, but the reflection of some substantial philosophic dispute about the nature of belief. Other logical tools for instance, manage to avoid this conflict with the Bayesian paradigm. In particular, a more semantically motivated logical model, using possible worlds, is perfectly compatible with probabilistic models of belief.[2]

In our third and last case, different frameworks are used interdependently, one building on the results or another. Within a complex target system, there is a variety of different factors and mechanisms at work. Naturally, some of them might call for, say, a probabilistic model while others are best fit with a logical model. An agent will have a probabilistic degree of belief whether it is going to rain tomorrow, but her actions are best described in an all-or-nothing, logical fashion: She will go to the beach or she won't. Furthermore, the individual parts of a complex system will, in general, not be isolated, but they interact in several ways. Thus, also the corresponding formalizations might need to interact in a similar or related way. To give a concrete example, consider informational cascades, an informational feedback phenomenon that can lead a group to a unanimous belief in some falsehood. Cascades crucially depend upon the agent's belief updating mechanism which is, for this purpose, best described with a probabilistic model. But it takes a logical model, supervening on this probabilistic belief model, to show that no degree of rationality can protect against cascades. Even completely rational agents with common knowledge of the situation and each others' rationality, may still run into a cascade that leads them to a uniformly false belief [10].

More broadly, most computational models depend upon prior formalizations of their respective input mechanisms. Computer simulations can track the emergence of trust (chapter 6), but also political opinions [77], radicalization [15],

[2]To give some more details: Every rational valued probability ascription to a finite set of propositions is representable as a set of equiprobable possible worlds and can thus be reasoned about in classical epistemic logic. It is folklore in epistemic logic that this analogy extends to Bayesian conditionalization corresponding to public announcements (of the formula conditioned on) and Jeffrey conditioning that can be replicated with product update models (see chapter 2 for a definition). In the last case, the exact choice of product update model will, however, need to depend upon the base model it is applied to.

or racial segregation [139] in a society. But the underlying learning or decision rules are, of course, both formulated and defended in a logical or probabilistic framework. Similarly, many of the theoretical discussions in setting up and defending agent based simulations are phrased in game theoretic terms, addressing the space of possible actions, the agent's goals, beliefs and preferences, which strategies of the players to include and how they should behave. These discussions finally results in identifying and defending a certain fragment of the game theoretic framework to be used in the actual implementation. Depending upon the scope of the simulation, this fragment might be little more than some trivial consequence of the game theoretic framework, some belief updating mechanism or a decision rule. For instance, the simulation in chapter 6 breaks the complex considerations involved in trusting others down to a single decision rule that translates the agent's past experience into her current behavior.

7.5 Formalizations and Dynamics

Political scientists want to understand how societies build up and destroy social capital, trust or shared norms. Social epistemology studies how preferences and beliefs gradually develop in social interaction. Philosophers such as Peirce or Levi and, later, the Bayesians assess the rationality of an epistemic agent through her reaction to incoming information. All these examples deal, directly or indirectly, with dynamic aspects of their target systems. Across various fields, dynamic or cross-temporal aspects of various kinds have become more and more central in recent research.

This emphasis on dynamical aspects reflects in a variety of developments in the formal realm, ranging from the emergence of computer simulations all the way to the *dynamic turn* in modal logic, producing a plethora of logical tools and frameworks related to various aspects of dynamic systems. In particular the emergence of simulations proved vital for the study of dynamic interaction. Much of our current understanding of social dynamics would not have been possible without the appropriate formal tools, such as agent based simulations.

The relationship between dynamics and formalization is manifold. Formal models appear in various flavors, using many different frameworks including all three major framework types discussed here. In the end, the temporal evolution of various systems becomes a target system in itself, to be addressed with formal tools from various perspectives. There are at least four different ways in which formalizations can relate to dynamic aspects of a target system. First, formal models can merely be suited for incorporating the effects of dynamic

events, allowing for the transition from some prior state to the posterior after some event. But, this is the second case, formalization can also focus on the representation of dynamic events and patterns themselves, rather than solely representing their effects on a situation. Third, formalizations can replicate the temporal evolution of dynamic systems. And fourth and last, formalizations can be used to reason about the limit behavior of some dynamic process and anticipate and explore different future routes the system could take. Let us explore these alternatives in a bit more detail.

At some times, this is the first case, a static representation of an agent's belief set is to be combined with a formal tool that automatizes the transition from one situation to the next, rather than having to compose a completely new model upon every bit of information the agents receive. Preparing a formalization for incoming changes comes at a price. Accommodating dynamic events poses new conditions and requirements that need to be taken into account while picking a formal framework. Some incoming information might contradict an agent's current beliefs, requiring her to have some fallback states to resort to. Reacting adequately to such information will require insight into the causal or probabilistic relationship between different features of the world or information about the relative strength of certain beliefs. These desiderata, sometimes dubbed the "logic of change" reflect in the development of various logical and mathematical frameworks such as knowledge-belief models in logic [12] or Bayesian Networks for probabilistic models [27].

In the second case, the rules, patterns and events governing the dynamic process themselves are subject to formalization. Depending upon the choice of target system, the dynamical patterns to be formalized can be actual changes in actions or beliefs, but also normative theories on how agents *should* ideally respond to incoming information. Formal frameworks related to the patterns of belief revision include the AGM axioms [3] or Bayes and Jeffrey updating in the probabilistic case. Other frameworks such as Bayesian sensor integration focus on incorporating information from several partially reliable and potentially contradictory sources at once, see [96] or the models presented in chapters 4 and 6.

Yet other formalizations concentrate on individual updating events, treating them as entities in their own right. To motivate this, imagine a politician composing different speeches for her upcoming election campaign. She wants to prepare various manuscripts, differing in style and content, to be prepared for the different possible audience she may encounter until election day. Every

speech updates the beliefs and attitudes of the respective audience in some way or another. In a certain sense, a political speech can hence be seen as an dynamic event in its own right that can be applied to various possible audience and occasion. That is, the candidate writing her speeches composes an updating event without referring to any particular listener or any particular informational state. Treating updating events as individual entities allow us to represent, compare and reason about dynamic events independent of any actual state of nature. Formal tools to model these events are product updates and public announcements in logic ([11] and chapter 2)or transition matrices (see chapter 5) within mathematical modeling.

In the third, case formalizations are aimed at depicting, predicting and understanding the entire dynamic process. This third aim is primarily pursued by computational simulations which are "designed to imitate the time-evolution of a real system" (Hartmann [75]). We focus on two roles served by computational models, solution or exploration and visualization. While dynamic systems are formulated in a mathematical or logical language, various aspects of dynamic systems are too complex for an analytic treatment. Many dynamic accounts of group behavior are formulated in an agent based language, describing the behavior of an individual agent at a single time step (see chapter 6, but also [15, 77, 119]). When applied to several interacting agents, these individual rules generate complex behavioral patterns, that are close to impossible to anticipate let alone classify with analytical means. By executing the underlying algorithm, computational models *solve* the formal interactive system, they determine the long term behavior that emerges from the chosen agentive and interaction rules. The second function served by computational models is visualization. Computational tools enable different perspectives on the target system. A simulation can produce numerical output, representing certain aspects of the dynamic system. But computer programs can also visualize the dynamic system, for instance by representing individual agents as functions or moving points on a map. Each such perspective allows for different insights into the target system. For instance, visually tracking the emergence of a multi-agent system on a two-dimensional grid can inform about typical interactive patterns or iterating clusters that are hard or impossible to identify in the numerical output stream produced by the simulation alone.

At yet other times, this is the fourth case, formalizations are to reason about global or limit properties of the target system. When encountering some new trend, say a newly created party, a technological standard or the informational

shocks studied in chapter 6, we might be interested whether this trend will manage to acquire a stable and lasting level of support. Or, as a second example, in dealing with a complex system or a computer program we want to be sure that the program does not crash along the way. These questions refer to global or limit properties of a dynamic system rather than to particular events at individual time steps. The limit behavior of a system can be anticipated through computer simulation, imitating the dynamic process itself. At other times, formalizations identify adequate structural properties of the underlying situation governing the convergence behavior of some system such as Nash equilibria in strategic contexts [99] or general properties of the communication graph underlying the adoption of certain trends. [39]. Similarly, the stability of complex systems can be tested, inter alia, through a logical analysis of some transition graph encoding the possible runs of the system [58].

7.6 Final Remarks

In this concluding chapter, we have offered some general remarks about the role of formal tools in philosophy. Our main motivation here was to provide some framing for the five main chapters in this thesis and to give some indications on how they could relate to other formal and informal work in their respective fields. In this discussion, we have concentrated on three types of formal tools used in this thesis, logical and probabilistic models and computer simulations. Of course, there are others and there are other roles of formal tools than the ones presented here [105]. We do, however, think that these remarks give a fair overview over some of the decisions and considerations involved in the use of formal tools.

Bibliography

[1] Abramsky, S. and L. Hardy (2012). Logical bell inequalities. *Physical Review A 85*(ARTN 062114), 1–11.

[2] Aceto, L., W. van der Hoek, A. Ingólfsdóttir, and J. Sack (2011). Sigma algebras in probabilistic epistemic dynamics. In *Proceedings of the Thirteenth conference on Theoretical Aspects of Rationality and Knowledge, ACM, 2011, pp. 191–199.*, pp. 191 – 199.

[3] Alchourrón, C. E., P. Gärdenfors, and D. Makinson (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic 50*(2), 510 – 530.

[4] Alexander, J. M. (2007). *The structural evolution of morality.* Cambridge University Press Cambridge.

[5] Åqvist, L. (2002). Deontic logic. In *Handbook of philosophical logic*, pp. 147–264. Springer.

[6] Aragones, E., I. Gilboa, and A. Weiss (2011). Making statements and approval voting. *Voting Theory and Decision 71*, 461–472.

[7] Armstrong, J. S. (2001). Combining forecasts. In *Principles of forecasting*, pp. 417–439. Springer.

[8] Arrow, K. J. (1950). A difficulty in the concept of social welfare. *The Journal of Political Economy*, 328–346.

[9] Aumann, R. (1999). Interactive epistemology I: Knowledge. *International Journal of Game Theory 28*, 263–300.

[10] Baltag, A., Z. Christoff, J. U. Hansen, and S. Smets (2013). Logical models of informational cascades. *University of Amsterdam, http://staff. science. uva. nl/˜ ulle/teaching/lolaco/2013/papers/smets. pdf*.

[11] Baltag, A., L. Moss, and S. Solecki (1998). The logic of common knowledge, public announcements and private suspicions. In I. Gilboa (Ed.), *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)*, pp. 43 – 56.

[12] Baltag, A. and S. Smets (2006). Conditional doxastic models: A qualitative approach to dynamic belief revision. *Electronic Notes in Theoretical Computer Science 165*, 5–21.

[13] Bates, J. M. and C. W. Granger (1969). The combination of forecasts. *Or*, 451–468.

[14] Baumann, M. R. and B. L. Bonner (2004). The effects of variability and expectations on utilization of member expertise and group performance. *Organizational Behavior and Human Decision Processes 93*(2), 89–101.

[15] Baurmann, M., G. Betz, and R. Cramm (2014). Meinungsdynamiken in fundamentalistischen gruppen. erklärungshypothesen auf der basis von simulationsmodellen. *Analyse & Kritik 36*(1).

[16] Ben-Zvi, I. and Y. Moses (2011). Known unknowns: time bounds and knowledge of ignorance. In *TARK*, pp. 161–169.

[17] Bereby-Meyer, Y. and I. Erev (1998). On learning to become a successful loser: a comparison of alternative abstractions of learning processes in the loss domain. *Journal of mathematical psychology 42*(2), 266–286.

[18] Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, reciprocity, and social history. *Games and economic behavior 10*(1), 122–142.

[19] Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

[20] Bicchieri, C., E. Xiao, and R. Muldoon (2011). Trustworthiness is a social norm, but trusting is not. *Politics, Philosophy & Economics 10*(2), 170–187.

[21] Birk, A. (2001). Learning to trust. In *Trust in Cyber-societies*, pp. 133–144. Springer.

[22] Blackburn, P., M. de Rijke, and Y. Venema (2002). *Modal Logic*. Cambridge University Press.

[23] Bonanno, G. and P. Battigalli (1999). Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics 53*(2), 149–225.

[24] Bonner, B. L. (2000). The effects of extroversion on influence in ambiguous group tasks. *Small Group Research 31*(2), 225–244.

[25] Bonner, B. L. (2004). Expertise in group problem solving: Recognition, social combination, and performance. *Group Dynamics: Theory, Research, and Practice 8*(4), 277.

[26] Bonner, B. L., M. R. Baumann, and R. S. Dalal (2002). The effects of member expertise on group decision-making and performance. *Organizational Behavior and Human Decision Processes 88*(2), 719–736.

[27] Bovens, L. and S. Hartmann (2004). Bayesian epistemology. *OUP Catalogue*.

[28] Brandenburger, A. (2008). Epistemic game theory: Complete information. In S. N. Durlauf and L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics*. Palgrave Macmillan.

[29] Brandenburger, A. (2010). Origins of epsitemic game theory. In V. F. Hendricks and O. Roy (Eds.), *Epistemic Logic: Five Questions*, pp. 59–69. Automatic Press.

[30] Brandenburger, A. and A. Friedenberg (2010). Self-admissible sets. *Journal of Economic Theory 145*, 785 – 811.

[31] Bray, J. R. and J. T. Curtis (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological monographs 27*(4), 325–349.

[32] Brennan, G. and L. Lomasky (1993). *Democracy and Decision – The Pure Theory of Electoral Choice*. Cambridge University Press.

[33] Burt, R. S. (2000). The network structure of social capital. *Research in organizational behavior 22*, 345–423.

[34] Buskens, V. (2002). *Social networks and trust*, Volume 30. Springer.

[35] Carnap, R., R. Carnap, and R. Carnap (1962). Logical foundations of probability.

[36] Castelfranchi, C. and R. Falcone (2000). Trust is much more than subjective probability: Mental components and sources of trust. In *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on*, pp. 10–pp. IEEE.

[37] Castelfranchi, C. and R. Falcone (2010). *Trust theory: A socio-cognitive and computational model*, Volume 18. John Wiley & Sons.

[38] Chandy, M. and J. Misra (1985). How processes learn. In *Proceedings of the 4th ACM Conference on Principles of Distributed Computing*, pp. 204 – 214.

[39] Christoff, Z. and J. U. Hansen (2013). A two-tiered formalization of social influence. pp. 68–81.

[40] Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting 5*(4), 559–583.

[41] Coleman, J. S. (1994). *Foundations of social theory*. Harvard University Press.

[42] Cooke, R. M. (1991). Experts in uncertainty: opinion and subjective probability in science.

[43] Cordón-Franco, A., H. van Ditmarsch, D. Fernández-Duque, J. J. Joosten, and F. Soler-Toscano (2012). A secure additive protocol for card players. *Australasian Journal of Combinatorics 54*, 163–175.

[44] Davis, J. H. (1973). Group decision and social interaction: A theory of social decision schemes.

[45] de Bruin, B. (2010). *Explaining Games: The Epistemic Programme in Game Theory*. Springer.

[46] Dean, W. and R. Parikh (2011). The logic of campaigning. In *Logic and Its Applications*, pp. 38–49. Springer.

[47] DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association 69*(345), 118–121.

[48] Downs, A. (1957). *An Economic Theory of Democracy*. Harper and Row.

[49] Einhorn, H. J., R. M. Hogarth, and E. Klempner (1977). Quality of group judgment. *Psychological Bulletin 84*(1), 158.

[50] Elga, A. (2007). Reflection and disagreement. *Noûs 41*(3), 478–502.

[51] Fagin, R. (1994). A quantitative analysis of modal logic. *Journal of Symbolic Logic 59*(1), 209 – 252.

[52] Fagin, R., J. Geanakoplos, J. Halpern, and M. Vardi (1999). The hierarchical approach to modeling knowledge and common knowledge. *International Journal of Game Theory 28*(3), 331 – 365.

[53] Fagin, R., J. Halpern, Y. Moses, and M. Vardi (1995). *Reasoning about Knowledge*. The MIT Press.

[54] Fagin, R., J. Halpern, and M. Vardi (1991). A model-theoretic analysis of knowledge. *Journal of the ACM 91*(2), 382 – 428.

[55] Fagin, R. and M. Vardi (1985). An internal semantics for modal logic: Preliminary report. In *Proc. 17th ACM SIGACT Symposium on Theory of Computing*, pp. 305 – 315.

[56] Falcone, R. and C. Castelfranchi (2001). Social trust: A cognitive approach. In *Trust and deception in virtual societies*, pp. 55–90. Springer.

[57] Falcone, R. and C. Castelfranchi (2004). Trust dynamics: How trust is influenced by direct experiences and by trust itself. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pp. 740–747. IEEE Computer Society.

[58] Fischer, M. J. and R. E. Ladner (1979). Propositional dynamic logic of regular programs. *Journal of computer and system sciences 18*(2), 194–211.

[59] Friedman, M. (1953). The methodology of positive economics. *Essays in positive economics 3*(3).

[60] Frigg, R. and S. Hartmann (2009). Models in science. *Stanford encyclopedia of philosophy*.

[61] Frigg, R. and J. Reiss (2009). The philosophy of simulation: hot new issues or same old stew? *Synthese 169*(3), 593–613.

[62] Fukuyama, F. (2006). *The end of history and the last man*. Simon and Schuster.

[63] Gaddafi, M. a. (1996). *The village, the village, the earth, the earth and the suicide of the astronaut*. GCI.

[64] Gibbard, A. and H. R. Varian (1978). Economic models.

[65] Giere, R. N. (2004). How models are used to represent reality. *Philosophy of science 71*(5), 742–752.

[66] Gigerenzer, G. and D. G. Goldstein (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review 103*(4), 650–669.

[67] Gilboa, I., A. Postlewaite, L. Samuelson, and D. Schmeidler (2014a). Economic models as analogies. *The Economic Journal 124*(578), F513–F533.

[68] Gilboa, I., A. Postlewaite, L. Samuelson, and D. Schmeidler (2014b). Economic models as analogies. *Cowles foundation discussion paper 1958*.

[69] Grüne-Yanoff, T. and P. Schweinzer (2008). The roles of stories in applying game theory. *Journal of Economic Methodology 15*(2), 131–146.

[70] Guiso, L., P. Sapienza, and L. Zingales (2008). Alfred marshall lecture social capital as good culture. *Journal of the European Economic Association 6*(2-3), 295–320.

[71] Halpern, J. and Y. Moses (1990). Knowledge and common knowledge in a distributed environment. *Journal of the ACM 37*(3), 549 – 587.

[72] Hansson, B. (1969). An analysis of some deontic logics. *Nous*, 373–398.

[73] Harsanyi, J. (1967). Games with incomplete informations played by 'bayesian' players. *Management Science 14*, 159–182, 320–334, 486–502.

[74] Hart, S., A. Heifetz, and D. Samet (1996). 'knowing whether', 'knowing that' and the cardinality of state spaces. *Journal of Economic Theory 70*(1), 249 – 256.

[75] Hartmann, S. (1996). The world as a process. In *Modelling and simulation in the social sciences from the philosophy of science point of view*, pp. 77–100. Springer.

[76] Hartmann, S. and J. Sprenger (2010). The weight of competence under a realistic loss function. *Logic Journal of IGPL 18*(2), 346–352.

[77] Hegselmann, R. and U. Krause (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation 5*(3).

[78] Henry, R. A. (1995). Improving group judgment accuracy: Information sharing and determining the best member. *Organizational Behavior and Human Decision Processes 62*(2), 190–197.

[79] Herzig, A. (2014). Logics of knowledge and action: critical analysis and challenges. *Autonomous Agents and Multi-Agent Systems*.

[80] Higman, G. (1952). Ordering by divisibility in abstract algebras. *Proceedings of the London Mathematics Society 2*, 326 – 336.

[81] Hill, G. W. (1982). Group versus individual performance: Are< em> n+ 1 heads better than one? *Psychological Bulletin 91*(3), 517.

[82] Hinsz, V. B. (1999). Group decision making with responses of a quantitative nature: The theory of social decision schemes for quantities. *Organizational Behavior and Human Decision Processes 80*(1), 28–49.

[83] Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance 21*(1), 40–46.

[84] Hooghe, M. and D. Stolle (2003). *Generating social capital: Civil society and institutions in comparative perspective*. Palgrave Macmillan.

[85] Hyyppä, M. T. (2010). *Healthy ties: Social capital, population health and survival*. Springer.

[86] James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 361–379.

[87] Jonker, C. M. and J. Treur (1999). Formal analysis of models for the dynamics of trust based on experiences. In *Multi-Agent System Engineering*, pp. 221–231. Springer.

[88] Kawachi, I., S. Subramanian, and D. Kim (2008). *Social capital and health*. Springer.

[89] Kets, W. (2014). Bounded reasoning and higher-order uncertainty. Available on the author's website.

[90] Klein, D. and J. Marx (2015). The dynamics of trust – emergence and destruction. *Proceedings of the* 17^{th} *International Workshop on Trust in Agent Societies*, forthcoming.

[91] Klein, D. and E. Pacuit (2014a). Changing types: Information dynamics for qualitative type spaces. *Studia Logica 102*(2), 297–319.

[92] Klein, D. and E. Pacuit (2014b). Focusing on campaigns. To appear.

[93] Klein, D. and J. Sprenger (2015). Modeling individual expertise in group judgments. *Economics and Philosophy*, forthcoming.

[94] Knack, S. (2002). Social capital, growth and poverty: A survey of cross-country evidence. *The role of social capital in development: An empirical assessment*, 42–82.

[95] Knack, S. and P. Keefer (1997). Does social capital have an economic payoff? a cross-country investigation. *The Quarterly journal of economics*, 1251–1288.

[96] Knill, D. C. and J. A. Saunders (2007). Bayesian models of sensory cue integration. *Bayesian brain: Probabilistic approaches to neural coding*, 189–206.

[97] Kornai, J., B. Rothstein, and S. Rose-Ackerman (2004). *Creating social trust in post-socialist transition*. Palgrave Macmillan.

[98] Korte, B. and J. Vygen (2002). *Combinatorial optimization*. Springer.

[99] Kummerfeld, E. and K. Zollman (2015). Conservatism and the scientific state of nature. *British Journal for Philosophy of Science*, forthcoming.

[100] Larrick, R. P., K. A. Burson, and J. B. Soll (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes 102*(1), 76–94.

[101] Laughlin, P. R. and A. L. Ellis (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology 22*(3), 177–189.

[102] Laver, M. and E. Sergenti (2011). *Party Competition: An Agent-Based Model*. Princeton University Press.

[103] Lehrer, K. and C. Wagner (1981). *Rational Consensus in Science and Society: A Philosophical and Mathematical Study*, Volume 21. Springer.

[104] Leitgeb, H. (2013a). Reducing belief simpliciter to degrees of belief. *Annals of Pure and Applied Logic 164*(12), 1338–1389.

[105] Leitgeb, H. (2013b). Scientific philosophy, mathematical philosophy, and all that. *Metaphilosophy 44*(3), 267–275.

[106] Lewis, D. (1969). *Convention*. Harvard University Press.

[107] Lewis, D. (1981). Probabilities of conditionals and conditional probabilities. In *Ifs*, pp. 129–147. Springer.

[108] Libby, R., K. T. Trotman, and I. Zimmer (1987). Member variation, recognition of expertise, and group performance. *Journal of Applied Psychology 72*(1), 81–87.

[109] Lin, H. and K. T. Kelly (2012). Propositional reasoning that tracks probabilistic reasoning. *Journal of philosophical logic 41*(6), 957–981.

[110] Lin, N. and B. H. Erickson (2008). *Social capital: an international research program*. oxford university press Oxford.

[111] Lindley, D. (1983). Reconciliation of probability distributions. *Operations Research 31*(5), 866–880.

[112] List, C. (2012). The theory of judgment aggregation: An introductory review. *Synthese 187*(1), 179–207.

[113] Littlepage, G. E., G. W. Schmidt, E. W. Whisler, and A. G. Frost (1995). An input-process-output analysis of influence and performance in problem-solving groups. *Journal of Personality and Social Psychology 69*(5), 877–889.

[114] Martini, C., J. Sprenger, and M. Colyvan (2013). Resolving disagreement through mutual respect. *Erkenntnis 78*(4), 881–898.

[115] Maynard-Reid II, P. and Y. Shoham (1998). From belief revision to belief fusion. In *Proceedings of LOFT-98*.

[116] McMullin, E. (1985). Galilean idealization. *Studies in History and Philosophy of Science Part A 16*(3), 247–273.

[117] Morgan, M. S. and M. Morrison (1999). *Models as mediators: Perspectives on natural and social science*, Volume 52. Cambridge University Press.

[118] Moses, Y., D. Dolev, and J. Y. Halpern (1986). Cheating husbands and other stories: a case study of knowledge, action, and communication. *Distributed computing 1*(3), 167–176.

[119] Muldoon, R., C. Lisciandra, C. Bicchieri, S. Hartmann, and J. Sprenger (2013). On the emergence of descriptive norms. *Politics, Philosophy & Economics*, 1470594X12447791.

[120] Myerson, R. (2004). Harsanyi's games with incomplete information. *Management Science 50*(12), 1818–1824.

[121] Nadeau, R., E. Cloutier, and J.-H. Guay (1993). New evidence about the existence of a bandwagon effect in the opinion formation process. *International Political Science Review 14*(2), 203–213.

[122] Nooteboom, B., T. Klos, and R. Jorna (2001). Adaptive trust and cooperation: An agent-based simulation approach. In *Trust in Cyber-societies*, pp. 83–109. Springer.

[123] Nozick, R. (1994). *The nature of rationality*. Princeton University Press.

[124] Pacuit, E. (2006). ESSLLI course on neighborhood semantics for modal logic. Course notes found at `ai.stanford.edu/~epacuit/nbhd_esslli.html`.

[125] Pacuit, E. and O. Roy (to appear). *Interactive Rationality*.

[126] Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press.

[127] Parikh, R. (2002). Social software. *Synthese 132*, 187 – 211.

[128] Parikh, R. (2003). Levels of knowledge, games, and group action. *Research in Economics 57*(3), 267 – 281. Logic and the Foundations of the Theory of Games and Decisions.

[129] Parikh, R. and P. Krasucki (1992). Levels of knowledge in distributed systems. *Sadhana 17*, 167–191.

[130] Pauly, M. (2008). On the role of language in social choice theory. *Synthese 163*(2), 227–243.

[131] Perea, A. (2012). *Epistemic game theory: reasoning and choice*. Cambridge University Press.

[132] Portes, A. (2000). Social capital: Its origins and applications in modern sociology. *LESSER, Eric L. Knowledge and Social Capital. Boston: Butterworth-Heinemann*, 43–67.

[133] Putnam, R. D. (1988). Diplomacy and domestic politics: the logic of two-level games. *International organization 42*(03), 427–460.

[134] Putnam, R. D. (1995). Tuning in, tuning out: The strange disappearance of social capital in america. *PS: Political science & politics 28*(04), 664–683.

[135] Putnam, R. D. (2000). *Bowling alone: The collapse and revival of American community*. Simon and Schuster.

[136] Putnam, R. D., R. Leonardi, and R. Y. Nanetti (1994). *Making democracy work: Civic traditions in modern Italy*. Princeton university press.

[137] Riker, W. H. (1962). *The theory of political coalitions*, Volume 578. Yale University Press New Haven.

[138] Satterthwaite, M. A. (1975). Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory 10*(2), 187–217.

[139] Schelling, T. C. (1971). Dynamic models of segregation. *Journal of mathematical sociology 1*(2), 143–186.

[140] Schneier, B. (2007). *Applied cryptography: protocols, algorithms, and source code in C*. john wiley & sons.

[141] Schumpeter, J. A. (2013). *Capitalism, socialism and democracy*. Routledge.

[142] Selvin, S. (1975). Problem in probability. *American Statistician 29*(1), 67–67.

[143] Sen, A. (1997). Maximization and the act of choice. *Econometrica: Journal of the Econometric Society 65*, 745–779.

[144] Sietsma, F. and J. van Eijck. Action emulation between canonical models. *preprint*.

[145] Siniscalchi, M. (2008). Epistemic game theory: Beliefs and types. In S. Durlauf and L. Blume (Eds.), *The New Palgrave Dictionary of Economics*. Basingstoke: Palgrave Macmillan.

[146] Soll, J. B. and R. P. Larrick (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition 35*(3), 780–805.

[147] Stokhof, M. and M. Van Lambalgen (2011). Abstractions and idealisations: The construction of modern linguistics. *Theoretical linguistics 37*(1-2), 1–26.

[148] Strevens, M. (2004). The causal and unification approaches to explanation unified causally. *Noûs 38*(1), 154–176.

[149] Sugden, R. (2000). Credible worlds: the status of theoretical models in economics. *Journal of Economic Methodology 7*(1), 1–31.

[150] Surowiecki, J. (2005). *The wisdom of crowds*. Random House LLC.

[151] Sztompka, P. (1999). *Trust: A sociological theory*. Cambridge University Press.

[152] Torche, F. and E. Valenzuela (2011). Trust and reciprocity: A theoretical distinction of the sources of social capital. *European Journal of Social Theory 14*(2), 181–198.

[153] Uslaner, E. M. (2002). *The moral foundations of trust*. Cambridge University Press.

[154] van Benthem, J. (2011). *Logical Dynamics of Information and Interaction*. Cambridge University Press.

[155] van Benthem, J., J. Gerbrandy, and B. Kooi (2009). Dynamic update with probabilities. *Studia Logica: An International Journal for Symbolic Logic 93*(1), 67 – 96.

[156] van Benthem, J., E. Pacuit, and O. Roy (2011). Toward a theory of play: A logical perspective on games and interaction. *Games 2*(1), 52 – 86.

[157] van der Hoek, W. and M. Pauly (2006). Modal logic for games and information. In P. Blackburn, J. van Benthem, and F. Wolter (Eds.), *Handbook of Modal Logic*, Volume 3 of *Studies in Logic*, pp. 1077 – 1148. Elsevier.

[158] van Ditmarsch, H. and T. French (2009). Simulation and information: Quantifying over epistemic events. In J.-J. Meyer and J. Broersen (Eds.), *Knowledge Representation for Agents and Multi-Agent Systems*, Volume 5605 of *Lecture Notes in Computer Science*, pp. 51–65. Springer Berlin / Heidelberg.

[159] van Ditmarsch, H., J. Ruan, and R. Verbrugge (2008). Sum and product in dynamic epistemic logic. *Journal of Logic and Computation 18*, 563–588.

[160] van Ditmarsch, H., W. van der Hoek, and B. Kooi (2007). *Dynamic Epistemic Logic*. Synthese Library. Springer.

[161] van Eijck, J., Y. Wang, and F. Sietsma (2010). Composing models. In *Proceedings of LOFT 2010*.

[162] Weisberg, M. (2007a). Three kinds of idealization. *The Journal of Philosophy*, 639–659.

[163] Weisberg, M. (2007b). Who is a modeler? *The British journal for the philosophy of science 58*(2), 207–233.

[164] Weisberg, M. (2012). *Simulation and similarity: using models to understand the world*. Oxford University Press.

[165] Wilf, H. S. (1985). Some examples of combinatorial averaging. *American Mathematical Monthly*, 250–261.

[166] Zollman, K. J., C. T. Bergstrom, and S. M. Huttegger (2013). Between cheap and costly signals: the evolution of partially honest communication. *Proceedings of the Royal Society B: Biological Sciences 280*(1750).