Modelling Democratic Deliberation

MSc Thesis *(Afstudeerscriptie)*

written by

Roosmarijn Goldbach
(born 12th of May 1988 in Breda, the Netherlands)

under the supervision of **Dr. Alexandru Baltag** and **Dr. Ulle Endriss**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
11th of March, 2015	Dr. Alexandru Baltag
	Prof. Dr. Johan van Benthem
	Prof. Dr. Jan van Eijck
	Dr. Ulle Endriss
	Prof. Dr. Christian List
	Prof. Dr. Fenrong Liu
	Dr. Soroush Rafiee Rad
	Dr. Jakub Szymanik

Institute for Logic, Language and Computation

Abstract

Deliberative democracy is a political theory that places deliberation at the heart of political decision making. In a deliberation, people justify their preferences to one another. They are confronted with new information and new perspectives, which might lead them to change their preferences. Therefore, deliberative democracy, unlike social choice theory, takes preferences to be secondary (derived) and dynamic rather than primary and static.

The first goal of this thesis is to formally model deliberation as aspired by deliberative democracy, henceforward referred to as democratic deliberation. This is done in two steps. Firstly, this thesis develops models for preference formation, since democratic deliberation is about justifying one's preferences. These models combine multi-agent plausibility models from dynamic epistemic logic with Dietrich and List's setting about reasons and rational choice. Combining these allows us to define the agents' preferences in terms of (i) their knowledge and belief, (ii) their motivational state or perspective and (iii) the properties that hold of the alternatives. Secondly, we introduce a model transformer for the preference formation models that models deliberation as a process in which all agents share all their information and all their perspectives. Together, the preference formation models and the model transformer for deliberation make up our formal framework. This framework is able to model two claims that are often made in the literature on deliberative democracy, namely that deliberation might lead to preference change and to a better understanding among the agents.

The second goal of this thesis is to use this formal framework to investigate the philosophical claim that deliberation provides an escape from social choice theory's impossibility results. The main result proved in this thesis is that in cases where the issue at stake is one-dimensional, deliberation is useful because it ensures single-peaked preferences via meta-agreement, and hence helps to circumvent Arrow's impossibility result.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors Alexandru Baltag and Ulle Endriss. Alexandru, thank you first of all for the inspiring course on modal logic. Your enthusiasm for the subject made me, and undoubtedly many of my fellow students, fall in love with logic. Furthermore, I am very grateful for your help on the thesis project, which has greatly benefited from your sharp comments and your many interesting insights. Ulle, thank you for continuous support throughout this project. I am happy that you were there for me even when it became clear that this thesis took a direction very different from our initial plan. During every phase of this project, you made me explain the intuition behind all the formalities which helped me to structure my ideas and to keep the bigger picture in mind.

Furthermore, I feel indebted to Christian List whose interesting work got me enthusiastic about the thesis topic in the first place. Christian, although we have only spoken for ten minutes, I am very thankful to you. In those ten minutes, you were willing to share your ideas on deliberative democracy with a complete stranger (me) and you gave me tons of ideas for my thesis. Therefore, I am honoured that you are part of my thesis committee. Moreover, my gratitude goes to the other members of the commitee Johan van Benthem, Jan van Eijck, Fenrong Liu, Jakub Szymanik and Soroush Rafiee Rad for taking the time to read this thesis.

Last but definitely not least, I would like to thank my friends and family for all their love and support before, during and hopefully long after this thesis project. Sylvia and Lara, thank you for all the coffee breaks, which always made me forget my thesis at some point or another. Irene and Maud, thank you for your support and encouragment. Special thanks go to Corinne, for all the long conversations during the harder times of this thesis and for understanding the things I struggle with so well. Lots of gratitude goes to my parents, Lia and Ferry, for their everlasting support and for giving me the opportunity to study so many things of the things that I am interested in. Finally, I would like to thank Thijs for his confidence in me, but mostly for just being there.

Contents

Chapter 1

Introduction

One of the most important challenges for modern democracies is how to cope with pluralism. As Berlin (1969: 169) argued, "if the ends of men ... are many and and not all of them are in principle compatible with each other, then the possibility of conflict ... can never wholly be eliminated from human life, either personal or social." The main challenge for democratic societies and, consequently, for democratic theories is to provide a way to handle conflicting interests.

Social choice theory is concerned with aggregating the manifold preferences of individuals in a democratic way, whereas deliberative democracy places the emphasis on the power of discourse. In deliberative processes, people communicate with one another which leads them to critically asses their own preferences, to better understand one another and to find common ground. Formal methods form an essential tool in the area of social choice theory, but they are hardly used in the literature on deliberative democracy. As working with formal methods forces one to structurally tackle a problem and helps make hidden assumptions explicit, it would be good to approach deliberative democracy from a more formal point of view. This thesis aims at doing so by developing a formal model for deliberation in line with the theory of deliberative democracy. Once this model has been developed, we want to use it to study the often heard claim that deliberative democracy can help circumvent social choice theory's impossibility results.

In order to formulate the goal of this thesis more precisely, some background is required. Section 1.1 contains a short introduction to social choice theory and its main results. In Section 1.2, the philosophical theory of deliberative democracy is introduced. Section 1.3 compares these two divergent ways of approaching pluralism in politics and discusses how they can be reconciled. Section 1.4 formulates the goal of this thesis in more detail and Section 1.5 describes its organisation.

1.1 Social Choice Theory

Dealing with pluralism is one of the most important challenges for modern democracy. As Dryzek & List (2003: 2) argue, "to social choice theorists, the democratic problem involves aggregation of views, interests or preferences across individuals". The question is how to aggregate these in a way that does justice to the will of the people, i.e. in a democratic way.

One of the pioneers of social choice theory, Nicolas de Condorcet (1785), extensively discussed a paradox that is still very relevant today. Suppose three individuals have to provide a linear order over a set $\mathcal{X} = \{a, b, c\}$ of three alternatives representing their preferences and suppose they do this in the following manner:

- Individual 1: $a \succ b \succ c$
- Individual 2: $b \succ c \succ a$
- Individual 3: $c \succ a \succ b$

Suppose that we want to determine the preferences of the group. One way to do this is the majority rule. Notice that two individuals prefer a to b, two prefer b to c and two prefer c to a. Thus, the collective preferences will be cyclical: $a \succ b \succ c \succ a$. In other words, whatever the collective choice is, there will always be another option which is preferred by a majority of the individuals.
This insight of Condorcet was generalised by Kenneth Arrow, another main pioneer of social choice theory. According to Arrow (1963), any aggregation method should at least satisfy the following democratic and seemingly intuitive constraints:[1]

- If all individuals prefer alternative a to b, then so does society.
- The social ranking of two alternatives a and b depends solely on the way the individuals rank these two alternatives.
- There does not exist an individual such that the collective ordering always equals the preferences of that individual. In other words, there is no dictator.

He was able to prove, however, that there is no aggregation method satisfying these constraints that always yields an acyclic social ordering. This surprising and influential result is known in social choice theory as Arrow's impossibility theorem.[2]

Another important impossibility result in the area of social choice theory is due to Gibbard (1973) and Satterthwaite (1975), who were concerned with strategy-proof aggregation methods. A method is strategy-proof if there never exists an individual that is tempted to misrepresent his preferences in order to get a more favourable collective outcome. Gibbard (1973) and Satterthwaite (1975) proved that if there are more than two alternatives, there does not exist a non-dictatorial aggregation method that is strategy-proof.[3]

A conclusion that can be drawn from these impossibility results is that no matter what aggregation method we use, democratic principles will be violated. Therefore, these theorems have important consequences not only for social choice theorists – whose task it is to come up with ways to escape these impossibility results – but also for political philosophers. Any theory of democracy has to respond to these uncomfortable results that demonstrate that aggregating the views, preferences or interests of a group of individuals in a democratic way is highly problematic.

1.2 Deliberative Democracy

Giving a proper definition of deliberative democracy is hardly possible, as the essence of this democratic theory is described differently by different proponents. As Elster (1998: 8) puts it:[4]

> There is a robust core of phenomena that count as deliberative democracy... All agree, I think, that the notion includes collective decision making with the participation of all who will be affected by the decision or their representatives: this is the democratic part. Also, all agree that it includes decision making by means of arguments offered *by* and *to* participants who are committed to the values of rationality and impartiality: this is the deliberative part.

In other words, deliberative democracy is an umbrella term for many political theories. Essential to all of them, however, is that they attribute a central role to deliberation in the process of political decision making.

[1] For a formal statement of these conditions, see Section 6.1.
[2] For a formal statement of Arrow's theorem, see Section 6.1. For the proof, see Arrow (1963) or Sen (1986).
[3] For a formal statement of the theorem and a proof, see Gibbard (1973) and Satterthwaite (1975).
[4] Italics in the original.

Since deliberation plays a central role in deliberative democratic theories, we will consider it in more detail. We start by turning to the goal of deliberation and, afterwards, consider the effects deliberation is believed to have by its adherents. Before doing so, however, it is important to note that representative forms of democracy are desirable from the perspective of deliberative democracy, because the number of people who can at the same time have a conversation together is limited. Thus, political debates that occur in democratic institutions form the standard, but definitely not the only case, of democratic deliberation.

The idea of deliberation is that people justify their preferences and views to one another by giving reasons that are "mutually acceptable and generally accessible" (Gutmann & Thompson, 2004: 7).[5] Thus, the participants of a deliberative process have to justify the following two things:

1. Their preferences.
2. Their views or opinions.

Firstly, the deliberators should justify their preferences. Thus, they have to argue why they prefer some alternative or policy over another. These reasons used for this should be mutually acceptable. That is, the other agents should, in principle, be able to accept the offered arguments. More specifically, although agents might disagree on the relative importance of the reasons offered, they should be able to acknowledge their intrinsic worth. Only reasons that meet this requirement are truly mutually acceptable. Suppose, for instance, that there is parliamentary debate about whether ritual animal slaughtering should be allowed or not. A representative of the green party might argue that we should ban religious slaughtering in the name of animal welfare. Even though someone might believe that in this particular case freedom of religion is more important than animal welfare, he cannot deny that animal welfare is worth pursuing. Hence, this reason is mutually acceptable. This example also shows that in justifying their preferences, agents usually turn to moral values or fundamental properties. Secondly, deliberators should defend their views and/or beliefs with reasons that are generally accessible. A reason is generally accessible if it can be tested with generally accepted methods of inquiry or, in cases where empirical evidence or logical inference are not appropriate, if it is not extremely implausible (Gutmann & Thompson, 2004: 72-73).

The fact that, in a democratic deliberation, people justify their views and preferences to one another has at least two important consequences that are essential for the theory of deliberative democracy. First of all, deliberation might lead to changes in opinion and preference, because the participants come into contact with new information, new perspectives and are critically questioned on their own views.[6] As a consequence, deliberative democrats consider the preferences of the agents as well as their beliefs to be dynamic in nature. Furthermore, deliberation leads to a better understanding among the participants, since it requires the justification of preferences and opinions with mutually acceptable and generally accessible reasons.[7] After deliberation, the agents know why the others prefer the options they prefer and why they believe what they believe.

In conclusion, deliberative democracy is a term used for political theories which attribute a central role to deliberation in the process of political decision making. In a deliberation, the participants have to justify their preferences and views to one another. By doing so, they learn, are confronted with new perspectives and their opinions are subject to critical assessment. Therefore, deliberation might induce preference change and leads to a better understanding among the participants.

[5] For similar ideas, see Cohen (1989), Habermas (1996) and Elster (1998).

[6] The idea that deliberation might lead to preference change can be found in any book or article on deliberative democracy. Some deliberative democrats, such as Elster (1986), Cohen (1989), Habermas (1996) and Bohman (1998), argue that deliberation tends to produce unanimous preferences. Others, such as Gutmann and Thompson (1996, 2004) and Dryzek (2000), argue that this is unrealistic, but they do not dispute the fact that deliberation might induce preference change. We will come back to this in Section 6.1.2.

[7] See, for instance, Gutmann & Thompson (1996, 2002) and Miller (1992).

1.3 The Relation between SCT and DD

This section focuses on the relationship between social choice theory and deliberative democracy. It starts by considering the main difference between the two areas and then looks at ways to reconcile them.

Both social choice theory and deliberative democracy consider ways to deal with the pluralism endemic to modern democracies. However, the two approaches diverge in the way they regard preferences. As Gutmann and Thompson (2002: 13) argue, social choice theory "takes preferences as given. It requires no justification for the preferences themselves, but seeks only to combine them in various ways that are efficient and fair". In other words, preferences are primary and static. The preferences are primary because they are exogenously given, which means that they are treated as inexplicable characteristics of the agent. The preferences are static, since they are treated as fixed and unchanging. Deliberative democracy, on the other hand, as Van Mill (1996: 746) argues, "focuses on the creation and alteration of preferences". In other words, preferences are secondary and dynamic. Preferences are secondary, because they require justification. People have to give reasons for them and these reasons form the fundamental meta-level from which preferences stem. Preferences are dynamic, because they might change during deliberation. Thus, the main difference between social choice theory and deliberative democracy is that the former takes preferences as primary and static, whereas the latter considers them to be secondary and dynamic.

As social choice theory is concerned with preference aggregation and deliberative democracy with preference formation and alteration, there has long been a gap between the two. These opposing camps hardly interacted with one another. This has changed with the works of Miller (1992), Van Mill (1996) and Dryzek and List (2003), who all sought to bring them together.

Miller (1992) argued that social choice theory and deliberative democracy are much less different than often assumed. He pointed out that they have the same starting point. Both search for ways to deal with pluralism and both want to do this in a manner that respects important democratic ideals. Social choice theory as well as deliberative democracy work from the core idea that all men are created equal. In social choice theory, this is reflected by the one man one vote principle, which meets the following three important requirements:[8]

1. Everyone has the possibility to express his or her preferences.
2. Everyone's vote has an equal weight.
3. Everyone's interest is taken into account.

Deliberative democracy requires the same things. Firstly, everyone should have the possibility to express his views, albeit often via a representative, in a deliberation. Secondly, people justify their views and preferences in a deliberative process. Deliberative democracy asks that these justifications be judged on their merits and not on the basis of existing power structures. Thirdly, in a deliberation "preferences are transformed to take account of the view of others" (Miller, 1992: 55).

Furthermore, social choice theory and deliberative democracy are not so far apart because almost all deliberative democrats agree that "deliberation concludes with voting" (Cohen, 1989: 23). This view is shared not only by deliberative democrats that do not believe that deliberation tends to lead to unanimity but also by the ones who do, as they acknowledge that this ideal might not always be achieved in practice.[9] Thus, although deliberation might induce preference change, the preferences of the agents after deliberation might still differ. In order to make a decision, we therefore need to make

[8] These three requirements are not completely independent of one another. In a properly functioning democracy, the second criterion presupposes the first and implies the third. This is, however, not always the case. Consider, for example, the elections in North Korea. In these elections, everyone's vote has an equal weight. However, due to the dictatorial regime, people cannot express their true preferences and it is definitely not the case that everyone's interest is taken into account.

[9] See, for example, Cohen (1989) and Habermas (1994).

use of aggregation mechanisms and, consequently, social choice theory's impossibility results become relevant again.

Starting from the fact that deliberation ends with voting, Dryzek and List (2003) have proposed an elegant way to reconcile social choice theory and deliberative democracy. The former has shown that preference aggregation is, in the words of Dryzek and List (2003: 2), "bedevilled by impossibility, instability and arbitrariness". However, instead of drawing pessimistic conclusions from these results, they believe that they show what deliberation minimally has to achieve for a democracy to function properly. In other words, "social choice theory shows exactly what deliberation must accomplish in order to render collective decision making tractable and meaningful, suggesting that democracy must in the end have a deliberative aspect" (Dryzek & List, 2003: 28). To make this claim more concrete, we use the Gibbard-Satterthwaite theorem as an example. This theorem says that if we have three or more alternatives, there does not exist a non-dictatorial and strategy-proof aggregation function. Dryzek and List (2003) hold that if deliberation can ensure that individuals always reveal their true preferences, then the condition of strategy-proofness can be weakened and we can circumvent the impossibility result. They argue that deliberation can indeed achieve this, as experimental results show that deliberation induces the agents' disposition to co-operate. Moreover, revealing false preferences may be costly in a deliberative democracy as the politicians will lose their credibility. Thus, what Dryzek and List (2003) argue is that deliberation can help to circumvent social choice theory's impossibility results and, hence, opens up the possibility of a meaningful democracy which combines deliberation with voting.

In conclusion, the main difference between social choice theory and deliberative democracy lies in the way they handle preferences. Social choice theory is concerned with preference aggregation and takes preferences to be primary and static. Deliberative democracy is concerned with preference formation and alteration and believes that preferences are secondary and dynamic. Because the focus of these two fields differ, they have ignored each other for a long time. More recently, however, Miller (1992) has shown that their starting point is the same and Dryzek and List (2003) reconciled them by proposing that deliberation before voting provides an escape from the impossibility results of social choice theory.

1.4 Goal of the Thesis

The goal of this thesis is twofold. Firstly and most importantly, this thesis aims at developing a formal framework for democratic deliberation. By a democratic deliberation, we mean a political deliberation that is ideal from the perspective of deliberative democracy. Secondly, once this framework has been developed, we want to use it to formally investigate the philosophical claim that deliberation can help circumvent social choice theory's impossibility results. This two-part goal will now be explained in more detail.

The main goal of this thesis is to develop a formal framework for democratic deliberation, in which people justify their preferences to one another and are confronted with new information and new perspectives. This goal consists of two subgoals. As deliberative democrats hold that preferences are derived, the first subgoal is to develop a formal model of preference formation that takes into account (i) the knowledge and belief of the agent in question, (ii) his motivational state or perspective and (iii) the properties that hold of the alternatives or the reasons that can be used to justify one's preference for a certain policy. The second subgoal is to model the fact that in, a deliberation, people are confronted with new information and with new perspectives, which might lead them to change their preferences.

The next goal of this thesis is to use the formal framework for democratic deliberation to investigate the claim that deliberation might help with circumventing the impossibility results of social choice theory. In particular, we will formally investigate the claim that deliberation provides a so-

lution to Arrow's theorem by ensuring single-peaked preferences.[10] For a thorough understanding of this claim, more background is needed. For now, the intuitive idea will suffice. Recall that Arrow's theorem says that the aggregation of individual preferences into a collective decision is impossible if we want our aggregation method to satisfy certain apparent democratic principles. This result does not apply, however, if the individual preference orderings that serve as an input to the aggregation mechanism are structured enough. The intuitive meaning of the claim is that deliberation structures the preferences of the individuals in such a way that Arrow's impossibility result is avoided.

This thesis contributes to the philosophical literature on deliberative democracy as well as to the literature on preference formation. It contributes to the preference formation literature, because our framework for preference formation takes into account the agent's epistemic doxastic state, his motivational state and the properties that hold of the alternatives. Although there are models for preference formation that consider two of these things, there does not exist, to the best of our knowledge, a model that takes all three of them into account. Furthermore, this thesis contributes to the literature on deliberative democracy because it makes explicit use of formal methods, which are usually eschewed in this area. By using formal methods to study, among others, the philosophical claim that deliberation might provide a solution to Arrow's impossibility theorem, the hidden assumptions underlying this claim are made explicit and we can see more clearly how tenable this claim really is.

1.5 Organisation of the Thesis

This thesis is organised as follows. As agents share information in a deliberation and use their knowledge and beliefs in determining their preferences, multi-agent plausibility models, which are commonly used in epistemic doxastic logic, form the basis for our framework. Chapter 2 introduces these models and Chapter 3 introduces a model transformer for them, which models deliberation as a process in which all agents share all their information. Chapter 4 focuses on preferences and their justifications. This chapter combines the models from Chapter 2 with Dietrich and List's (2013a, 2013b) setting for preference formation in order to create preference formation models that take into account the knowledge and belief of the agents, their motivational state or perspective and the properties that hold of the alternatives or the reasons usable for justifying one's preference for a certain policy. In Chapter 5, the model transformer from Chapter 3 is modified in such a way that it can by applied to the models for preference formation. This model transformer models deliberation as a process in which agents share not only their information, but also their perspectives. Our formal framework for democratic deliberation consists of the preference formation models together with the model transformer for deliberation. With the formal framework in place, Chapter 6 formally investigates the claim that deliberation can induce single-peaked preferences and, thereby, the idea that democratic deliberation provides a solution to social choice theory's impossibility results. Lastly, Chapter 7 concludes with a discussion of this thesis and ideas for future work.

[10] Arrow's theorem, the definition of single-peakedness and this claim are formally introduced in Chapter 6. Readers with a background in social choice theory may recall that a set of individual preference orderings is single-peaked if there exists a dimension such that each individual has a most preferred position on that dimension (his peak), with decreasing preferences as the options are further removed from this peak.

Chapter 2

The Epistemic Doxastic Basis

The essence of deliberative democracy is that people justify their preferences and views to one another in a deliberation. According to Von Wright (1963), preferences come in two flavours. On the one hand, there are preferences that people hold for no particular reason. Examples are someone's preference for lemon over strawberry ice cream or for tea over coffee. Preferences of this kind are intrinsic. On the other hand, there are extrinsic preferences. In the case of extrinsic preferences, "a judgment of betterness serves as a ground or reason for preference" (Von Wright, 1963: 14). Democratic deliberations are typically about this latter kind. Otherwise there would be no point to deliberation in the first place. Extrinsic preferences are based on judgements. This entails that an agent's knowledge and beliefs play a role in the formation of his preferences. Because of this, the basis for the formal framework for democratic deliberation will be an epistemic doxastic logic.

This chapter introduces that logic and, in doing so, it takes a semantic approach. Section 2.1 introduces multi-agent plausibility frames and discusses some basic epistemic doxastic attitudes. Common prior frames, a special kind of multi-agent plausibility frames, are discussed in Section 2.2. Section 2.3 considers different concepts of group knowledge and belief. In Section 2.4, we formally introduce a language for multi-agent plausibility models and obtain our basic epistemic doxastic logic (EDL), of which we prove soundness and completeness.

2.1 Multi-Agent Plausibility Frames, Knowledge and Belief

This section introduces the basics of epistemic doxastic logic. In Section 2.1.1, multi-agent plausibility frames are formally defined. Section 2.1.2 considers the basic epistemic doxastic attitudes knowledge and belief and discusses their properties. Section 2.1.3 shows how the notion of belief can be interpreted in a standard frames, a special kind of multi-agent plausibility frames.

2.1.1 Multi-Agent Plausibility Frames

The appearance of his book *Knowledge and Belief: an introduction to the logic of the two notions* made Hintikka (1962) one of the founding fathers of epistemic doxastic logic. He proposed the use of standard Kripke semantics to model knowledge and belief. Here we take a somewhat different approach, based on Baltag and Smets (2006a, 2006b), who introduced plausibility frames:

Definition 2.1. Let $\mathcal{N}$ be a finite set of agents. A *multi-agent plausibility frame* is a tuple $F = (W, \sim_i, \leq_i)_{i \in \mathcal{N}}$, where W is a set of possible worlds, $\sim_i \subseteq W \times W$ is an equivalence relation and $\leq_i \subseteq W \times W$ is reflexive and transitive relation, satisfying the following constraint:

- For all $i \in \mathcal{N} : \leq_i \subseteq \sim_i$.

For all agents $i \in \mathcal{N}$, $\sim_i$ is the relation for epistemic indistinguishability and $\leq_i$ for plausibility. A proposition $P \subseteq W$ is a set of possible worlds. Given a proposition P, we define $\neg P := \overline{P} := W \setminus P$ as the complement of P. In this thesis, we use $\neg P$ and $\overline{P}$ interchangeably. Given two propositions P and Q, we define $P \wedge Q := P \cap Q$ as their intersection and $P \vee Q := P \cup Q$ as their union. Furthermore, we define the tautological proposition $\top := W$ and the inconsistent proposition $\bot := \emptyset$.

2.1.2 Basic Epistemic Doxastic Notions

A multi-agent plausibility frame consists of a set of possible worlds and two relations for each agent, one denoting epistemic indistinguishability and one denoting plausibility. These relations play a crucial role in defining epistemic doxastic attitudes. This section starts with a brief introduction of Kripke modalities. Afterwards, we discuss the most widely used notions of knowledge and belief.

In multi-agent plausibility frames, epistemic doxastic attitudes can be modelled as Kripke modalities. In fact, any binary relation $R \subseteq W \times W$ gives rise to a Kripke modality $[R]$:

Definition 2.2. Let F be a multi-agent plausibility frame and $R \subseteq W \times W$ a binary relation. R gives rise to a *Kripke modality* $[R] : \mathcal{P}(W) \to \mathcal{P}(W)$ defined by $[R]P := \{w \in W \mid \forall w' \in W(wRw' \Rightarrow w' \in P)\}$ for any $P \subseteq W$.

Thus, Kripke modalities in general and epistemic doxastic attitudes in particular can be thought of as operators that take each proposition to a specific set of worlds. Regardless of the constraints on the binary relation R, monotonicity and generalisation always hold:

Proposition 2.3. Let F be a multi-agent plausibility frame and $R \subseteq W \times W$ a binary relation. Let $P, Q \subseteq W$ be arbitrary. The following hold:

1. Monotonicity: $P \subseteq Q$ implies $[R]P \subseteq [R]Q$.
2. Generalisation: $[R]W = W$.

Proof. Left to the reader. □

Monotonicity and generalisation together are the semantic counterparts of the **K**-axiom. The **K**-axiom says that for any two syntactic propositions φ, ψ it is the case that $[R](\varphi \to \psi) \to ([R]\varphi \to [R]\psi)$. This holds no matter what conditions we put on R. When defining specific epistemic doxastic notions, the constraints on R do matter. In multi-agent plausibility frames, these are defined in terms of the agent-specific relations $\sim_i$ and $\leq_i$.

For defining knowledge, we use the equivalence relation $\sim_i$. This relation corresponds to epistemic indistinguishability and is the accessibility relation for hard or infallible knowledge. That is, if agent i cannot distinguish between worlds w and w' on the basis of his knowledge, i.e. if w and w' are epistemically indistinguishable for agent i, then the two worlds are related by $\sim_i$. This allows for the following definition:

Definition 2.4. Let F be a multi-agent plausibility frame and let $i \in \mathcal{N}$. We define *hard knowledge* as the Kripke modality for the epistemic indistinguishability relation: $K_i^F := [\sim_i]$. Whenever the frame is fixed or understood, we simply use the notation K_i.

Notice that using the definitions of a Kripke modality and of hard knowledge, we get $K_i P = \{w \in W \mid \forall w' \in W(w \sim_i w' \Rightarrow w' \in P)\}$ for all $P \subseteq W$. In other words, a certain proposition is infallible knowledge at world w for agent i iff it is true in all the worlds that are epistemically indistinguishable from w for i. This concept of knowledge corresponds to an absolute interpretation of knowledge.

According to Hintikka (1962: 20), by saying one knows something one "implicitly denies that any further information would have led him to alter his view". That is, if an agent infallibly knows a proposition P, this knowledge will never be defeated, even if the agent receives false information.

Notice that $\sim_i$ is an equivalence relation. Hence, it partitions the state space in equivalence classes. Worlds within the same equivalence class cannot be distinguished on the basis of the agents hard knowledge, whereas worlds that belong to a different equivalence class can. These equivalence classes can be thought of as information cells. The fact that $\sim_i$ is an equivalence relation gives us the following properties of hard knowledge:

Proposition 2.5. Let F be a multi-agent plausibility frame, $i \in \mathcal{N}$ and $P \subseteq W$. The following hold:

1. $K_i P \subseteq P$.
2. $K_i P = K_i K_i P$.
3. $\neg K_i P = K_i \neg K_i P$.

Proof. Left to the reader. □

As mentioned, the inclusions in Proposition 2.5 follow directly from the properties of the underlying relation $\sim_i$. The inclusion $K_i P \subseteq P$ follows from the reflexivity of $\sim_i$ and is the semantic version of the **T**-axiom, which captures the *factivity of knowledge*: If you know something, it is true. In other words, the agents can only know true things. The inclusion $K_i P \subseteq K_i K_i P$ follows from the transitivity of $\sim_i$ and is the semantic version of axiom **4**. Philosophically, this corresponds to *positive introspection*. That is, whenever you know something you know that you know it. The inclusion $\neg K_i P \subseteq K_i \neg K_i P$ follows from the Euclideanness of $\sim_i$ and is the semantic version of axiom **5**.[1] Axiom **5** corresponds to *negative introspection*: if you do not know something, you at least know that you do not know it. If both positive and negative introspection are satisfied, we speak of *full introspection*. This means that agents are fully aware of the extent of their knowledge. Given any proposition P, the agent knows whether he does or does not know it. Lastly, any Kripke modality satisfies the **K**-axiom. Thus, $K_i(P \to Q) \to (K_i P \to K_i Q)$. Philosophically, this captures the idea that agents know all the logical consequences of their knowledge. In other words, they are *logically omniscient*.

Logical omniscience and negative introspection are highly debated in the epistemological literature.[2] Logical omniscience is clearly debatable. Just think of a mathematical theorem that took you hours to prove. If only you had been logically omniscient then. There are also many counterexamples to negative introspection. Suppose, for instance, that the president has decided to resign but has not told this to anyone. Obviously, it is not the case that a random citizen knows that he does not know that the president is resigning. Despite these objections, the most widely used concept of knowledge in decision theory, economics and artificial intelligence is **S5**-knowledge, which validates **K, T, 4** and **5**. The popularity of this concept of knowledge is partly explained by its nice mathematical properties, as the underlying relations are equivalence relations. Since we want our framework to be in line with the standard account of knowledge in the above mentioned areas as much as possible, we work with the concept of infallible knowledge throughout this thesis.

In order to define belief, we need to introduce the Kripke modality for the plausibility relation $\leq_i$. Two worlds w and w' are related by $\leq_i$, if agent i considers world w' to be at least as plausible as world w. The Kripke modality for $\leq_i$ is called the plausibility modality:

[1] Recall that, given a set X and a relation $R \subseteq X \times X$, R is Euclidean iff for all $x, y, z \in X, xRy$ and xRz imply yRz. In addition, recall that a binary relation is reflexive, transitive and symmetric iff it is reflexive, transitive and Euclidean. Thus, every equivalence relation satisfies Euclideanness.

[2] For a good overview of the contemporary debates in epistemology, see Steup & Sosa (2005). For the problem of logical omniscience, see Stalnaker (1991, 1999), Fagin et al. (2003): Chapter 9 and Hendricks (2006): 80-114. For a discussion of negative introspection, among other things, see Lenzen (1979, 1980).

Definition 2.6. Let F be a multi-agent plausibility frame and let $i \in \mathcal{N}$. *The plausibility modality* is defined as the Kripke modality for the plausibility relation: $\Box_i^F = [\leq_i]$. Whenever the frame is fixed or understood, we simply use the notation $\Box_i$.

The plausibility modality can best be understood in terms of its dual $\Diamond_i P := \overline{\Box_i \overline{P}}$, which says that there exists a world just as plausible as the actual world in which proposition P holds. That is, proposition P is plausible.

The condition on multi-agent plausibility expressing that $w \leq_i w'$ implies $w \sim_i w'$ gives us some insight into the relation between plausibility and epistemic indistinguishability. It is best understood by looking at the contrapositive. Suppose an agent can distinguish two worlds w and w' on the basis of his hard knowledge, i.e. one of these worlds is impossible for him. Then it makes no sense to relate this world by plausibility, because the impossible world is implausible as well. In contrast to the plausibility models introduced by Baltag and Smets (2006a, 2006b), we do not require the plausibility relations to be connected. Therefore, it may be the case that the agent is unable to compare two worlds that are both epistemically possible. The fact that $\leq_i \subseteq \sim_i$ gives rise to the following observation:

Observation 2.7. Let F be a multi-agent plausibility frame, $i \in \mathcal{N}$ and $P \subseteq W$. Then $K_i P \subseteq \Box_i P$.

The plausibility modality has the following properties:

Proposition 2.8. Let F be a multi-agent plausibility frame, $i \in \mathcal{N}$ and $P \subseteq W$. The following hold:

1. $\Box_i P \subseteq P$.
2. $\Box_i P = \Box_i \Box_i P$.

Proof. Left to the reader. □

According to Propositions 2.3 and 2.8, the plausibility modality satisfies the axioms **K**, **T** and **4**. In other words, it satisfies all the **S4**-axioms. Therefore, if one insists, one can think of the plausibility modality as representing some form of soft, unconscious knowledge which is factive, positively introspective and closed under logical consequence. In fact, in the case where the plausibility relations are locally connected, the plausibility modality represents *defeasible knowledge*.[3] Defeasible knowledge is preserved under true information, i.e. an agent "knows that φ if and only if, for any ψ that is true, she would still believe that φ after learning ψ" (Stalnaker, 2006: 189). Thus, in contrast to infallible knowledge, defeasible knowledge does not have to be preserved in light of false evidence.

With knowledge and plausibility in place, it is possible to define belief:

Definition 2.9. Let F be a multi-agent plausibility frame, $i \in \mathcal{N}$ and $P \subseteq W$. We define *belief* as follows: $B_i^F P := K_i^F \Diamond_i^F \Box_i^F P$, where $\Diamond_i^F Q := \overline{\Box_i^F \overline{Q}}$ for any $Q \subseteq W$. Whenever the frame is fixed or understood, we simply use the notation B_i.

Observation 2.10. Let F be a multi-agent plausibility frame, $i \in \mathcal{N}$ and $P \subseteq W$. Then $w \in B_i P$ iff for all $w' \in W$ such that $w \sim_i w'$ there exists a $w'' \in W$ such that $w' \leq_i w''$ and for all $w''' \in W$ such that $w'' \leq_i w'''$ it holds that $w''' \in P$.

Thus, this definition says that an agent believes proposition P in the actual world w if from all the worlds in the agent's information cell containing w one can go via the plausibility relation to some point and from that point onwards P remains true. In English, an agent believes proposition P if P is true in all "plausible enough" worlds consistent with the agent's knowledge. This definition of belief might seem rather complicated at first sight. Section 2.1.3 shows, however, that this definition of belief simply amounts to truth in all the most plausible worlds in frames where the existence of

[3] Recall that a binary relation $R \subseteq X \times X$ is locally connected if for all $x, y \in X$ it holds that xRy or yRx. The concept of defeasible knowledge was mainly championed by Lehrer & Paxson (1969), Klein (1971) and Lehrer (1990). Versions of it were formalised by Rott (2004), Stalnaker (2006) and Baltag & Smets (2006b).

such worlds is guaranteed.

Definition 2.9 immediately gives us the following relations between knowledge and belief:

Proposition 2.11. Let F be a multi-agent plausibility frame, $i \in \mathcal{N}$ and $P \subseteq W$. The following hold:

1. $K_i P \subseteq B_i P$
2. $B_i P = K_i B_i P$
3. $\neg B_i P = K_i \neg B_i P$

Proof. Item 1 follows from Observation 2.10 and the fact that $\leq_i \subseteq \sim_i$. Items 2 and 3 follow immediately from Definition 2.9 and Proposition 2.5. □

Proposition 2.11 captures some basic intuitions about knowledge and belief. The first item says that knowledge implies belief. That is, if an agent knows P he believes it as well. The second and third item together say that agents are fully introspective with respect to their beliefs. Given any proposition, the agents know whether they do or do not believe it.

One can now easily show that belief has the following properties:

Proposition 2.12. Let F be a multi-agent plausibility frame, $i \in \mathcal{N}$ and $P \subseteq W$. The following are valid:

1. $B_i \bot = \emptyset$.
2. $B_i P = B_i B_i P$.
3. $\neg B_i P = B_i \neg B_i P$.

Proof. For item 1, suppose towards a contradiction that $B_i \bot \neq \emptyset$. Then there exists a $w \in W$ such that $w \in B_i \bot$. Since $\sim_i$ is reflexive, it follows from Observation 2.10 that there exists a $v \in W$ such that $w \leq_i v$ and for all v' such that $v \leq_i v'$ it holds that $v' \in \bot = \emptyset$. Since $\leq_i$ is reflexive, it follows that $v \in \emptyset$. Contradiction. Thus, $B_i \bot = \emptyset$. The $\subseteq$-inclusions of items 2 and 3 follow immediately from Proposition 2.11. The proofs of the $\supseteq$-inclusions are by contradiction and use the fact that $B_i \bot = \emptyset$. □

Notice that Proposition 2.12.2 and 2.12.3 are the doxastic counterparts of Proposition 2.5.2 and 2.5.3. Thus, agents are fully introspective with respect to their beliefs. Moreover, the **K**-axiom holds, i.e. agents believe all the logical consequences of their beliefs. The only difference between knowledge and belief is that belief is not factive. And indeed, it should not be. Agents may well believe things that are not actually the case. However, the beliefs of the agents do have to be consistent. This is what the identity $B_i \bot = \emptyset$, the semantic counterpart of the **D**-axiom, expresses. Since **K**, **D**, **4** and **5** are all satisfied, this notion of belief is called **KD45**-belief. Although this notion of belief is controversial for logical omniscience, among other things, we use it in this thesis because of its standardness in decision theory, economics and artificial intelligence.

Lastly, we show that in the case where the plausibility relations are locally connected, belief can by defined solely in terms of $\Diamond_i$ and $\Box_i$:

Theorem 2.13. Let F be a multi-agent plausibility frame such that for all $i \in \mathcal{N}$ the relation $\leq_i$ is locally connected, i.e. for all $w, w' \in W$, $w \sim_i w'$ implies $w \leq_i w'$ or $w' \leq_i w$. Let $i \in \mathcal{N}$ and $P \subseteq W$. Then $B_i P = \Diamond_i \Box_i P$.[4]

[4]The fact that B_i is equivalent to $\Diamond_i \Box_i$ in models where the plausibility relations are locally connected, has first been pointed out by Stalnaker (2006). These types of models are particularly interesting for modelling belief revision, because they validate all the AGM-axioms (Baltag and Smets, 2006b).

Proof. According to Definition 2.9, $B_iP = K_i\Diamond_i\Box_iP$. Therefore, the $\subseteq$-inclusion follows from the factivity of knowledge. For the $\supseteq$-inclusion, let $w \in \Diamond_i\Box_iP$. Let $w' \in W$ be such that $w \sim_i w'$. From $w \in \Diamond_i\Box_iP$, it follows that there exists a $v \in W$ such that $w \leq_i v$ and $v \in \Box_iP$. By the fact that $\leq_i \subseteq \sim_i$ and the Euclideanness of $\sim_i$, we get $w' \sim_i v$. By local connectedness, it follows that $w' \leq_i v$ or $v \leq_i w'$. If the former, we immediately get that $w' \in \Diamond_i\Box_iP$. If the latter, let $w'' \in W$ be such that $w' \leq_i w''$. By transitivity of $\leq_i$, it follows that $v \leq_i w''$ and hence that $w'' \in P$. Thus, $w' \in \Box_iP$. Due to reflexivity of $\leq_i$, it holds that $w' \in \Diamond_i\Box_iP$. In both cases, $w' \in \Diamond_i\Box_iP$ and, thus, $w \in K_i\Diamond_i\Box_iP = B_iP$. Thus, $\Diamond_i\Box_iP \subseteq B_iP$. □

2.1.3 Belief in Standard Multi-Agent Plausibility Frames

For readers who are not familiar with epistemic doxastic logic, the definition of belief introduced in the previous section might seem quite complicated. Therefore, this section introduces standard multi-agent plausibility frames and shows that, in these frames, belief has a more intuitive interpretation as truth in all the most plausible worlds.

Standard multi-agent plausibility frames are defined as follows:

Definition 2.14. Let $\mathcal{N}$ be a finite set of agents and $F = (W, \sim_i, \leq_i)_{i\in\mathcal{N}}$ a multi-agent plausibility frame. F is called a *standard multi-agent plausibility frame* if for all $i \in \mathcal{N} : \leq_i$ is converse well-founded.

Recall that a relation is converse well-founded if there does not exist an infinitely ascending chain. Therefore, the converse well-foundedness of the plausibility relations ensures that there always exists a most plausible world.

In standard multi-agent plausibility frames, belief is usually defined as truth in all the most plausible worlds. We will now formally define this alternative notion of belief and show that it is equivalent to notion introduced earlier. In order to do so, we first define the maximal elements of a relation:

Definition 2.15. Let F be a multi-agent plausibility frame and $R \subseteq W \times W$. The set of *R-maximal* elements is defined as $\mathrm{Max}(R) := \{w \in W \mid \forall w' \in W(wRw' \Rightarrow w'Rw)\}$.

Notice that if a relation is converse well-founded, it always has maximal elements:

Observation 2.16. Let F be a multi-agent plausibility frame. Let $R \subseteq W \times W$ be such that R is a converse well-founded relation. Then $\mathrm{Max}(R) \neq \emptyset$.

In standard multi-agent plausibility frames, the plausibility relations of all the agents are converse well-founded. Thus, each agent is endowed with a set of worlds that he considers to be most plausible. This allows for the following definition of belief:

Definition 2.17. Let F be a standard multi-agent plausibility frame. We define the *belief relation* $\rightarrow_i \subseteq W \times W$ to be such that for all $w, w' \in W : w \rightarrow_i w'$ iff $w \sim_i w'$ and $w' \in \mathrm{Max}(\leq_i)$.

Definition 2.18. Let F be a standard multi-agent plausibility frame and let $i \in \mathcal{N}$. We define *belief* as the Kripke modality for the belief relation: $\mathbb{B}_i^F = [\rightarrow_i]$. Whenever the frame is fixed or understood, we simply use the notation $\mathbb{B}_i$.

Thus, an agent believes proposition P iff it is true in all the worlds that he considers to be most plausible. The next theorem shows that this definition of belief is equivalent to the one from Section 2.1.2:

Theorem 2.19. Let F be a standard multi-agent plausibility frame, $i \in \mathcal{N}$ and $P \subseteq W$. Then $\mathbb{B}_iP = K_i\Diamond_i\Box_iP =: B_iP$, where $\Diamond_iQ := \overline{\Box_i\overline{Q}}$ for any $Q \subseteq W$.

Proof. Observe the following: for any $Q \subseteq W$, $\Diamond_i Q = \overline{\Box_i \overline{Q}} = \{w \in W \mid w \notin \Box_i \overline{Q}\} = \{w \in W \mid \exists w' \in W(w \leq_i w' \wedge w' \notin \overline{Q}\} = \{w \in W \mid \exists w' \in W(w \leq_i w' \wedge w' \in Q)\}$.

For the $\subseteq$-inclusion, let $w \in \mathbb{B}_i P$. That is, for all $v \in W$ such that $w \sim_i v$ and $v \in \mathrm{Max}(\leq_i)$, $v \in P$. Let $w' \in W$ be such that $w \sim_i w'$. Since $\leq_i$ is converse well-founded and serial, it follows that there exists a $w'' \in W$ such that $w' \leq_i w''$ and $w'' \in \mathrm{Max}(\leq_i)$.[5] Let $w''' \in W$ be such that $w'' \leq_i w'''$. Since w'' is maximal, w''' must be as well. Moreover, by the transitivity of $\sim_i$ and the fact that $\leq_i \subseteq \sim_i$, we can conclude from $w \sim_i w' \leq_i w'' \leq_i w'''$ that $w \sim_i w'''$. Thus, $w''' \in P$. Consequently, $w'' \in \Box_i P$. From the observation on $\Diamond_i$ and the fact that $w' \leq_i w''$, we get $w' \in \Diamond_i \Box_i P$ and, hence, $w \in K_i \Diamond_i \Box_i P$. Thus, $\mathbb{B}_i P \subseteq K_i \Diamond_i \Box_i P$

For the $\supseteq$-inclusion, let $w \in K_i \Diamond_i \Box_i P$. Let $w' \in W$ be such that $w \sim w'$ and $w' \in \mathrm{Max}(\leq_i)$. $w \sim_i w'$ implies $w' \in \Diamond_i \Box_i P$. Thus, there exists a w'' such that $w' \leq_i w''$ and $w'' \in \Box_i P$. Notice that from $w' \leq_i w''$ and $w' \in \mathrm{Max}(\leq_i)$, it follows that $w'' \leq_i w'$. Hence $w' \in P$ and $w \in \mathbb{B}_i P$. Thus, $K_i \Diamond_i \Box_i P \subseteq \mathbb{B}_i P$. □

2.2 Common Prior Frames

A standard assumption in economics and game theory is the *common prior assumption*. This assumption "says that differences in beliefs among agents can be completely explained by differences in information" (Halpern, 1998: 133). That is, at some (fictional) initial state all agents had the same beliefs. The differences in belief at the current state are solely due to the fact that different agents have experienced and learned different things throughout their lives and, thus, have different information about the state of the world.

In order to make our framework easily compatible with standard settings in game and decision theory, we define common prior frames. In common prior frames, the current plausibility relations of the agents can be obtained from their knowledge and a common prior plausibility relation, which is the plausibility relation before the agents got to learn any private information:

Definition 2.20. Let $\mathcal{N}$ be a finite set of agents. A multi-agent plausibility frame $F = (W, \sim_i, \leq_i)_{i \in \mathcal{N}}$ is a *common prior frame* iff there exists a reflexive and transitive binary relation $\leq \subseteq W \times W$ such that for all $i \in \mathcal{N} : \leq_i = \leq \cap \sim_i$. The relation $\leq$ is called the *common prior plausibility relation*.

2.3 Epistemic and Doxastic Group Notions

In multi-agent systems in general and deliberation in particular, even more interesting than the epistemic doxastic attitudes of individuals might be the knowledge and belief of the entire group. Different notions of group knowledge and belief have been studied in formal epistemology, such as common knowledge, distributed knowledge and common belief.[6] This section introduces distributed knowledge and distributed belief. As, in multi-agent plausibility frames, all the agents are endowed with an epistemic indistinguishability and a plausibility relation, these group notions can be defined in terms of the relations of the individual agents.

Distributed knowledge is typically characterised as the knowledge a group of agents would have if they were to combine all their knowledge. Notice that distributed knowledge of φ does not mean that at least one of the agents knows φ. For example, suppose Alice knows φ and Bob knows $\varphi \rightarrow \psi$. If they were to combine their knowledge, they would know ψ although neither Alice nor Bob knows it at present (Fagin et al. 2003). On the one hand, we can see distributed knowledge as a form of

[5] Recall that, given a set X and a binary relation $R \subseteq X \times X$, a relation is serial if for all $x \in X$ there exists a $y \in X$ such that xRy.

[6] The concept of common knowledge was introduced in formal epistemology by Lewis (1969) and Aumann (1976). After their pioneering work, many other concepts of group knowledge and belief have been studied. For a good overview, see Fagin et al. (1995), Chapters 5 and 11.

group knowledge, because it is the knowledge that is latently present within the group. On the other hand, distributed knowledge can be seen as a form of potential knowledge, since it is the knowledge the agents would have *if* they were to share all their hard information.

In a multi-agent plausibility frame, each agent is endowed with an epistemic indistinguishability relation which represents the agent's knowledge. If two worlds cannot be distinguished on the basis of agent i's hard knowledge, they are related by $\sim_i$. Intuitively, the group of agents cannot distinguish between two worlds if none of the agents can. Therefore, the epistemic indistinguishability relation for the group of agents $\sim_{\mathcal{N}}$ can be defined as $\sim_{\mathcal{N}} := \bigcap_{i\in\mathcal{N}} \sim_i$. Notice that $\sim_{\mathcal{N}}$ is an equivalence relation, as it is the intersection of equivalence relations. We now obtain the following definition for distributed knowledge:

Definition 2.21. Let F be a multi-agent plausibility frame. *Distributed knowledge* is defined as the Kripke modality for the intersection of the epistemic indistinguishability relations of the individuals: $D_K^F = [\sim_{\mathcal{N}}]$, where $\sim_{\mathcal{N}} := \bigcap_{i\in\mathcal{N}} \sim_i$. Whenever the frame is fixed or understood, we simply use the notation D_K.

As the accessibility relation for D_K is the intersection of the individual indistinguishability relations, it can really be thought of as distributed knowledge. Recall that a multi-agent plausibility frame consists of a set of worlds, and that each agent is endowed with a partition of these worlds in information cells. A proposition is hard information for agent i at world w, if it is true throughout the information cell of i containing w. Similarly, it is the case that a certain proposition is distributed knowledge at world w if it is true in the intersection of the information cells of all the agents that contain w.

In multi-agent plausibility frames, it makes sense to not only consider the intersection of all the epistemic indistinguishability relations, but also the intersection of the plausibility relations. Therefore, we define $\leq_{\mathcal{N}} := \bigcap_{i\in\mathcal{N}} \leq_i$. Notice that $\leq_{\mathcal{N}}$ is reflexive and transitive, because it is the intersection of reflexive and transitive relations. The relation $\leq_{\mathcal{N}}$ can be used to define distributed plausibility:

Definition 2.22. Let F be a multi-agent plausibility frame. *Distributed plausibility* is the Kripke modality for the intersection of the plausibility relations of the individuals: $D_\Box^F = [\leq_{\mathcal{N}}]$, where $\leq_{\mathcal{N}} := \bigcap_{i\in\mathcal{N}} \leq_i$. Whenever the frame is fixed or understood, we simply use the notation $D_\Box$.

Distributed plausibility is the plausibility the agents would have if they were to share their plausibility relations. The philosophical meaning of this is not immediately clear. We come back to this at the end of this section, where we show that in some circumstances distributed plausibility is the plausibility the agents would have if they were to share all their information.

As mentioned, $\sim_{\mathcal{N}}$ is an equivalence relation and $\leq_{\mathcal{N}}$ is reflexive and transitive. Additionally, the fact that $\leq_i \subseteq \sim_i$ for any $i \in \mathcal{N}$ implies that $\leq_{\mathcal{N}} = \bigcap_{i\in\mathcal{N}} \leq_i \subseteq \bigcap_{i\in\mathcal{N}} \sim_i = \sim_{\mathcal{N}}$. Because of this, the distributed knowledge and plausibility satisfy the same properties as their non-distributed variants:

Observation 2.23. Let F be a multi-agent plausibility frame and $P \subseteq W$. The following hold:

1. D_K satisfies the **S5**-axioms.
2. $D_\Box$ satisfies the **S4**-axioms.
3. $D_K P \subseteq D_\Box P$.

Furthermore, some special relations hold between the knowledge and plausibility of the individual agents and the distributed variants:

Proposition 2.24. Let F be a multi-agent plausibility frame, $i \in \mathcal{N}$ and $P \subseteq W$. The following hold:

1. $K_i P \subseteq D_K P$
2. $\Box_i P \subseteq D_\Box P$.

Proof. Left to the reader. □

Since distributed knowledge and plausibility satisfy the same properties as their non-distributed variants, we can define " distributed belief" in terms of D_K and $D_\Box$ in the same way as the B_i's were defined in terms of K_i's and $\Box_i$'s:

Definition 2.25. Let F be a multi-agent plausibility frame, $i \in \mathcal{N}$ and $P \subseteq W$. We define *distributed belief* as follows: $D_B^F P := D_K^F D_\Diamond^F D_\Box^F P$, where $D_\Diamond^F Q := \overline{D_\Box^F \overline{Q}}$ for any $Q \subseteq W$. Whenever the frame is fixed or understood, we simply use the notation D_B.

Observation 2.26. Let F be a multi-agent plausibility frame, $i \in \mathcal{N}$ and $P \subseteq W$. Then $w \in D_B P$ iff for all w' such that $w \sim_{\mathcal{N}} w'$ there exists a w'' such that $w' \leq_{\mathcal{N}} w''$ and for all w''' such that $w'' \leq_{\mathcal{N}} w'''$ it holds that $w''' \in P$.

Technically speaking, distributed belief is not an ideal name since it is not defined in terms of the beliefs of the individuals. For instance, $B_i P \subseteq D_B P$ does not hold. This is because belief is non-monotonic. An agent's knowledge can only increase, but that does not hold for an agent's belief. For instance, if an agent believes a proposition that goes against the knowledge of another agent, the belief in this proposition has to be given up when agents share information.

The name potential group belief would be more apt, because under some circumstances distributed belief is the belief the agents would have if they were to share all their knowledge, as is shown at the end of this section. We have decided to use the term distributed belief, however, because D_B is the counterpart of the B_i's in the same way as D_K and $D_\Box$ are of the K_i's and $\Box_i$'s respectively. Therefore, it is also the case that distributed knowledge implies distributed belief:

Observation 2.27. Let F be a multi-agent plausibility frame and $P \subseteq W$. Then $D_K P \subseteq D_B P$.

We now come back to the philosophical interpretation of distributed plausibility and distributed belief. Recall that distributed knowledge is the knowledge the agents would have if they were to share all their hard information. Hard knowledge is fully introspective. Given any proposition, the agent knows whether he does or does not know it. Thus, sharing all your hard information with somebody else is achievable. For distributed plausibility, things are more complicated. For starters, it is not clear what plausibility philosophically means. As mentioned, we could think of it as some kind of soft, unconscious knowledge. This form of knowledge is, however, not fully introspective. In other words, given a proposition, it does not have to be the case that the agent softly knows whether he knows it or not. This begs the question how it is possible to share one's unconscious knowledge. This difficulty disappears under the common prior assumption, because in that case $\leq_{\mathcal{N}}$ is independent of the individual plausibility relations:

Proposition 2.28. Let F be a common prior frame and let $\leq$ be the common prior plausibility relation. Then $\leq_{\mathcal{N}} = \leq \cap \sim_{\mathcal{N}}$.

Proof. $\leq_{\mathcal{N}} = \bigcap_{i \in \mathcal{N}} \leq_i = \bigcap_{i \in \mathcal{N}} (\leq \cap \sim_i) = \leq \cap \bigcap_{i \in \mathcal{N}} \sim_i = \leq \cap \sim_{\mathcal{N}}$. □

The interpretation of Proposition 2.28 is that, under the common prior assumption, distributed plausibility can be thought of as the plausibility the group of agents would have if they were to share all their hard information. As distributed belief is defined in terms of distributed knowledge and distributed plausibility, this result also holds for distributed belief. That is, distributed belief can be interpreted as the common belief of the group of agents after sharing all their hard information.

Concludingly, we can say that under the common prior assumption distributed belief/plausibility can be interpreted in the same way as distributed knowledge, i.e. as the belief/plausibility the agents would have if they were to share all their information. As sharing one's hard information is achievable, we may also speak of potential rather than distributed knowledge and belief. Notice that the common prior assumption says that differences in plausibility are solely due to differences in the hard information of the agents. Therefore, if the differences in hard information disappear, the plausibility relations of all agents become the same. As a consequence, after sharing their information, all agents have the same epistemic doxastic state.

2.4 Epistemic Doxastic Logic with Group Knowledge

After discussing multi-agent plausibility frames and the concepts of distributed knowledge and belief, this section introduces a formal language for these frames. By doing this, we obtain an epistemic doxastic logic with group knowledge. As this logic forms the basis for the rest of this thesis, we simply refer to it as EDL. Section 2.4.1 introduces both the syntax and the semantics of EDL. The proof system Λ_{EDL} is introduced in Section 2.4.2. Soundness of Λ_{EDL} is proven in Section 2.4.3 and completeness in Section 2.4.4.

2.4.1 Syntax and Semantics of EDL

The language of EDL is defined recursively as follows:

Definition 2.29. Let P be a finite set of propositional letters and $\mathcal{N}$ a finite set of agents. The set $\mathcal{L}_{EDL}$ of formulas of φ of EDL is defined recursively:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid \Box_i\varphi \mid D_K\varphi \mid D_\Box\varphi$$

where $p \in P$ and $i \in \mathcal{N}$. We define $\bot := p \wedge \neg p$ and $\top := \neg\bot$. The Boolean connectives $\vee$ and $\rightarrow$ are defined in terms of $\neg$ and $\wedge$ in the standard manner. The duals of the modal operators are defined in the following way: $\hat{K_i} := \neg K_i \neg$ and $\Diamond_i := \neg\Box_i\neg$. Belief is defined as $B_i := K_i\Diamond_i\Box_i$ and distributed belief as $D_B := D_K D_\Diamond D_\Box$, where $D_\Diamond := \neg D_\Box \neg$.

This language is interpreted over multi-agent plausibility models, which consist of a multi-agent plausibility frame together with a valuation function:

Definition 2.30. A *model for EDL* or a *multi-agent plausibility model* is a tuple $M = (F, V)$ where $F = (W, \sim_i, \leq_i)_{i\in\mathcal{N}}$ is a multi-agent plausibility frame and $V : P \rightarrow \mathcal{P}(W)$ is a valuation function. If F is a common prior frame, M is called a *common prior model*.

Truth in models is defined as follows:

Definition 2.31. Let $M = (W, \sim_i, \leq_i, V)_{i\in\mathcal{N}}$ be a model for EDL, $w \in W$ and $i \in \mathcal{N}$. The *satisfaction* relation $\models$ between pairs (M, w) and formulas $\varphi \in \mathcal{L}_{EDL}$ is defined as follows:

- $M, w \models p$ iff $w \in V(p)$
- $M, w \models \neg\varphi$ iff $M, w \not\models \varphi$
- $M, w \models \varphi_1 \wedge \varphi_2$ iff $M, w \models \varphi_1$ and $M, w \models \varphi_2$
- $M, w \models K_i\varphi$ iff for all $w' \in W(w \sim_i w'$ implies $M, w' \models \varphi)$
- $M, w \models \Box_i\varphi$ iff for all $w' \in W(w \leq_i w'$ implies $M, w' \models \varphi)$
- $M, w \models D_K\varphi$ iff for all $w' \in W(w \sim_{\mathcal{N}} w'$ implies $M, w' \models \varphi)$
- $M, w \models D_\Box\varphi$ iff for all $w' \in W(w \leq_{\mathcal{N}} w'$ implies $M, w' \models \varphi)$

A formula φ is *valid* if for all $M = (W, \sim_i, \leq_i, V)_{i\in\mathcal{N}}$ and all $w \in W$, we have $M, w \models \varphi$. We denote validity by $\models \varphi$.

Next, we define a truth map from formulas to sets of worlds:

Definition 2.32. Let $M = (W, \sim_i, \leq_i, V)_{i\in\mathcal{N}}$ be a model for EDL. We define a *truth map* from formulas to sets of worlds $\|\cdot\| : \mathcal{L}_{EDL} \rightarrow \mathcal{P}(W)$ such that $\|\varphi\| := \{w \in W \mid M, w \models \varphi\}$.

Observe that this truth map relates the Kripke modalities from the previous sections with the corresponding syntactic symbols:

Observation 2.33. Let $M = (W, \sim_i, \leq_i, V)_{i \in \mathcal{N}}$ be a model for EDL and let $\|\cdot\| : \mathcal{L}_{EDL} \to \mathcal{P}(W)$ be the truth map. Then the following hold for all $\varphi, \psi \in \mathcal{L}_{EDL}$: :

- $\|p\| = V(p)$,
- $\|\neg\varphi\| = \overline{\|\neg\varphi\|}$,
- $\|\varphi \wedge \psi\| = \|\varphi\| \cap \|\psi\|$,
- $\|K_i\varphi\| = K_i\|\varphi\|$,
- $\|\Box_i\varphi\| = \Box_i\|\varphi\|$,
- $\|D_K\varphi\| = D_K\|\varphi\|$,
- $\|D_\Box\varphi\| = D_\Box\|\varphi\|$,

where on the left side K_i, $\Box_i$, D_K and $D_\Box$ denote the syntactic symbol and on the right side the Kripke modality.

2.4.2 The Proof System of EDL

Definition 2.34. The proof system of EDL (notation: Λ_{EDL}) includes the following *axioms* for all formulas φ and ψ in the language of EDL and all $i \in \mathcal{N}$:

1. All tautologies of propositional logic.
2. The **S5**-axioms for individual and distributed knowledge:
   (a) $K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$
   (b) $K_i\varphi \to \varphi$
   (c) $K_i\varphi \to K_iK_i\varphi$
   (d) $\neg K_i\varphi \to K_i\neg K_i\varphi$
   (e) $D_K(\varphi \to \psi) \to (D_K\varphi \to D_K\psi)$
   (f) $D_K\varphi \to \varphi$
   (g) $D_K\varphi \to D_KD_K\varphi$
   (h) $\neg D_K\varphi \to D_K\neg D_K\varphi$
3. The **S4**-axioms for individual and distributed plausibility:
   (a) $\Box_i(\varphi \to \psi) \to (\Box_i\varphi \to \Box_i\psi)$
   (b) $\Box_i\varphi \to \varphi$
   (c) $\Box_i\varphi \to \Box_i\Box_i\varphi$
   (d) $D_\Box(\varphi \to \psi) \to (D_\Box\varphi \to D_\Box\psi)$
   (e) $D_\Box\varphi \to \varphi$
   (f) $D_\Box\varphi \to D_\Box D_\Box\varphi$
4. Knowledge implies plausibility:
   (a) $K_i\varphi \to \Box_i\varphi$

(b) $D_K\varphi \to D_\Box\varphi$

5. All individual knowledge/plausibility is distributed knowledge/plausibility:
   (a) $K_i\varphi \to D_K\varphi$
   (b) $\Box_i\varphi \to D_\Box\varphi$

The proof system of EDL includes the following *inference rules* for all formulas φ and ψ in the language of EDL and all $i \in \mathcal{N}$:

1. Modus Ponens: From φ and $\varphi \to \psi$, infer ψ.
2. Necessitation for all the modalities:
   (a) From φ, infer $K_i\varphi$.
   (b) From φ, infer $\Box_i\varphi$.
   (c) From φ, infer $D_K\varphi$.
   (d) From φ, infer $D_\Box\varphi$.

2.4.3 Soundness of EDL

Soundness captures the idea that everything that can be proven in a proof system is valid. In other words, the proof system cannot prove falsities:

Theorem 2.35. Λ_{EDL} is sound with respect to the class of all EDL-models. That is, for all $\varphi \in \mathcal{L}_{EDL}$: $\vdash \varphi$ implies $\models \varphi$.

Proof. Soundness is proven by *induction on the length of the proof*. Therefore, it suffices to show that each axiom is sound and that the inference rules preserve truth. The soundness of the axioms follows from the semantics of the modalities, Observation 2.33 and the properties of the epistemic and doxastic notions. The soundness of the **K**-axiom for any of the modal operators is trivial. For the soundness of the **S5**-axioms for K_i and D_K, see Proposition 2.5 and Observation 2.23.1. For the soundness of the **S4**-axioms for $\Box_i$ and $D_\Box$, see Proposition 2.8 and Observation 2.23.2. The soundness of $K_i\varphi \to \Box_i\varphi$ and $D_K\varphi \to D_\Box\varphi$ follows from Observations 2.7 and 2.23.3 respectively. Lastly, Proposition 2.24 gives us the soundness of $K_i\varphi \to D_K\varphi$ and $\Box_i\varphi \to D_\Box\varphi$.

It remains to be shown that the inference rules preserve validity. For modus ponens, suppose φ is valid and $\varphi \to \psi$ is valid. That is, in all models and all worlds both φ and $\varphi \to \psi$ hold. Obviously, ψ holds in all models and all worlds as well, i.e. ψ is valid. For generalisation, suppose φ is valid. That is, φ is valid in all models and all worlds. Let M be an EDL-model and w a world. No matter what accessibility relation we use to get from w to world w', it is the case that $M, w \models \varphi$. Hence necessitation preserves validity. □

Since Λ_{EDL} is sound with respect to the class of all EDL-models, we get the following soundness result as a corollary:

Corollary 2.36. Λ_{EDL} is sound with respect to the class of all common prior models.

2.4.4 Completeness of EDL

This section proves weak completeness of EDL. The idea behind weak completeness is that a formula φ is provable in Λ_{EDL}, if φ is valid on all models belonging to a certain class $\mathcal{C}$:

Definition 2.37. Let Λ be a logic and $\mathcal{C}$ a class of models. Let $\mathcal{C} \models \varphi$ denote the fact that for all $M \in \mathcal{C}$ and all $w \in W$, $M, w \models \varphi$. Λ is *weakly complete* with respect to $\mathcal{C}$ if for any formula φ, $\mathcal{C} \models \varphi$ implies $\vdash \varphi$.

As is shown in Blackburn et al. (2001: 194-195), this is equivalent to saying that every consistent formula φ is satisfiable on some model in $\mathcal{C}$:

Proposition 2.38. Let Λ be a logic and $\mathcal{C}$ a class of models. Λ is weakly complete with respect to $\mathcal{C}$ iff every Λ-consistent formula is satisfiable on some $M \in \mathcal{C}$.

This section shows that Λ_{EDL} is weakly complete with respect to the class of all common prior models. Since the language of EDL cannot distinguish between common prior and non common prior models, this immediately gives us weak completeness with respect to EDL-models in general.

Outline of the Completeness Proof

The goal is to show that Λ_{EDL} is weakly complete with respect to the class of all common prior models. The ideas behind the completeness proof for EDL are based on the method used by Fagin et al. (1992) to prove completeness for the logic of distributed knowledge. We briefly describe the main ideas behind the completeness proof, as this helps to keep the bigger picture in mind when going through the technical details of the proof.

The completeness proof consists of three steps. In step 1, the notion of a model is generalised to that of a pseudo-model. In a pseudo-models, the modal operators $K_i, \Box_i, D_K$ and $D_\Box$ each receive its own accessibility relation. These relations satisfy all the conditions that hold in normal models *except* for the following:

- $\bigcap_{i \in \mathcal{N}} \sim_i \subseteq \sim_{\mathcal{N}}$
- $\bigcap_{i \in \mathcal{N}} \leq_i \subseteq \leq_{\mathcal{N}}$

Step 1 proves completeness for pseudo-models with the standard canonical model – or more appropriately, *canonical pseudo-model* – construction, as discussed in Blackburn et al. (2001: 196-201). In other words, we show that given a certain consistent EDL-formula φ, it is possible to construct a pseudo-model that satisfies φ. The next steps transform the canonical pseudo-model into a common prior model in a truth preserving way

In step 2, the canonical pseudo-model is *unravelled into a tree*. The worlds in this tree correspond to finite "histories" from a fixed world w_0 to some other world w_n in the canonical pseudo-model, via any combination of relations. This tree is neither a model nor a pseudo-model, but it does satisfy the same EDL-formulas as the canonical pseudo-model, because there is a bounded morphism from the tree to the canonical pseudo-model.[7] Furthermore, the tree satisfies the property of *non-redundant path uniqueness*, which means that there is a unique non-redundant path between any two worlds in the tree. This property plays in important role when we modify the tree into a common prior model in the next step.

Step 3 *redefines the accessibility relations* of the tree in such a way that we obtain a common

[7] Technically speaking, it does not make sense to speak about a bounded morphism between the canonical pseudo-model and the tree since their structure differs. However, one can introduce the notion of a general model $M = (W, R_i^1, R_i^2, R^3, R^4, V)_{i \in \mathcal{N}}$, where W is a set of worlds, the relations are binary and V is a valuation function. Satisfaction in general models can be defined as in Definition 2.31, where R_i^1 is a the accessibility relation for K_i, R_i^2 for $\Box_i$, R^3 for D_K and R^4 for $D_\Box$. Furthermore, one can define the notion of a bounded morphism between general models in the spirit of Blackburn et al. (2001), Definition 2.10. As both the canonical pseudo-model and the tree are general models, we can then sensibly speak about a bounded morphism between the two. In the outline of the proof, however, we skip these details because it is the intuitive idea that matters.

prior model in the true sense of the world. In particular, the uniqueness of non-redundant paths gives us that $\sim_{\mathcal{N}} = \bigcap_{i\in\mathcal{N}} \sim_i$ and $\leq_{\mathcal{N}} = \bigcap_{i\in\mathcal{N}} \leq_i$. The modification of the relations happens in such a way that the truth of EDL-formulas is preserved. Thus, we have a common prior model that satisfies the same EDL-formulas as the canonical pseudo-model.

In conclusion, suppose φ is a consistent EDL-formula. The canonical pseudo-model construction gives us a pseudo-model satisfying φ. This model can be transformed in a truth preserving way into a common prior model. Thus, given a consistent EDL-formula φ, there is common prior model satisfying it. In other words, Λ_{EDL} is weakly complete with respect to the class of all common prior models.

Step 1: The Canonical Pseudo-Model Construction

Step 1 proves completeness of EDL with respect to pseudo-models. Step 1a defines pseudo-models and introduces the notion of a bounded morphism. Step 1b is concerned with the actual construction of the canonical pseudo-model.

Step 1a: Pseudo-Models

A pseudo-model is defined as follows:

Definition 2.39. A *pseudo-model* for EDL is a tuple $K = (W, \sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}}, V)_{i\in\mathcal{N}}$ such that W is a set of possible worlds, V is a valuation function, $\sim_i, \sim_{\mathcal{N}} \subseteq W \times W$ are equivalence relations, and $\leq_i, \leq_{\mathcal{N}} \subseteq W \times W$ are reflexive and transitive relations, such that:

1. For all $i \in \mathcal{N}$: $\leq_i \subseteq \sim_i$.
2. For all $i \in \mathcal{N}$: $\sim_{\mathcal{N}} \subseteq \sim_i$.
3. For all $i \in \mathcal{N}$: $\leq_{\mathcal{N}} \subseteq \leq_i$.
4. $\leq_{\mathcal{N}} \subseteq \sim_{\mathcal{N}}$.

Satisfaction in pseudo-models is defined as in Definition 2.31.

Notice that all models for EDL are pseudo-models. The only difference between models and pseudo-models is that in latter $\sim_{\mathcal{N}} = \bigcap_{i\in\mathcal{N}} \sim_i$ and $\leq_{\mathcal{N}} = \bigcap_{i\in\mathcal{N}} \leq_i$ do not hold in general.

The idea of the completeness proof is to construct a canonical pseudo-model and transform it in several steps into a common prior model, which satisfies the same EDL-formulas as the canonical pseudo-model. For this, the notion of a bounded morphism between pseudo-models is important, as they preserve the truth of EDL-formulas:

Definition 2.40. Let $K = (W, \sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}}, V)_{i\in\mathcal{N}}$ and $K' = (W', \sim'_i, \leq'_i, \sim'_{\mathcal{N}}, \leq'_{\mathcal{N}}, V')_{i\in\mathcal{N}}$ be two pseudo-models for EDL. A mapping $f : K \to K'$ is a *bounded morphism* if the following hold:

1. For all $w \in W$: w and $f(w)$ satisfy the same propositional letters.
2. Forth condition: For all $\preceq \in \{\sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}} \mid i \in \mathcal{N}\}$ and all $w, v \in W$, if $w \preceq v$ then $f(w) \preceq' f(v)$.
3. Back condition: For all $\preceq \in \{\sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}} \mid i \in \mathcal{N}\}$, all $w \in W$ and all $v' \in W'$, if $f(w) \preceq' v'$ then there exists a $v \in W$ such that $w \preceq v$ and $f(v) = v'$.

Proposition 2.41. Let $K = (W, \sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}}, V)_{i\in\mathcal{N}}$ and $K' = (W', \sim'_i, \leq'_i, \sim'_{\mathcal{N}}, \leq'_{\mathcal{N}}, V')_{i\in\mathcal{N}}$ be two pseudo-models for EDL and let $f : K \to K'$ be a bounded morphism. Let φ be a formula and $w \in W$. Then $K, w \models \varphi$ iff $K', f(w) \models \varphi$.

Proof. By induction on the complexity of φ:

- Base Case: This follows from condition 1 on bounded morphisms.
- Inductive Step: Suppose that for all $w \in W$ and all formulas ψ of lower complexity than φ the following holds: $K, w \models \psi$ iff $K', f(w) \models \psi$. The Boolean cases where $\varphi = \neg\psi$ and $\varphi = \psi_1 \wedge \psi_2$ follow immediately from the induction hypothesis. Only the modalities remain. Let $\varphi = [R]\psi$, where $[R] \in \{K_i, \Box_i, D_K, D_\Box |\ i \in \mathcal{N}\}$. Let $\preceq$ be the accessibility relation corresponding to the Kripke modality $[R]$.

  For the left-to-right direction, suppose $K, w \models [R]\psi$. Consider $f(w)$ and let $v' \in W'$ be such that $f(w) \preceq' v'$. According to the back condition, there exists a $v \in W$ such that $w \preceq v$ and $f(v) = v'$. Since $K, w \models [R]\psi$, it follows that $K, v \models \psi$. By the induction hypothesis, it follows that $K', v' \models \psi$. Hence $K', f(w) \models [R]\psi$. For the right-to-left direction, suppose $K', f(w) \models [R]\psi$. Let $v \in W$ be such that $w \preceq v$. By the forth condition, it follows that $f(w) \preceq' f(v)$. Since $K', f(w) \models [R]\psi$, we get $M', f(v) \models \psi$. By the induction hypothesis, $K, v \models \psi$. Thus, $K, w \models [R]\psi$.

□

Step 1b: The Canonical Pseudo-Model

The idea of this step is to construct a canonical pseudo-model as described in Blackburn et al. (2001: 196-201). This construction guarantees that every consistent EDL-formula is satisfiable on the canonical pseudo-model. We start with some basic definitions:

Definition 2.42. A formula φ is *consistent* iff $\nvdash \neg\varphi$. Otherwise it is inconsistent.
Definition 2.43. A set of formulas $\Gamma \subseteq \mathcal{L}_{EDL}$ is consistent iff $\Gamma \nvdash \bot$.
Definition 2.44. A set of formulas Γ is *maximal consistent* if Γ is consistent and for all $\varphi \notin \Gamma$, $\Gamma \cup \{\varphi\}$ is inconsistent.

As the following lemma shows, any consistent set of formulas can be extended to a maximal consistent set:

Lemma 2.45. (Lindenbaum's Lemma) If Γ is a consistent set of EDL-formulas, then there is a a set Γ^+ such that $\Gamma \subseteq \Gamma^+$ and Γ^+ is maximal consistent.

Proof. The proof is a special case of Blackburn et al. (2001: 197). The language of EDL is countable. Let $\varphi_0, \varphi_1, ...$ be an enumeration of the formulas in the language. Γ^+ is defined as the union of a chain of consistent sets:

- $\Gamma_0 := \Gamma$
- $\Gamma_{n+1} := \begin{cases} \Gamma_n \cup \{\varphi_n\} & \text{if this is consistent} \\ \Gamma_n \cup \{\neg\varphi_n\} & \text{otherwise} \end{cases}$
- $\Gamma^+ := \bigcup_{n \in \mathbb{N}} \Gamma_n$

By construction, $\Gamma \subseteq \Gamma^+$. One can easily check that Γ^+ is maximal consistent. □

Maximal consistent sets have some handy properties:

Proposition 2.46. Let Γ be a maximal consistent set. The following hold:

1. Γ is closed under modus ponens: if $\varphi_1, \varphi_1 \rightarrow \varphi_2 \in \Gamma$, then $\varphi_2 \in \Gamma$.
2. If φ is a tautology, then $\varphi \in \Gamma$.
3. For all formulas φ: $\varphi \in \Gamma$ or $\neg\varphi \in \Gamma$.
4. $\varphi_1 \wedge \varphi_2 \in \Gamma$ iff $\varphi_1 \in \Gamma$ and $\varphi_2 \in \Gamma$.

5. Let $[R] \in \{K_i, \Box_i, D_K, D_\Box \mid i \in \mathcal{N}\}$. If $\{[R]\varphi \mid [R]\varphi \in \Gamma\} \vdash \psi$, then $[R]\psi \in \Gamma$.

Proof. The proof of items 1-4 are left to the reader. For item 5, suppose $\{[R]\varphi \mid [R]\varphi \in \Gamma\} \vdash \psi$. That is, there exists a finite subset $\Gamma' \subseteq \{[R]\varphi \mid [R]\varphi \in \Gamma\}$ such that $\Gamma' \vdash \psi$. Let the formulas in Γ' be denoted by $[R]\gamma_1, ..., [R]\gamma_n$. Thus, $\vdash [R]\gamma_1 \wedge ... \wedge [R]\gamma_n \to \psi$. By necessitation and the K-axiom, it follows that $\vdash [R]([R]\gamma_1 \wedge ... \wedge [R]\gamma_n) \to [R]\psi$. Moreover, it holds that $\vdash [R][R]\gamma_1 \wedge .... \wedge [R][R]\gamma_n \to [R]([R]\gamma_1 \wedge ... \wedge [R]\gamma_n)$. Combining these gives us that $\vdash [R][R]\gamma_1 \wedge ... \wedge [R][R]\gamma_n \to [R]\psi$. Notice that the proof system contains the axioms $[R]\varphi \to [R][R]\varphi$ for all formulas φ. Thus, for all $1 \leq j \leq n$ it holds that $[R][R]\gamma_j \in \Gamma$. Since Γ is closed under modus ponens, it follows that $[R]\psi \in \Gamma$. □

Maximal consistent sets are essential, because they are the building blocks of the canonical pseudo-model:

Definition 2.47. The *canonical pseudo-model* is a tuple $K^\Omega = (W^\Omega, \sim_i^\Omega, \leq_i^\Omega, \sim_\mathcal{N}^\Omega, \leq_\mathcal{N}^\Omega, V^\Omega)$ such that

- W^Ω is the set of all maximal consistent sets,
- for all $i \in \mathcal{N}$, $\sim_i$ is a binary relation defined by: $w \sim_i^\Omega w'$ iff $\forall\varphi(K_i\varphi \in w \Rightarrow \varphi \in w')$,
- for all $i \in \mathcal{N}$, $\leq_i$ is a binary relation defined by: $w \leq_i^\Omega w'$ iff $\forall\varphi(\Box_i\varphi \in w \Rightarrow \varphi \in w')$,
- $\sim_\mathcal{N}$ is a binary relation defined by: $w \sim_\mathcal{N}^\Omega w'$ iff $\forall\varphi(D_K\varphi \in w \Rightarrow \varphi \in w')$,
- $\leq_\mathcal{N}$ is a binary relation defined by: $w \leq_\mathcal{N}^\Omega w'$ iff $\forall\varphi(D_\Box\varphi \in w \Rightarrow \varphi \in w')$,
- V^Ω is a valuation function defined by: $V^\Omega(p) := \{w \in W^\Omega \mid p \in w\}$.

The following proposition shows that the canonical pseudo-model K^Ω is indeed a pseudo-model for EDL:

Proposition 2.48. $K^\Omega = (W^\Omega, \sim_i^\Omega, \leq_i^\Omega, \sim_\mathcal{N}^\Omega, \leq_\mathcal{N}^\Omega, V^\Omega)$ is a pseudo-model for EDL.

Proof. In order to show that K^Ω is a pseudo-model, we need to show the following:

1. For all $i \in \mathcal{N}$, $\sim_i^\Omega$ is an equivalence relation.
2. For all $i \in \mathcal{N}$, $\leq_i^\Omega$ is reflexive and transitive.
3. $\sim_\mathcal{N}$ is an equivalence relation.
4. $\leq_\mathcal{N}$ is reflexive and transitive.
5. For all $i \in \mathcal{N}$: $\leq_i^\Omega \subseteq \sim_i^\Omega$.
6. For all $i \in \mathcal{N}$: $\sim_\mathcal{N}^\Omega \subseteq \sim_i^\Omega$.
7. For all $i \in \mathcal{N}$: $\leq_\mathcal{N}^\Omega \subseteq \leq_i^\Omega$.
8. $\leq_\mathcal{N}^\Omega \subseteq \sim_\mathcal{N}^\Omega$.

For item 1, let $i \in \mathcal{N}$. Firstly, let $w \in W^\Omega$ and φ be such that $K_i\varphi \in w$. Since w contains all tautologies, it follows that $K_i\varphi \to \varphi \in w$. The fact that maximal consistent sets are closed under modus ponens gives us that $\varphi \in w$. By definition of $\sim_i^\Omega$, it follows that $w \sim_i^\Omega w$. Hence $\sim_i^\Omega$ is reflexive. Secondly, let $w, w', w'' \in W^\Omega$ be such that $w \sim_i^\Omega w'$ and $w' \sim_i^\Omega w''$. Let φ be such that $K_i\varphi \in w$. Notice that $K_i\varphi \to K_iK_i\varphi \in w$. By closure under modus ponens, it follows that $K_iK_i\varphi \in w$. From $w \sim_i^\Omega w'$ it follows that $K_i\varphi \in w'$ and from $w' \sim_i^\Omega w''$ that $\varphi \in w''$. Thus, $w \sim_i^\Omega w''$. Hence $\sim_i^\Omega$ is transitive. Thirdly, let $w, w' \in W^\Omega$ be such that $w \sim_i^\Omega w'$ and let φ be such that $K_i\varphi \in w'$. Suppose, for contradiction, that $K_i\varphi \notin w$. Because w is a maximal consistent set, it follows that $\neg K_i\varphi \in w$. Moreover, $\neg K_i\varphi \to K_i\neg K_i\varphi \in w$. By closure under modus ponens, $K_i\neg K_i\varphi \in w$. Since $w \sim_i^\Omega w'$, it follows that $\neg K_i\varphi \in w'$. This is a contradiction. Thus, it must be the case that $K_i\varphi \in w$. From the reflexivity of $\sim_i^\Omega$, it follows that $\varphi \in w$. From the definition of $\sim_i^\Omega$, we get that $w' \sim_i^\Omega w$. Hence

$\sim_i^\Omega$ is symmetric. In conclusion, $\sim_i^\Omega$ is an equivalence relation because it is reflexive, transitive and symmetric. Items 2-4 can be proven in a similar way.

For item 5, let $i \in \mathcal{N}$ be arbitrary and let $w, w' \in W^\Omega$ be such that $w \leq_i^\Omega w'$. Let φ be such that $K_i\varphi \in w$. Notice that $K_i\varphi \to \Box_i\varphi \in w$. Closure under modus ponens gives us that $\Box_i\varphi \in w$. From the definition of $\leq_i^\Omega$, it follows that $\varphi \in w'$. Thus, $w \sim_i^\Omega w'$. Hence $\leq_i^\Omega \subseteq \sim_i^\Omega$. For item 6, let $i \in \mathcal{N}$ be arbitrary and let $w, w' \in W^\Omega$ be such that $w \sim_\mathcal{N} w'$. Let φ be such that $K_i\varphi \in w$. Notice that $K_i\varphi \to D_K\varphi \in w$, as w contains all tautologies. Thus, $D_K\varphi \in w$, by closure under modus ponens. By definition of $\sim_\mathcal{N}$, it follows from $w \sim_\mathcal{N} w'$ and $D_K\varphi \in w$, that $\varphi \in w'$. By definition, $w \sim_i^\Omega w'$. Hence $\sim_\mathcal{N}^\Omega \subseteq \sim_i^\Omega$. The proof of item 7 is similar to that of item 6 and the proof of item 8 to that of item 5. □

In the canonical pseudo-model, the valuation function is defined in such a way that truth of a propositional letter at a certain world is the same as membership of that world. The following lemma lifts this property from propositional letters to arbitrary EDL-formulas:

Lemma 2.49. (Truth Lemma) For all formula φ and all $w \in W^\Omega$, $w \models \varphi$ iff $\varphi \in w$.

Proof. We prove this by induction on the complexity of φ.

- Base Case: This follows from the definition of V^Ω.
- Inductive Case: Suppose that for all $w \in W$ and all formulas ψ of lower complexity than φ, we have $w \models \psi$ iff $\psi \in w$. The Boolean cases where $\varphi = \neg\psi$ and $\varphi = \psi_1 \wedge \psi_2$ follow from the induction hypothesis together with Proposition 2.46.3 and 2.46.4 respectively. Only the modalities remain. Let $\varphi = [R]\psi$, where $[R] \in \{K_i, \Box_i, D_K, D_\Box \mid i \in \mathcal{N}\}$. Let $\preceq$ be the accessibility relation corresponding to the Kripke modality $[R]$.

  For the left-to-right direction, suppose $w \models [R]\psi$. First, we show that $\{[R]\chi \mid [R]\chi \in w\} \vdash \psi$. Suppose, for contradiction, that $\{[R]\chi \mid [R]\chi \in w\} \nvdash \psi$. That is, the set $\Gamma = \{\neg\psi\} \cup \{[R]\chi \mid [R]\chi \in w\}$ is consistent. By Lindenbaum's lemma, Γ can be extended to a maximal consistent set Γ^+. Let $w' := \Gamma^+$. From the fact $\{[R]\chi \mid [R]\chi \in w\} \subseteq w'$ and the reflexivity of $\preceq^\Omega$, it follows that $\chi \in w'$ for all $[R]\chi \in w$. Thus, $w \preceq^\Omega w'$. From $w \models [R]\psi$, we get $w' \models \psi$. By the induction hypothesis, $\psi \in w'$. This contradicts the fact that $\neg\psi \in w'$. Thus, $\{[R]\chi \mid [R]\chi \in w\} \vdash \psi$. From Proposition 2.46.5, it follows that $[R]\psi \in w$.

  For the right-to-left direction, suppose $[R]\psi \in w$. Suppose, for contradiction, that $w \nvDash [R]\psi$. Then, there exists a $w' \in W^\Omega$ such that $w \preceq^\Omega w'$ and $w' \nvDash \psi$. By the induction hypothesis, it follows that $\psi \notin w'$. However, from the fact that $w \preceq^\Omega w'$ and $[R]\psi \in w$, it follows that $\psi \in w'$. This is a contradiction. Thus, $w \models [R]\psi$.

□

With the truth lemma established, one can show that Λ_{EDL} is complete with respect to the class of all pseudo-models:

Lemma 2.50. Λ_{EDL} is weakly complete with respect to the class of all pseudo-models.

Proof. By Proposition 2.38, it suffices to show that every consistent EDL-formula is satisfiable on some pseudo-model. Let φ be a consistent formula. By Lindenbaum's Lemma, there is a maximal consistent set Γ such that $\{\varphi\} \subseteq \Gamma$. By definition of the canonical pseudo-model, there exists a $w \in W^\Omega$ such that $\Gamma = w \ni \varphi$. By the truth lemma, we get that $K^\Omega, w \models \varphi$. □

Step 2: Unravelling the Canonical Pseudo-Model

Step 1 has shown that EDL is weakly complete with respect to the class of all pseudo-models. The ultimate goal is, however, to prove weak completeness with respect to the class of all common prior models. Therefore, we transform the canonical pseudo-model in several steps into a common prior model in a truth preserving way. After the introduction of some preliminary notions in step 2a, step 2b unravels the canonical pseudo-model into a tree.

Step 2a: Preliminaries

We start by introducing some useful notation:

Notation. Let X be a set and $R \subseteq X \times X$ a binary relation. We use the following notation:

- $R^{-1} := \{(x, y) \in X \times X \mid (y, x) \in R\}$.
- R^* is called the reflexive, transitive closure of R and denotes the smallest reflexive and transitive relation containing R.
- $R^\sim$ is called the equivalence closure of R and denotes the smallest reflexive, transitive and symmetric relation containing R.

Observation 2.51. Let X be a set and $R \subseteq X \times X$ a binary relation. $R^\sim = (R \cup R^{-1})^*$.

In addition, we define the notion of a labelled tree:

Definition 2.52. Let $F = (W, R_0, ..., R_n)$ with $n \in \mathbb{N}$ be a tuple such that W is a set of possible worlds and for all $k \leq n, R_k \subseteq W \times W$ is a binary relation. Define $S := \bigcup_{k \leq n} R_n$. F is a *labelled tree* iff the following hold:

1. Root: There exists a unique $r \in W$ such that for all $w \in W : rS^*w$.
2. No Cycles: There does not exists a sequence $wR_0...R_mw'$ such that for all $k \leq m : R_k \in S$ and $w = w'$.
3. Unique Predecessor and Relation: For all $w, w', v \in W$ and all R_k, R_m with $k, m \leq n$, wR_kv and $w'R_mv$ implies $w = w'$ and $R_k = R_m$.

Step 2b: Unravelling

Before getting to the actual unravelling, histories and their compositions are introduced:

Definition 2.53. Let $K^\Omega = (W^\Omega, \sim_i^\Omega, \leq_i^\Omega, \sim_\mathcal{N}^\Omega, \leq_\mathcal{N}^\Omega, V^\Omega)_{i \in \mathcal{N}}$ be the canonical pseudo-model. Let $w \in W^\Omega$. A *history with origin* w is a finite sequence $\bar{h} := (w_0, R_0, w_1, ..., R_{n-1}, w_n)$ such that:

- for all $k \leq n : w_k \in W^\Omega$,
- $w_0 = w$,
- for all $k < n : R_k \in \{\sim_i^\Omega, \leq_i^\Omega, \sim_\mathcal{N}^\Omega, \leq_\mathcal{N}^\Omega \mid i \in \mathcal{N}\}$,
- for all $k < n : w_kR_kw_{k+1}$.

For any history $\bar{h} = (w_0, R_0, w_1, ..., R_{n-1}, w_n)$, let $\mathit{first}(\bar{h}) := w_0$ and $\mathit{last}(\bar{h}) := w_n$.

Definition 2.54. Let $\bar{h} = (w_0, R_0, w_1, ..., R_{n-1}, w_n)$ and $\bar{h}' = (w'_0, R'_0, w'_1, ..., R'_{m-1}, w'_m)$ be two histories such that $\mathit{last}(\bar{h}) = \mathit{first}(\bar{h}')$. The *composed history* $\bar{h}+\bar{h}'$ is defined as $\bar{h}+\bar{h}' := (w_0, R_0, w_1, ..., R_{n-1}, w_n = w'_0, R'_0, w'_1, ..., R'_{m-1}, w'_m)$.

The histories from the canonical pseudo-model will be the worlds in the unravelled tree, which is defined as follows:

Definition 2.55. Let $K^\Omega = (W^\Omega, \sim_i^\Omega, \leq_i^\Omega, \sim_\mathcal{N}^\Omega, \leq_\mathcal{N}^\Omega, V^\Omega)$ be the canonical pseudo-model. Let $w \in W^\Omega$. The unravelling of K^Ω around w is a tuple $\vec{K} = (\vec{W}, R_{\sim_i}, R_{\leq_i}, R_{\sim_\mathcal{N}}, R_{\leq_\mathcal{N}}, \vec{V})_{i\in\mathcal{N}}$ such that:

- $\vec{W}$ is the set of all histories with origin w.
- For any $\preceq \in \{\sim_i^\Omega, \leq_i^\Omega, \sim_\mathcal{N}^\Omega, \leq_\mathcal{N}^\Omega \mid i \in \mathcal{N}\}$, $R_\preceq \subseteq \vec{W} \times \vec{W}$ is a binary relation defined as follows: $\bar{h} R_\preceq \bar{h}'$ iff $\bar{h} + (last(\bar{h}), \preceq, w') = \bar{h}'$.
- $\vec{V} : P \to \mathcal{P}(\vec{W})$ is a valuation function such that $\vec{V}(p) := \{\bar{h} \in \vec{W} \mid last(\bar{h}) \in V^\Omega(p)\}$.

The resulting structure is indeed a labelled tree:

Proposition 2.56. $\vec{K} = (\vec{W}, R_{\sim_i}, R_{\leq_i}, R_{\sim_\mathcal{N}}, R_{\leq_\mathcal{N}})_{i\in\mathcal{N}}$ is a labelled tree.

Proof. Firstly, notice that (w) is the root of the tree. Secondly, the definitions of the relations ensure that we do not have any cycles, because whenever a history $\bar{h}$ is related to $\bar{h}'$, $\bar{h}'$ is a strictly longer sequence than $\bar{h}$. Thirdly, every history $\bar{h}$ different from (w) has precisely one unique predecessor which is uniquely related to $\bar{h}$. This predecessor can be obtained by removing the last two elements, i.e. the last world and the last relation, from $\bar{h}$. □

The unravelled tree structure is no longer a pseudo-model, because all the relational properties have been destroyed. However, one can show that the tree structure and the canonical pseudo-model satisfy the same EDL-formulas.[8] Although this is important for the bigger picture, it is not strictly needed for the completeness proof. Therefore, we do not elaborate on it any further. The next step in the completeness proof is to transform the unravelled tree structure into a common prior model, which will be done in step 3. In order to show in this step that there exists a bounded morphism between the resulting common prior model and the canonical pseudo-model, we need an important property of the unravelled tree structure: non-redundant path uniqueness. We start by defining paths and their compositions:

Definition 2.57. Let $\vec{K} = (\vec{W}, R_{\sim_i}, R_{\leq_i}, R_{\sim_\mathcal{N}}, R_{\leq_\mathcal{N}}, \vec{V})_{i\in\mathcal{N}}$ be the unravelling of the canonical pseudo-model K^Ω around some world $w \in W^\Omega$. Let $\mathcal{R} \subseteq \{R_{\sim_i}, R_{\sim_i}^{-1}, R_{\leq_i}, R_{\leq_i}^{-1}, R_{\sim_\mathcal{N}}, R_{\sim_\mathcal{N}}^{-1}, R_{\leq_\mathcal{N}}, R_{\leq_\mathcal{N}}^{-1} \mid i \in \mathcal{N}\} =: Rel$. An $\mathcal{R}$-*path* from $\bar{h}$ to $\bar{h}'$ is a finite sequence $\bar{p} := (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$ such that:

- for all $k \leq n : \bar{h}_k \in \vec{W}$,
- $\bar{h}_0 = \bar{h}$,
- $\bar{h}_n = \bar{h}'$,
- for all $k < n : R_k \in \mathcal{R}$,
- for all $k < n : \bar{h}_k R_k \bar{h}_{k+1}$.

If $\mathcal{R}$ is not specified, we speak of a *path*. For any path $\bar{p} := (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$, we define $first(\bar{p}) := \bar{h}_0$ and $last(\bar{p}) := \bar{h}_n$.

[8] In order to prove this, one has to introduce the notion of a general model $M = (W, R_i^1, R_i^2, R^3, R^4, V)_{i\in\mathcal{N}}$, where W is a set of worlds, the relations are binary and V is a valuation function. Define satisfaction as in Definition 2.31, where R_i^1 is a the accessibility relation for K_i, R_i^2 for $\Box_i$, R^3 for D_K and R^4 for $D_\Box$. In the spirit of Blackburn et al. (2001), one needs to introduce the notion of a bounded morphism between general models and show that there exists a bounded morphism between K^Ω and $\vec{K}$. See Definition 2.10 and Lemma 4.52 from Blackburn et al. (2001) respectively. It then follows from from the fact that bounded morphism between general models preserve truth that K^Ω and $\vec{K}$ satisfy the same EDL-formulas. For this, see Proposition 2.14 of Blackburn et al. (2001).

Definition 2.58. Let $\bar{p} = (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$ and $\bar{p}' = (\bar{h}'_0, R'_0, \bar{h}'_1, ..., R'_{m-1}, \bar{h}'_m)$ be two paths such that $last(\bar{p}) = first(\bar{p}')$. The *composed path* $\bar{p}+\bar{p}'$ is defined as $\bar{p}+\bar{p}' := (\bar{h}'_0, R'_0, \bar{h}'_1, ..., R'_{n-1}, \bar{h}'_n = \bar{h}'_0, R'_0, \bar{h}'_1, ..., R'_{m-1}, \bar{h}'_m)$.

Notice that histories and paths are defined in a similar manner. The difference between them is that histories consists of worlds and relations that come from the pseudo-model, whereas paths consist of worlds and (the converse of) relations that come from the unravelled tree structure. Different names are used in order to avoid confusion.

The following observation about paths is useful:

Observation 2.59. Let $\mathcal{R} \subseteq Rel$ and let $S := (\bigcup_{R\in\mathcal{R}} R)^*$. For any $\bar{h}, \bar{h}' \in \vec{W}$ the following holds: $\bar{h}S\bar{h}'$ and $\bar{h} \neq \bar{h}'$ iff there exists an $\mathcal{R}$-path from $\bar{h}$ to $\bar{h}'$.

We now introduce the notion of a non-redundant path:

Definition 2.60. Let $\mathcal{R} \subseteq Rel$ and let $\bar{p} = (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$ be an $\mathcal{R}$-path. $\bar{p}$ is a non-redundant path if there is no $k < n-1$ such that $\bar{h}_k = \bar{h}_{k+2}$ and $R_{k+1} = R_k^{-1}$.

Intuitively, a non-redundant path is a path where we do not go forward along some edge and immediately backward on the same edge. In a labelled tree structure, there is precisely one non-redundant path between any two nodes:

Lemma 2.61. Let $\vec{K} = (\vec{W}, R_{\sim_i}, R_{\leq_i}, R_{\sim_\mathcal{N}}, R_{\leq_\mathcal{N}}, \vec{V})_{i\in\mathcal{N}}$ be the unravelling of the canonical pseudo-model K^Ω around some world $w \in W^\Omega$. Let $\bar{h}, \bar{h}' \in \vec{K}$ be such that $\bar{h} \neq \bar{h}'$. There is exactly one non-redundant path $\bar{p} = (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$ from $\bar{h}$ to $\bar{h}'$.

Proof. Recall that $\vec{K} = (\vec{W}, R_{\sim_i}, R_{\leq_i}, R_{\sim_\mathcal{N}}, R_{\leq_\mathcal{N}})_{i\in\mathcal{N}}$ is a labelled tree. Let $\bar{h}, \bar{h}' \in \vec{W}$ be such that $\bar{h} \neq \bar{h}'$. If $\bar{h}'$ is a subsequence of $\bar{h}$, the non-redundant path is obtained by going from $\bar{h}$ to $\bar{h}'$ via the unique immediate predecessors and the corresponding unique edges. Notice that this non-redundant path is unique. If $\bar{h}$ is a subsequence of $\bar{h}'$, we obtain the unique non-redundant path by taking the non-redundant path from $\bar{h}'$ to $\bar{h}$ and reversing it. If it is not the case that one is a subsequence of the other, we go to the longest history $\bar{h}^*$ that is a subsequence of both. $\bar{h}^*$ always exists since each world has only finitely many predecessors and (w) is a predecessor of any world different from itself. In addition, $\bar{h}^*$ is unique. We construct a non-redundant path from $\bar{h}$ to $\bar{h}'$ by composing the non-redundant path from $\bar{h}$ to $\bar{h}^*$ with the non-redundant path from $\bar{h}^*$ to $\bar{h}'$. As the composition of two non-redundant paths is non-redundant, this path is non-redundant as well. Since $\bar{h}^*$, the non-redundant paths from $\bar{h}$ to $\bar{h}^*$ and from $\bar{h}^*$ to $\bar{h}'$ are all unique, the resulting non-redundant path is as well. □

Thus, there is a unique non-redundant path between any two histories in the tree structure. Furthermore, one can show that any path between two histories contains the non-redundant path between them:

Lemma 2.62. Let $\vec{K} = (\vec{W}, R_{\sim_i}, R_{\leq_i}, R_{\sim_\mathcal{N}}, R_{\leq_\mathcal{N}}, \vec{V})_{i\in\mathcal{N}}$ be the unravelling of the canonical pseudo-model K^Ω around some world $w \in W^\Omega$. Let $\bar{h}, \bar{h}' \in \vec{W}$ and $\bar{p} = (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$ a path from $\bar{h}$ to $\bar{h}$. $\bar{p}$ contains the unique non-redundant path $\bar{p}^*$ from $\bar{h}$ and $\bar{h}$.

Proof. Let $\bar{p}^* = (\bar{h}^*_0, R^*_0, \bar{h}^*_1, ..., R^*_{m-1}, \bar{h}^*_m)$ be the unique non-redundant path from $\bar{h}$ and $\bar{h}'$. Suppose, for contradiction, that $\bar{p}$ does not contain $\bar{p}^*$. This means that there exists an edge in the non-redundant path $\bar{p}^*$ that does not occur in $\bar{p}$. In other words, there exists a $k < m$ such that $(\bar{h}^*_k, R^*_k, \bar{h}^*_{k+1})$ is not a subsequence of $\bar{p}$. Notice that the following algorithm reduces $\bar{p}$ to a non-redundant path:

- Step 1: Remove all subsequences of the form $(\bar{h}_l, R_l, \bar{h}_{l+1}, R_{l+1}, \bar{h}_{l+2})$ such that $\bar{h}_l = \bar{h}_{l+2}$ and $R_{l+1} = R_l^{-1}$.

- Step 2: If the obtained path is non-redundant, stop. If the obtained path is redundant, go back to step 1.

Notice that this algorithm always ends, because the path is finite. Let $\bar{p}'$ be the reduction of $\bar{p}$. Notice that $(\bar{h}^*_k, R^*_k, \bar{h}^*_{k+1})$ is also not a subsequence of $\bar{p}'$. Thus, $\bar{p}'$ is non-redundant path from $\bar{h}$ and $\bar{h}'$ and $\bar{p}' \neq \bar{p}^*$. This contradicts Lemma 2.61, which says that there is a unique non-redundant path between any two nodes. Thus, every path from $\bar{h}$ to $\bar{h}'$ contains the unique non-redundant path $\bar{p}^*$. □

Step 3: Defining the Common Prior Model

Step 2 unravelled the canonical pseudo-model from step 1 in a truth preserving way. Unfortunately, all the relational properties of the canonical pseudo-model were destroyed. This step fixes this by keeping the worlds of the tree structure, but redefining the relations. These new relations give us a pseudo-model where $\sim_{\mathcal{N}} = \bigcap_{i\in\mathcal{N}} \sim_i$ and $\leq_{\mathcal{N}} = \bigcap_{i\in\mathcal{N}} \leq_i$. In order words, we get a multi-agent plausibility model. Moreover, we can show that this model is in fact a common prior model, which satisfies the same EDL-formulas as the canonical pseudo-model. Let us start by defining the new model:

Definition 2.63. Let $\vec{K} = (\vec{W}, R_{\sim_i}, R_{\leq_i}, R_{\sim_{\mathcal{N}}}, R_{\leq_{\mathcal{N}}}, \vec{V})_{i\in\mathcal{N}}$ be the unravelling of the canonical pseudo-model K^Ω around some world $w \in W^\Omega$. We define $M = (\vec{W}, \sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}}, \vec{V})_{i\in\mathcal{N}}$ to be such that:

- $\leq_{\mathcal{N}} := (R_{\leq_{\mathcal{N}}})^*$,
- $\sim_{\mathcal{N}} := (R_{\sim_{\mathcal{N}}} \cup R_{\leq_{\mathcal{N}}})^\sim$,
- For all $i \in \mathcal{N}$, $\leq_i := (R_{\leq_i} \cup R_{\leq_{\mathcal{N}}})^*$,
- For all $i \in \mathcal{N}$, $\sim_i := (R_{\sim_i} \cup R_{\leq_i} \cup R_{\sim_{\mathcal{N}}} \cup R_{\leq_{\mathcal{N}}})^\sim$.

Before showing that M is an EDL-model, we prove the following identities:

Proposition 2.64. For all $i \in \mathcal{N}$, the following hold:

1. $\leq_i = (R_{\leq_i} \cup \leq_{\mathcal{N}})^*$.
2. $\sim_i = (R_{\sim_i} \cup \leq_i \cup \sim_{\mathcal{N}})^\sim$.

Proof. Let $i \in \mathcal{N}$ be arbitrary. For item 1, we get $\leq_i = (R_{\leq_i} \cup R_{\leq_{\mathcal{N}}})^*$. Because we take the reflexive, transitive closure of the union in the end anyway, it holds that $(R_{\leq_i} \cup R_{\leq_{\mathcal{N}}})^* = (R_{\leq_i} \cup (R_{\leq_{\mathcal{N}}})^*)^* = (R_{\leq_i} \cup \leq_{\mathcal{N}})^*$. For item 2, notice that $(R_{\sim_i} \cup \leq_i \cup \sim_{\mathcal{N}})^\sim = (R_{\sim_i} \cup (R_{\leq_i} \cup R_{\leq_{\mathcal{N}}})^* \cup (R_{\sim_{\mathcal{N}}} \cup R_{\leq_{\mathcal{N}}})^\sim)^\sim = (R_{\sim_i} \cup R_{\leq_i} \cup R_{\sim_{\mathcal{N}}} \cup R_{\leq_{\mathcal{N}}})^\sim = \sim_i$. □

The next lemmas shows that M is common prior model for EDL:

Lemma 2.65. $M = (\vec{W}, \sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}}, \vec{V})_{i\in\mathcal{N}}$ is a model for EDL.

Proof. By definition of the relations, it follows that $\sim_i$ and $\sim_{\mathcal{N}}$ are equivalence relations and that $\leq_i$ and $\leq_{\mathcal{N}}$ are reflexive and transitive. Therefore, in order to prove that M is an EDL-model, it remains to show the following:

1. For all $i \in \mathcal{N} : \leq_i \subseteq \sim_i$.
2. $\leq_{\mathcal{N}} = \bigcap_{i\in\mathcal{N}} \leq_i$.
3. $\sim_{\mathcal{N}} = \bigcap_{i\in\mathcal{N}} \sim_i$.

Item 1 follows directly from Proposition 2.64.2.

The $\subseteq$-inclusion of item 2 follows from Proposition 2.64.1. For the $\supseteq$-inclusion, let $\bar{h}, \bar{h}' \in \vec{W}$ be such that $(\bar{h}, \bar{h}') \in \bigcap_{i \in \mathcal{N}} \leq_i$. That is, $\bar{h} \leq_i \bar{h}'$ for all $i \in \mathcal{N}$. Let $i \in \mathcal{N}$ be arbitrary. If $\bar{h} = \bar{h}'$, then $\bar{h} \leq_{\mathcal{N}} \bar{h}'$ by reflexivity of $\leq_{\mathcal{N}}$. Suppose $\bar{h} \neq \bar{h}'$. Notice that $\leq_i = (R_{\leq_i} \cup R_{\leq_{\mathcal{N}}})^*$. By Observation 2.59, there exists an $\mathcal{R}$-path in $\vec{K}$ from $\bar{h}$ to $\bar{h}'$ with $\mathcal{R} = \{R_{\leq_i}, R_{\leq_{\mathcal{N}}}\}$. By Lemma 2.62, this path contains the non-redundant $\mathcal{R}$-path $\bar{p} = (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$ from $\bar{h}$ to $\bar{h}'$. In fact, we can show that $\mathcal{R} = \{R_{\leq_{\mathcal{N}}}\}$. Suppose, for contradiction, that this is not the case. Then $\bar{p}$ contains the relation $R_{\leq_i}$. Recall that $(\bar{h}, \bar{h}') \in \bigcap_{i \in \mathcal{N}} \leq_i$. Let $j \in \mathcal{N}$ be such that $i \neq j$. Since $\bar{h} \leq_j \bar{h}'$, there is a non-redundant $\mathcal{R}'$-path $\bar{p}' = (\bar{h}'_0, R'_0, \bar{h}'_1, ..., R'_{m-1}, \bar{h}'_m)$ with $\mathcal{R}' = \{R_{\leq_j}, R_{\leq_{\mathcal{N}}}\}$ from $\bar{h}$ to $\bar{h}'$. Notice that $\bar{p} \neq \bar{p}'$, because $\bar{p}$ uses $R_{\leq_i}$ whereas $\bar{p}'$ does not. This contradicts the uniqueness of non-redundant paths, as shown in Lemma 2.61. Because path $\bar{p}$ contains only the $R_{\leq_{\mathcal{N}}}$ relation, it follows that $(\bar{h}, \bar{h}') \in (R_{\leq_{\mathcal{N}}})^* = \leq_{\mathcal{N}}$. Thus, $\bigcap_{i \in \mathcal{N}} \leq_i \subseteq \leq_{\mathcal{N}}$.

The $\subseteq$-inclusion of item 3 follows from Proposition 2.64.2. For the $\supseteq$-inclusion, let $\bar{h}, \bar{h}' \in \vec{W}$ be such that $(\bar{h}, \bar{h}') \in \bigcap_{i \in \mathcal{N}} \sim_i$. That is, $\bar{h} \sim_i \bar{h}'$ for all $i \in \mathcal{N}$. Let $i \in \mathcal{N}$ be arbitrary. If $\bar{h} = \bar{h}'$, then $\bar{h} \sim_{\mathcal{N}} \bar{h}'$ by reflexivity of $\sim_{\mathcal{N}}$. Suppose $\bar{h} \neq \bar{h}'$. Notice that $\sim_i = (R_{\sim_i} \cup R_{\leq_i} \cup R_{\sim_{\mathcal{N}}} \cup R_{\leq_{\mathcal{N}}})^\sim$. By Observations 2.51 and 2.59, it follows that there exists an $\mathcal{R}$-path in $\vec{K}$ from $\bar{h}$ to $\bar{h}'$ with $\mathcal{R} = \{R_{\sim_i}, R^{-1}_{\sim_i}, R_{\leq_i}, R^{-1}_{\leq_i}, R_{\sim_{\mathcal{N}}}, R^{-1}_{\sim_{\mathcal{N}}}, R_{\leq_{\mathcal{N}}}, R^{-1}_{\leq_{\mathcal{N}}}\}$. By Lemma 2.62, this path contains the non-redundant $\mathcal{R}$=path $\bar{p} = (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$ from $\bar{h}$ to $\bar{h}'$. In fact, we claim that $\mathcal{R} = \{R_{\leq_{\mathcal{N}}}, R^{-1}_{\leq_{\mathcal{N}}}, R_{\sim_{\mathcal{N}}}, R^{-1}_{\sim_{\mathcal{N}}}\}$. This claim can be proven by using an argument similar to the one used in the previous item. Because $\bar{p}$ only uses the relations $R_{\leq_{\mathcal{N}}}, R^{-1}_{\leq_{\mathcal{N}}}, R_{\sim_{\mathcal{N}}}$ and $R^{-1}_{\sim_{\mathcal{N}}}$, it follows $(\bar{h}, \bar{h}') \in (R_{\leq_N} \cup R_{\sim_{\mathcal{N}}})^\sim = \sim_{\mathcal{N}}$. Thus, $\bigcap_{i \in \mathcal{N}} \sim_i \subseteq \sim_{\mathcal{N}}$. □

Lemma 2.66. *M is a common prior model.*

Proof. In order to show that M is a common prior model, we need to prove that there exists a relation $\leq$ such that $\leq \cap \sim_i = \leq_i$ for all $i \in \mathcal{N}$. Let $\leq := (\bigcup_{i \in \mathcal{N}} R_{\leq_i} \cup R_{\leq_{\mathcal{N}}})^*$. We show that for all $i \in \mathcal{N}$, $\leq \cap \sim_i = \leq_i$. Let $i \in \mathcal{N}$ be arbitrary. Recall the definitions of the involved relations:

- $\leq_i = (R_{\leq_i} \cup R_{\leq_{\mathcal{N}}})^*$
- $\sim_i = (R_{\sim_i} \cup R_{\leq_i} \cup R_{\sim_{\mathcal{N}}} \cup R_{\leq_{\mathcal{N}}})^\sim$
- $\leq = (\bigcup_{i \in \mathcal{N}} R_{\leq_i} \cup R_{\leq_{\mathcal{N}}})^*$

Firstly, notice that $\leq_i \subseteq \leq \cap \sim_i$ follows from the fact that $\leq_i = (R_{\leq_i} \cup R_{\leq_{\mathcal{N}}})^* \subseteq (\bigcup_{i \in \mathcal{N}} R_{\leq_i} \cup R_{\leq_{\mathcal{N}}})^* = \leq$ and $\leq_i = (R_{\leq_i} \cup R_{\leq_{\mathcal{N}}})^* \subseteq (R_{\sim_i} \cup R_{\leq_i} \cup R_{\sim_{\mathcal{N}}} \cup R_{\leq_{\mathcal{N}}})^\sim = \sim_i$. It remains to be shown that $\leq \cap \sim_i \subseteq \leq_i$. Let $\bar{h}, \bar{h}' \in \vec{W}$ be such that $(\bar{h}, \bar{h}') \in \leq \cap \sim_i$. If $\bar{h} = \bar{h}'$, then $\bar{h} \leq_i \bar{h}'$ by reflexivity of $\leq_i$. If $\bar{h} \neq \bar{h}'$, it follows from the definition of $\leq$ and Observation 2.59 that there exists an $\mathcal{R}$-path in $\vec{K}$ from $\bar{h}$ to $\bar{h}'$ with $\mathcal{R} = \{R_{\leq_i}, R_{\leq_{\mathcal{N}}} \mid i \in \mathcal{N}\}$. By Lemma 2.62, this path contains the non-redundant $\mathcal{R}$-path $\bar{p} = (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$. Similarly, we can use the definition of $\sim_i$, Observations 2.51 and 2.59 and Lemma 2.62 to conclude that there exists a non-redundant $\mathcal{R}'$-path $\bar{p}' = (\bar{h}'_0, R'_0, \bar{h}'_1, ..., R'_{n-1}, \bar{h}'_m)$ from $\bar{h}$ to $\bar{h}'$ with $\mathcal{R}' = \{R_{\sim_i}, R^{-1}_{\sim_i}, R_{\leq_i}, R^{-1}_{\leq_i}, R_{\sim_{\mathcal{N}}}, R^{-1}_{\sim_{\mathcal{N}}}, R_{\leq_{\mathcal{N}}}, R^{-1}_{\leq_{\mathcal{N}}}\}$. By the uniqueness of non-redundant paths, it follows that $\bar{p} = \bar{p}'$. Thus, path $\bar{p}$ only uses the relations $R_{\leq_i}$ and $R_{\leq_{\mathcal{N}}}$. Consequently, $(\bar{h}, \bar{h}') \in (R_{\leq_i} \cup R_{\leq_{\mathcal{N}}})^* \subseteq \leq_i$. Thus, $\leq \cap \sim_i \subseteq \leq_i$. □

Recall that bounded morphisms between pseudo-models preserve the truth of EDL-formulas. Hence in order to show that M and the canonical pseudo-model satisfy the same EDL-formulas, it suffices to show that there exists a bounded morphism between them:

Lemma 2.67. Let $K^\Omega = (W^\Omega, \sim^\Omega_i, \leq^\Omega_i, \sim^\Omega_{\mathcal{N}}, \leq^\Omega_{\mathcal{N}}, V^\Omega)$ be the canonical pseudo-model and let $\vec{K} = (\vec{W}, R_{\sim_i}, R_{\leq_i}, R_{\sim_{\mathcal{N}}}, R_{\leq_{\mathcal{N}}}, \vec{V})_{i \in \mathcal{N}}$ be the unravelling of K^Ω around some $w \in W^\Omega$. Let $M = (\vec{W}, \sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}}, \vec{V})_{i \in \mathcal{N}}$ be the generated common prior model. Let $f : M \to K^\Omega$ be a mapping such that for all histories $\bar{h} \in \vec{W} : f(\bar{h}) := \mathit{last}(\bar{h})$. f is a bounded morphism.

Proof. First, notice that for all $p \in P$, it is the case that $\vec{V}(p) = \{\bar{h} \in \vec{W} \mid last(\bar{h}) \in V^{\Omega}(p)\} = \{\bar{h} \in \tilde{W} \mid f(\bar{h}) \in V^{\Omega}(p)\}$. Thus, for all $\bar{h} \in \vec{W}$, it holds that $\bar{h}$ and $f(\bar{h})$ satisfy the same propositional letters. Furthermore, the back condition can be easily proven. Let $\bar{h} \in \vec{W}$, $w \in W^{\Omega}$ and $\preceq \in \{\sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}} \mid i \in \mathcal{N}\}$. Suppose $f(\bar{h}) \preceq^{\Omega} w$, i.e. $last(\bar{h}) \preceq^{\Omega} w$. Define $\bar{h}' := \bar{h} + (last(h), \preceq^{\Omega}, w)$. Obviously, $\bar{h}' \in \vec{W}$ and $f(\bar{h}') = last(\bar{h}') = w$. By definition of $R_{\preceq}$, we get $\bar{h} R_{\preceq} \bar{h}'$. Furthermore, since $R_{\preceq} \subseteq \preceq$, it follows that $\bar{h} \preceq \bar{h}'$. Hence the back condition is satisfied.

The forth condition requires more work. Let $\preceq \in \{\sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}} \mid i \in \mathcal{N}\}$ and $\bar{h}, \bar{h}' \in \vec{W}$ such that $\bar{h} \preceq \bar{h}'$. We need to show that $f(\bar{h}) \preceq^{\Omega} f(\bar{h}')$. We prove all cases separately:

- Suppose $\preceq = \leq_{\mathcal{N}}$. From the definition of $\leq_{\mathcal{N}}$ and Observation 2.59, it follows that there exists an $\mathcal{R}$-path from $\bar{h}$ to $\bar{h}'$ in $\vec{K}$ with $\mathcal{R} = \{R_{\leq_{\mathcal{N}}}\}$. According to Lemma 2.62, this path contains the non-redundant $\mathcal{R}$-path $\bar{p} = (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$ from $\bar{h}$ to $\bar{h}'$. Unfolding the definition of $R_{\leq_{\mathcal{N}}}$, we get $last(h_0) \leq^{\Omega}_{\mathcal{N}} ... \leq^{\Omega}_{\mathcal{N}} last(h_n)$, i.e. $f(\bar{h}) \leq^{\Omega}_{\mathcal{N}} ... \leq^{\Omega}_{\mathcal{N}} f(\bar{h}')$. The transitivity of $\leq^{\Omega}_{\mathcal{N}}$ gives us that $f(\bar{h}) \leq^{\Omega}_{\mathcal{N}} f(\bar{h}')$.

- Suppose $\preceq = \leq_i$ for some $i \in \mathcal{N}$. From the definition of $\leq_i$ and Observation 2.59, it follows that there exists an $\mathcal{R}$-path from $\bar{h}$ to $\bar{h}'$ in $\vec{K}$ with $\mathcal{R} = \{R_{\leq_i}, R_{\leq_{\mathcal{N}}}\}$. According to Lemma 2.62, this path contains the non-redundant $\mathcal{R}$-path $\bar{p} = (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$ from $\bar{h}$ to $\bar{h}'$. Using the definition of $R_{\leq_i}$ and $R_{\leq_{\mathcal{N}}}$, we get that $last(\bar{h}_0) \triangleleft ... \triangleleft last(\bar{h}_n)$, i.e. $f(\bar{h}) \triangleleft ... \triangleleft f(\bar{h}')$ with $\triangleleft \in \{\leq^{\Omega}_i, \leq^{\Omega}_{\mathcal{N}}\}$. Since K^{Ω} is a pseudo-model, it is the case that $\leq^{\Omega}_{\mathcal{N}} \subseteq \leq^{\Omega}_i$. Thus, $f(\bar{h}) \leq^{\Omega}_i ... \leq^{\Omega}_i f(\bar{h}')$. By transitivity of $\leq^{\Omega}_i$, it follows that $f(\bar{h}) \leq^{\Omega}_i f(\bar{h}')$.

- Suppose $\preceq = \sim_{\mathcal{N}}$. From the definition of $\sim_{\mathcal{N}}$ and Observations 2.51 and 2.59, it follows that there exists an $\mathcal{R}$-path from $\bar{h}$ to $\bar{h}'$ in $\vec{K}$ with $\mathcal{R} = \{R_{\sim_{\mathcal{N}}}, R^{-1}_{\sim_{\mathcal{N}}}, R_{\leq_{\mathcal{N}}}, R^{-1}_{\leq_{\mathcal{N}}}\}$. According to Lemma 2.62, this path contains the non-redundant $\mathcal{R}$-path $\bar{p} = (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$ from $\bar{h}$ to $\bar{h}'$. That is, for al $k < n$ we have $(\bar{h}_k, \bar{h}_{k+1}) \in (R_{\sim_{\mathcal{N}}} \cup R_{\leq_{\mathcal{N}}})$ or $(\bar{h}_{k+1}, \bar{h}_k) \in (R_{\sim_{\mathcal{N}}} \cup R_{\leq_{\mathcal{N}}})$. Using the definitions of $R_{\sim_{\mathcal{N}}}$ and $R_{\leq_{\mathcal{N}}}$, it follows that for all $k < n$ we have $(last(\bar{h}_k), last(\bar{h}_{k+1})) = (f(\bar{h}_k), f(\bar{h}_{k+1})) \in (\sim^{\Omega}_{\mathcal{N}} \cup \leq^{\Omega}_{\mathcal{N}})$ or $(last(\bar{h}_{k+1}), last(\bar{h}_k)) = (f(\bar{h}_{k+1}), f(\bar{h}_k)) \in (\sim^{\Omega}_{\mathcal{N}} \cup \leq^{\Omega}_{\mathcal{N}})$. K^{Ω} is a pseudo-model. Thus, $\leq^{\Omega}_{\mathcal{N}} \subseteq \sim^{\Omega}_{\mathcal{N}}$. Consequently, for all $k < n$ it holds that $(f(\bar{h}_k), f(\bar{h}_{k+1})) \in \sim^{\Omega}_{\mathcal{N}}$ or $(f(\bar{h}_{k+1}), f(\bar{h}_k)) \in \sim^{\Omega}_{\mathcal{N}}$. Moreover, since $\sim^{\Omega}_{\mathcal{N}}$ is an equivalence relation, it is symmetric. Consequently, for all $k < n$ it holds that $(f(\bar{h}_k), f(\bar{h}_{k+1})) \in \sim^{\Omega}_{\mathcal{N}}$. By transitivity, it follows that $f(\bar{h}) = f(\bar{h}_0) \sim^{\Omega}_{\mathcal{N}} f(\bar{h}_n) = f(\bar{h}')$.

- Suppose $\preceq = \sim_i$ for some $i \in \mathcal{N}$. From the definition of $\sim_i$ and Observations 2.51 and 2.59, it follows that there exists an $\mathcal{R}$-path from $\bar{h}$ to $\bar{h}'$ in $\vec{K}$ with $\mathcal{R} = \{R_{\sim_i}, R^{-1}_{\sim_i}, R_{\leq_i}, R^{-1}_{\leq_i}, R_{\sim_{\mathcal{N}}}, R^{-1}_{\sim_{\mathcal{N}}}, R_{\leq_{\mathcal{N}}}, R^{-1}_{\leq_{\mathcal{N}}}\}$. According to Lemma 2.62, this path contains the non-redundant $\mathcal{R}$-path $\bar{p} = (\bar{h}_0, R_0, \bar{h}_1, ..., R_{n-1}, \bar{h}_n)$ from $\bar{h}$ to $\bar{h}'$. That is, for al $k < n$ we have $(\bar{h}_k, \bar{h}_{k+1}) \in (R_{\sim_i} \cup R_{\leq_i} \cup R_{\sim_{\mathcal{N}}} \cup R_{\leq_{\mathcal{N}}})$ or $(\bar{h}_{k+1}, \bar{h}_k) \in (R_{\sim_i} \cup R_{\leq_i} \cup R_{\sim_{\mathcal{N}}} \cup R_{\leq_{\mathcal{N}}})$. Using the definitions of the relations in $\mathcal{R}$, it follows that for all $k < n$ we have $(last(\bar{h}_k), last(\bar{h}_{k+1})) = (f(\bar{h}_k), f(\bar{h}_{k+1})) \in (\sim^{\Omega}_i \cup \leq^{\Omega}_i \cup \sim^{\Omega}_{\mathcal{N}} \cup \leq^{\Omega}_{\mathcal{N}})$ or $(last(\bar{h}_{k+1}), last(\bar{h}_k)) = (f(\bar{h}_{k+1}), f(\bar{h}_k)) \in (\sim^{\Omega}_i \cup \leq^{\Omega}_i \cup \sim^{\Omega}_{\mathcal{N}} \cup \leq^{\Omega}_{\mathcal{N}})$. K^{Ω} is a pseudo-model. Consequently, $\leq^{\Omega}_{\mathcal{N}} \subseteq \leq^{\Omega}_i \subseteq \sim^{\Omega}_i$ and $\sim^{\Omega}_{\mathcal{N}} \subseteq \sim^{\Omega}_i$. Moreover, $\sim^{\Omega}_i$ is symmetric, as it is an equivalence relation. Consequently, for all $k < m$ it holds that $(f(\bar{h}_k), f(\bar{h}_{k+1})) \in \sim^{\Omega}_{\mathcal{N}}$. By transitivity, it follows that $f(\bar{h}) = f(\bar{h}_0) \sim^{\Omega}_{\mathcal{N}} f(\bar{h}_n) = f(\bar{h}')$.

□

Since there exists a bounded morphism between the common prior model M and the canonical pseudo-model K^{Ω}, we get the following result as a corollary:

Corollary 2.68. K^{Ω} and M satisfy the same EDL-formulas.

Proof. The proof follows from Lemma's 2.41 and 2.67. □

Conclusion

After going through these steps, we can show that Λ_{EDL} is weakly complete with respect to the class of all common prior models:

Theorem 2.69. Λ_{EDL} is weakly complete with respect to the class of all common prior models.

Proof. According to Proposition 2.38, it suffices to show that any consistent EDL-formula is satisfiable on some common prior model. Let φ be an consistent EDL-formula. By Lindenbaum's Lemma, $\{\varphi\}$ can be extended to a maximal consistent set Φ. By definition of the canonical pseudo-model, $\Phi \in W^{\Omega}$. Let $w := \Phi$. According to the truth lemma, it follows from $\varphi \in w$ that $K^{\Omega}, w \models \varphi$. Let $\vec{K}$ be the unravelling of K^{Ω} around w and let $M = (\vec{W}, \sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}}, \vec{V})_{i \in \mathcal{N}}$ be the generated common prior model. Notice that the history $(w) \in W$. Let $f : M \to K^{\Omega}$ be such that $f(\bar{h}) := \mathit{last}(\bar{h})$ for all $\bar{h} \in \vec{W}$. As was shown in Lemma 2.67, this is a bounded morphism. From Lemma 2.41, it follows that $M, (w) \models \varphi$ iff $K^{\Omega}, w \models \varphi$. Thus, $M, (w) \models \varphi$. Hence φ is satisfiable on a common prior model. Thus, Λ_{EDL} is weakly complete with respect to the class of all common prior models. □

Every common prior model is a model for EDL. Since the language of EDL cannot distinguish between models that do and models that do not satisfy the common prior assumption, we get completeness for all EDL-models as a corollary:

Corollary 2.70. Λ_{EDL} is weakly complete with respect to the class of all EDL-models.

Chapter 3

Communication as Information Sharing

People communicate with one another continuously and in a variety of ways. Due to communication, people might learn new things, change their beliefs or come into contact with new perspectives. All of these are interesting from the perspective of deliberative democracy. This chapter introduces a model transformer for the multi-agent plausibility models from the previous chapter, which mimics a communicative situation where all the agents share all their hard information. This is interesting because distributed knowledge and, in special circumstances, distributed belief can be realised if the agents share their knowledge. Futhermore, this model transformer is used in Chapter 5 as a basis for the model transformer for democratic deliberation.

Section 3.1 introduces several types of model transformers from the dynamic epistemic logic literature, including the one we just mentioned and which we will call "deliberation". Section 3.2 proves some interesting results about this model transformer. In Section 3.3, we formally extend EDL with a dynamic operator corresponding to "deliberation" and obtain the epistemic doxastic logic with deliberation (EDLD), of which we prove soundness and completeness.

3.1 Public Announcement, Tell All You Know and Deliberation

Kripke models in general, and multi-agent plausibility models in particular, consist of a set of possible worlds. Each agent is endowed with a partition of these possible worlds in information cells. An agent's hard information at a designated world w is composed of everything that is true throughout the information cell containing w, i.e. of everything that is true in all the worlds the agent considers to be epistemically possible. One of the first ideas to model updates with hard information comes from public announcement logic, which was developed by Plaza (1989), Gerbrandy & Groeneveld (1997), Baltag et al. (1998) and Van Ditmarsch (2000). The idea of public announcement logic is that if somebody publicly announces a piece of hard information φ, all the worlds that do not satisfy φ become epistemically impossible for all of the agents. Hence, they can be deleted from the model:

Definition 3.1. Let $\mathcal{N}$ be a finite set of agents, $M = (W, \sim_i, \leq_i, V)_{i\in\mathcal{N}}$ a multi-agent plausibility model and $\varphi \in \mathcal{L}_{EDL}$. The *updated model after a public announcement of* φ is defined as $M^{!\varphi} := (W^{!\varphi}, \sim_i^{!\varphi}, \leq_i^{!\varphi}, V^{!\varphi})_{i\in\mathcal{N}}$ where $W^{!\varphi} := \{w \in W \mid M, w \models \varphi\}$, $\sim_i^{!\varphi} := \sim_i \cap (W^{!\varphi} \times W^{!\varphi})$, $\leq_i^{!\varphi} := \leq_i \cap (W^{!\varphi} \times W^{!\varphi})$ and $V^{!\varphi}(p) = V(p) \cap W^{!\varphi}$ for all $p \in P$.

This model transformer is denoted by $[!\varphi]$. The idea is that after somebody publicly announces a piece of hard information φ, all the worlds that do not satisfy φ are removed from the model. This way of transforming the model is syntactical in nature, because the updated model is dependent on the specific formula φ and, hence, on the valuation function V.

One can also consider a situation where an agent does not publicly announce a specific formula φ, but everything that he knows. This idea was proposed by Baltag, who interprets this situation in a

language independent manner.[1] If an agent tells all he knows at a designated world w, he publically announces all the worlds that he cannot distinguish from w on the basis of his hard information. Or equivalently, he announces all the worlds that are epistemically impossible for him. That is, the agent shares his epistemic indistinguishability relation:

Definition 3.2. Let $\mathcal{N}$ be a finite set of agents, $M = (W, \sim_i, \leq_i, V)_{i\in\mathcal{N}}$ a multi-agent plausibility model and $j \in \mathcal{N}$. The *updated model after j tells everything he knows* is defined as $M^{!j} = (W^{!j}, \sim_i^{!j}, \leq_i^{!j}, V^{!j})_{i\in\mathcal{N}}$ where $W^{!j} := W$, $\sim_i^{!j} := \sim_i \cap \sim_j$, $\leq_i^{!j} := \leq_i \cap \sim_i^{!j} = \leq_i \cap \sim_j$ and $V^{!j} := V$.

This model transformer is denoted by $[!j]$. One might object that it is strange to interpret $[!j]$ in a language independent manner. If an agent knows something that is not expressible in language, how can he ever communicate this piece of information to the other agents? It is important, however, to make a distinction between natural language and the formal language of EDL. Obviously, natural language is much richer than any formal language. The formal language used so far is $\mathcal{L}_{EDL}$. However, we by no means claim that this is the language the agents use. Thus, interpreting "tell all you know" in a language independent manner solely means that this model transformer is independent of the formal language we use. What we do assume is that, in reality, agents are able to communicate all they know to one another in some way. Although this assumption is idealising, it is not problematic as the goal of our thesis is the development of a formal framework for an ideal situation of democratic deliberation.

One can even go a level higher and introduce a model transformer that models the situation in which all the agents share all their knowledge:

Definition 3.3. Let $\mathcal{N}$ be a finite set of agents and $M = (W, \sim_i, \leq_i, V)_{i\in\mathcal{N}}$ a multi-agent plausibility model. The *updated model after deliberation* is defined as $M^! = (W^!, \sim_i^!, \leq_i^!, V^!)_{i\in\mathcal{N}}$ where $W^! := W$, $\sim_i^! := \bigcap_{j\in\mathcal{N}} \sim_j$, $\leq_i^! := \leq_i \cap \sim_i^!$ and $V^! = V$.

This model transformer is denoted by $[!]$ and called "deliberation", as the model transformer used in Chapter 5 to model deliberation is an adaptation of this one. Notice that $[!]$ models the entire communication process in one go. We can show that deliberation indeed amounts to letting all the agents tell all they know:

Proposition 3.4. Let $\mathcal{N} = \{1, ..., n\}$ be a finite set of agents and $f : \mathcal{N} \to \mathcal{N}$ be a permutation among the agents. Let $M = (W, \sim_i, \leq_i, V)_{i\in\mathcal{N}}$ be a multi-agent plausibility model. Then $M^! = (M^{!f(1)})...)^{!f(n)}$.

Proof. According to the definitions of deliberation and tell all you know, in order to show that $M^! = (M^{!f(1)})...)^{!f(n)}$, it suffices to prove that $\sim_i^! = (\sim_i^{!f(1)})...)^{!f(n)}$. This follows from the fact that $\sim_i^! = \bigcap_{j\in\mathcal{N}} \sim_j = \sim_1 \cap ... \cap \sim_n = \sim_i \cap \sim_{f(1)} \cap ... \cap \sim_{f(n)} = (\sim_i^{!f(1)})...)^{!f(n)}$.[2] □

The rest of this chapter focuses on the model transformer for deliberation, which models a situation in which all the agents share all their information. This model transformer is interesting for philosophical as well as for technical reasons. In Chapter 5, we argue why (an adaptation of) this model transformer is a nice formal tool for modelling democratic deliberation. The main reason is already clear at this point. In a democratic deliberation, everyone has a say. Therefore, we can model deliberation as a situation in which all the agents speak up. This can be modelled by using the model transformer $[!j]$ and letting all the agents speak in turn. However, Proposition 3.4 shows that $[!]$ yields the same result. For matters of technical simplicity, we therefore prefer $[!]$.

[1] Baltag proposed this idea in the course *Topics in Dynamic Epistemic Logic*, taught at ILLC in spring 2013.

[2] The $(M^{!f(1)})...)^{!f(n)}$ represents the protocol that all agents speak in turn and essentially amounts to the algorithm presented in Van Benthem (2002).

3.2 Realising Distributed Knowledge and Belief

This section focuses on some interesting results about the model transformer [!]. More specifically, we prove that [!] realises distributed knowledge and that, in certain circumstances, it also realises distributed plausibility and hence distributed belief.

Recall that distributed knowledge is the knowledge a group of agents would have if they were to share all their hard information. Since [!] models precisely this, it should be the case that all agents have the same knowledge after deliberation and that this knowledge equals the distributed knowledge before any communication took place. This is indeed the case:

Proposition 3.5. Let $\mathcal{N}$ be a finite set of agents, M a multi-agent plausibility model and $M^!$ the updated model after deliberation. Then for all $i \in \mathcal{N} : \sim_i^! = \sim_{\mathcal{N}}$.

Proof. By definition of $\sim_i^!$. □

As a corollary, we get:

Corollary 3.6. Let $\mathcal{N}$ be a finite set of agents, M a multi-agent plausibility model and $M^!$ the updated model after deliberation. Then for all $i, j \in \mathcal{N}$ and all $P \subseteq W = W^!$, the following hold:

1. $K_i^{F^!} P = K_j^{F^!} P$.
2. $D_K^F P = K_i^{F^!} P$.

Corollary 3.6 shows that deliberation turns distributed knowledge into actual knowledge. Therefore, we can really think of distributed knowledge as the potential knowledge of a group of agents. After sharing all their hard information, the knowledge of the agents is the same. Their plausibility relations and their beliefs, however, might still differ. The deliberation operation [!] does not, in general, realise distributed plausibility and distributed belief. However, as we anticipated in Section 2.3, it does in common prior models. First, consider the updated plausibility relations:

Proposition 3.7. Let $\mathcal{N}$ be a finite set of agents, M a common prior model, $\leq$ the common prior plausibility relation and $M^!$ the updated model after deliberation. Then for all $i \in \mathcal{N} : \leq_i^! = \bigcap_{i \in \mathcal{N}} \leq_i = \leq_{\mathcal{N}}$.

Proof. Let $i \in \mathcal{N}$ be arbitrary. Unfolding the definitions yields $\leq_i^! = \leq_i \cap \sim_i^! = (\leq \cap \sim_i) \cap \bigcap_{i \in \mathcal{N}} \sim_i = \leq \cap \bigcap_{i \in \mathcal{N}} \sim_i = \bigcap_{i \in \mathcal{N}} (\leq \cap \sim_i) = \bigcap_{i \in \mathcal{N}} \leq_i = \leq_{\mathcal{N}}$. □

The interpretation of this proposition is that if the differences in the plausibility of the agents are solely due to the fact that they have learned and experienced different things in their lives, i.e. solely due to differences in information, then sharing information yields agreement with respect to their plausibilities. Or more formally, [!] realises distributed plausibility in common prior models, as is reflected in the following corollary:

Corollary 3.8. Let $\mathcal{N}$ be a finite set of agents, M a multi-agent plausibility model and $M^!$ the updated model after deliberation. For all $i, j \in \mathcal{N}$ and all $P \subseteq W = W^!$ the following hold:

1. $\Box_i^{F^!} P = \Box_j^{F^!} P$.
2. $D_\Box^F P = \Box_i^{F^!} P$.

Recall that distributed belief is defined in terms of distributed knowledge and distributed plausibility. Both are realised in common prior models. Therefore, distributed belief is realised in these models as well:

Theorem 3.9. (Agreeing to Disagree) Let $\mathcal{N}$ be a finite set of agents, M a common prior model and $M^!$ the updated model after deliberation. Then for all $i, j \in \mathcal{N}$ and all $P \subseteq W = W^!$, the following hold:

1. $B_i^{F^!} P = B_j^{F^!} P$.
2. $D_B^F P = B_i^{F^!} P$.

Proof. By Observation 2.10, it follows that $B_i^{F^!} P$ iff for all w' such that $w \sim_i^! w'$ there exists a w'' such that $w' \leq_i^! w''$ and for all w''' such that $w'' \leq_i^! w'''$ it holds that $w''' \in P$. According to Propositions 3.5 and 3.7, it holds for all $i \in \mathcal{N}$ that $\sim_i^! = \sim_{\mathcal{N}}$ and $\leq_i^! = \leq_{\mathcal{N}}$. Item 2 now follows directly from Observation 2.26. Item 1 follows from the fact that all agents have the same epistemic indistinguishability and plausibility relations. □

The result of Theorem 3.9.1 is a qualitative version of Aumann's (1976) *agreeing to disagree* theorem, which says that if differences in the beliefs of the agents are solely due to differences in information, then sharing of information leads them to agree on their beliefs.[3] Thus, in common prior models distributed belief can really be interpreted as form of potential group belief.

In standard common prior models, it is even more clear that distributed belief is a form of potential group belief, as one can show that the agents' belief relations collapse after deliberation. Before proving this, we define the common belief relation and show that this relation can be thought of as the accessibility relation for distributed belief. Furthermore, we define the updated belief relations:

Definition 3.10. Let $\mathcal{N}$ be a finite set of agents and M a standard common prior model. Let the *common belief relation* $\rightarrow_{\mathcal{N}} \subseteq W \times W$ be such that for all $w, w' \in W$: $w \rightarrow_{\mathcal{N}} w'$ iff $w \sim_{\mathcal{N}} w'$ and $w' \in \mathrm{Max}(\leq_{\mathcal{N}})$.

Proposition 3.11. Let $\mathcal{N}$ be a finite set of agents, M a standard common prior model and $P \subseteq W$. Then $D_B P = [\rightarrow_{\mathcal{N}}]P$.

Proof. Left to the reader. The proof is similar to the proof of Theorem 2.19. □

Definition 3.12. Let $\mathcal{N}$ be a finite set of agents, M a standard common prior model and $M^!$ the updated model after deliberation. Let the *updated belief relation* $\rightarrow_i^! \subseteq W^! \times W^!$ be such that for all $w, w' \in W^!$: $w \rightarrow_i^! w'$ iff $w \sim_i^! w'$ and $w' \in \mathrm{Max}(\leq_i^!)$.

The next proposition shows that the belief relations of all the individuals after deliberation equal the distributed belief relation:

Proposition 3.13. (Agreeing to Disagree) Let $\mathcal{N}$ be a finite set of agents, M a standard common prior model and $M^!$ the updated model after deliberation. Then for all $i \in \mathcal{N}$ the following holds: $\rightarrow_i^! = \rightarrow_{\mathcal{N}}$.

Proof. Let $i \in \mathcal{N}$ be arbitrary. Recall that $w \rightarrow_{\mathcal{N}} w'$ iff $w \sim_{\mathcal{N}} w'$ and $w' \in \mathrm{Max}(\leq_{\mathcal{N}})$. Propositions 3.5 and 3.7 give us that $\sim_i^! = \sim_{\mathcal{N}}$ and $\leq_i^! = \leq_{\mathcal{N}}$. By definition of $\rightarrow_i^!$, it follows that $\rightarrow_i^! = \rightarrow_{\mathcal{N}}$. □

It is immediately clear from this proposition that distributed belief is a form of potential group belief, because after the agents share all their information in a deliberation, they end up with the same belief relation and hence with the same beliefs. Thus, Proposition 3.13 is a qualitative verion of Aumann's (1976) *agreeing to disagree* theorem in standard models.

[3] Dégremont & Roy (2012) were one of the first to introduce a qualitative version of Aumann's theorem to dynamic epistemic logic.

3.3 Epistemic Doxastic Logic with Deliberation

The language of epistemic doxastic logic with deliberation (EDLD) is obtained by extending the language of EDL with the dynamic modality [!]:

Definition 3.14. Let P be a finite set of propositional letters and $\mathcal{N}$ a finite set of agents. The set $\mathcal{L}_{EDLD}$ of formulas of φ of EDLD is defined recursively:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid \Box_i\varphi \mid D_K\varphi \mid D_\Box\varphi \mid [!]\varphi$$

where $p \in P$ and $i \in \mathcal{N}$. We define $\bot := p \wedge \neg p$ and $\top := \neg\bot$. The Boolean connectives $\vee$ and $\rightarrow$ are defined in terms of $\neg$ and $\wedge$ in the standard manner. The duals of the modal operators are defined in the following way: $\hat{K}_i := \neg K_i \neg$ and $\Diamond_i := \neg\Box_i\neg$. Belief is defined as $B_i := K_i\Diamond_i\Box_i$ and distributed belief as $D_B := D_K D_\Diamond D_\Box$, where $D_\Diamond := \neg D_\Box \neg$.

The intended meaning of $[!]\varphi$ is that after deliberation, i.e. after all the agents have shared all their information, φ is the case. The models of EDLD are common prior models:

Definition 3.15. A tuple $M = (W, \sim_i, \leq_i, V)_{i\in\mathcal{N}}$ is a model for EDLD iff $M = (W, \sim_i, \leq_i, V)_{i\in\mathcal{N}}$ is a common prior model for EDL.[4]

As we shall see in a moment, the common prior assumption is needed for proving completeness. The truth conditions for EDLD are those of EDL, extended with one for formulas of the form $[!]\varphi$:

Definition 3.16. Let $M = (W, \sim_i, \leq_i, V)_{i\in\mathcal{N}}$ be a model for EDLD, $w \in W$ and $i \in \mathcal{N}$. The *satisfaction* relation $\models$ between pairs (M, w) and formulas $\varphi \in \mathcal{L}_{EDLD}$ is defined as in Definition 2.31, extended with the following clause:

- $M, w \models [!]\varphi$ iff $M^!, w \models \varphi$.

A formula φ is *valid* if for all $M = (W, \sim_i, \leq_i, V)_{i\in\mathcal{N}}$ and all $w \in W$, we have $M, w \models \varphi$. We denote validity by $\models \varphi$.

In order to prove completeness of EDLD, we add the following reduction axioms to the proof system of EDL:

Definition 3.17. The proof system of EDLD (notation: Λ_{EDLD}) is the proof system Λ_{EDL} extended with the following *reduction axioms*:

1. $[!]p \leftrightarrow p$
2. $[!]\neg\varphi \leftrightarrow \neg[!]\varphi$
3. $[!](\varphi_1 \wedge \varphi_2) \leftrightarrow ([!]\varphi_1 \wedge [!]\varphi_2)$
4. $[!]K_i\varphi \leftrightarrow D_K[!]\varphi$
5. $[!]\Box_i\varphi \leftrightarrow D_\Box[!]\varphi$
6. $[!]D_K\varphi \leftrightarrow D_K[!]\varphi$
7. $[!]D_\Box\varphi \leftrightarrow D_\Box[!]\varphi$
8. $[!][!]\varphi \leftrightarrow [!]\varphi$

[4] See Definition 2.20 for the definition of a common prior frame and Definition 2.30 for the definition of a common prior model.

The following theorem shows that these reduction axioms are sound on common prior models:

Theorem 3.18. Λ_{EDLD} is sound with respect to the class of all common prior models.

Proof. The proof system of EDLD is the one of EDL extended with the reduction axioms. According to Corollary 2.36, Λ_{EDL} is sound with respect to the class of all common prior models. Hence it suffices to show that the reduction axioms are sound on common prior models. Let $M = (W, \sim_i, \leq_i, V)_{i \in \mathcal{N}}$ be a common prior model, $w \in W$ and $i \in \mathcal{N}$.

The soundness of reduction axiom 1 follows from the fact that [!] does not change ontic facts about the world. The soundness of the Boolean reduction axioms 2 and 3 can be proven by unfolding the definitions. The same holds for reduction axiom 8. In order to prove the soundness of 4 and 5, we need the soundness of reduction axioms 6 and 7.

For the left-to-right-direction of reduction axiom 6, suppose $M, w \models [!]D_K\varphi$. Let $w' \in W$ be such that $w \sim_{\mathcal{N}} w'$. From the definition of $M, w \models [!]D_K\varphi$, we get that $M^!, w \models D_K\varphi$. Notice that $\sim^!_{\mathcal{N}} = \bigcap_{i \in \mathcal{N}} \sim^!_i = \sim_{\mathcal{N}}$. Therefore, it follows that $w \sim^!_{\mathcal{N}} w'$ and, thus, that $M^!, w' \models \varphi$. By definition we get that $M, w' \models [!]\varphi$ and, consequently, that $M, w \models D_K[!]\varphi$. For the right-to-left direction, suppose $M, w \models D_K[!]\varphi$. In order to show that $M, w \models [!]D_K\varphi$, we need to show that $M^!, w \models D_K\varphi$. Let $w' \in W$ be such that $w \sim^!_{\mathcal{N}} w'$. Because $\sim^!_{\mathcal{N}} = \sim_{\mathcal{N}}$, it is the case that $w \sim_{\mathcal{N}} w'$. Hence it follows from $M, w \models D_K[!]\varphi$ that $M, w' \models [!]\varphi$. Hence $M^!, w' \models \varphi$. Consequently, $M^!, w \models D_K\varphi$ and, thus, $M, w \models [!]D_K\varphi$. In conclusion, reduction axiom 6 is sound. The soundness of reduction axiom 7 can be proven in a similar manner, but uses the fact that $\leq^!_{\mathcal{N}} = \bigcap_{i \in \mathcal{N}} \leq^!_i = \bigcap_{i \in \mathcal{N}} (\leq_i \cap \sim^!_i) = \sim_{\mathcal{N}} \cap \bigcap_{i \in \mathcal{N}} \leq_i = \sim_{\mathcal{N}} \cap \leq_{\mathcal{N}} = \leq_{\mathcal{N}}$.

For axiom 4, we get $M, w \models [!]K_i\varphi$ iff $M^!, w \models K_i\varphi$ by definition. Since $\sim^!_i = \sim_{\mathcal{N}} = \sim^!_{\mathcal{N}}$, it follows that $M^!, w \models K_i\varphi$ iff $M^!, w \models D_K\varphi$ iff $M, w \models [!]D_K\varphi$. By reduction axiom 6, we get that this holds iff $M, w \models D_K[!]\varphi$. Thus, $M, w \models [!]K_i\varphi$ iff $M, w \models D_K[!]\varphi$. The proof of the soundness of reduction axiom 5 is similar to that of 4, but uses a fact that only holds in common prior models, namely that $\leq^!_i = \leq_{\mathcal{N}} = \leq^!_{\mathcal{N}}$ for all $i \in \mathcal{N}$, as was shown in Proposition 3.7. □

The proof of the soundness of reduction axioms 1-4 and 6-8 did *not* make use of the common prior assumption. Thus, these axioms are sound on all multi-agent plausibility models. For the soundness of $[!]\Box_i\varphi \leftrightarrow D_\Box[!]\varphi$, however, we appeal to the fact that the updated plausibility relations of all agents are the same and equal to the intersection of the initial plausibility relations. This is typical for common prior models. Without the common prior assumption, we cannot reduce dynamic formulas containing $\Box_i$ to static formulas in the language of EDL. For proving completeness, however, it is important that we are able to do so. Before moving to completeness, we need to show that EDL and EDLD are co-expressive. In order to do this, we start by defining a map from $\mathcal{L}_{EDLD}$ to $\mathcal{L}_{EDL}$:

Definition 3.19. We define a map $! : \mathcal{L}_{EDLD} \to \mathcal{L}_{EDL}$ by recursion on the complexity of φ:

- $!(p) := p$,
- $!(\neg\varphi) := \neg!(\varphi)$,
- $!(\varphi \wedge \psi) :=!(\varphi) \wedge !(\psi)$,
- $!(K_i\varphi) := D_K!(\varphi)$,
- $!(\Box_i\varphi) := D_\Box!(\varphi)$,
- $!(D_K\varphi) := D_K!(\varphi)$,
- $!(D_\Box\varphi) := D_\Box!(\varphi)$,
- $!([!]\varphi) :=!\varphi$.

Observation 3.20. The map $! : \mathcal{L}_{EDLD} \to \mathcal{L}_{EDL}$ is well-defined.

One can use the reduction axioms to prove the following:

Proposition 3.21. For all formulas $\varphi \in \mathcal{L}_{EDLD}$, the following holds: $\vdash_{EDLD} [!]\varphi \leftrightarrow !(\varphi)$.

Proof. The proof is by induction on the complexity of φ. □

We now define a function which translates formulas in $\mathcal{L}_{EDLD}$ to ones in $\mathcal{L}_{EDL}$:

Definition 3.22. Let $tr : \mathcal{L}_{EDLD} \to \mathcal{L}_{EDL}$ be the *translation function* defined by recursion on the complexity of φ:

- $tr(p) := p$,
- $tr(\neg\varphi) := \neg tr(\varphi)$,
- $tr(\varphi \wedge \psi) := tr(\varphi) \wedge tr(\psi)$,
- $tr(K_i\varphi) := D_K tr(\varphi)$,
- $tr(\Box_i\varphi) := D_\Box tr(\varphi)$,
- $tr(D_K\varphi) := D_K tr(\varphi)$,
- $tr(D_\Box\varphi) := D_\Box tr(\varphi)$,
- $tr([!]\varphi) :=!\varphi$.

Observation 3.23. The translation function $tr : \mathcal{L}_{EDLD} \to \mathcal{L}_{EDL}$ is well-defined.

Notice that if a formula φ does not contain any dynamic modalities, then it equals its own translation:

Observation 3.24. For all $\varphi \in \mathcal{L}_{EDL}$, it holds that $tr(\varphi) = \varphi$.

We can now show EDLD proves that every formula $\varphi \in \mathcal{L}_{EDLD}$ is equivalent to a formula $\psi \in \mathcal{L}_{EDL}$, namely its translation:

Proposition 3.25. For all $\varphi \in \mathcal{L}_{EDLD}$, it holds that $\vdash_{EDLD} \varphi \leftrightarrow tr(\varphi)$.

Proof. The proof is by induction on the complexity of φ. The proof for the base case is trivial. For the inductive case, the Boolean cases and cases where $\varphi = [R]\psi$ with $[R] \in \{K_i, \Box_i, D_K, D_\Box \mid i \in \mathcal{N}\}$ can be proven using the induction hypothesis. For formulas of the form $[!]\varphi$, use Proposition 3.21. □

With this in place, we can show that EDL and EDLD are co-expressive:

Theorem 3.26. EDL and EDLD are co-expressive.

Proof. According to Proposition 3.25, it holds that for all $\varphi \in \mathcal{L}_{EDLD}$ there exists a $\psi \in \mathcal{L}_{EDL}$ such that $\vdash_{EDLD} \varphi \leftrightarrow \psi$, namely $\psi = tr(\varphi)$. By the soundness of Λ_{EDLD}, it follows that $\models_{EDLD} \varphi \leftrightarrow \psi$. Furthermore, it holds that for all $\varphi \in \mathcal{L}_{EDL}$ there exists a $\psi \in \mathcal{L}_{EDLD}$ such that $\vdash_{EDL} \varphi \leftrightarrow \psi$, namely $\psi = \varphi$. By the soundness Λ_{EDL}, it follows that $\models_{EDL} \varphi \leftrightarrow \psi$. In conclusion, EDL and EDLD are co-expressive. □

Using the co-expressivity of EDL and EDLD, one can easily that Λ_{EDLD} is complete:

Theorem 3.27. Λ_{EDLD} is weakly complete with respect to the class of all common prior models.

Proof. According to Proposition 2.38, it suffices to show that any consistent EDLD-formula is satisfiable on some common prior model. Let φ be a consistent EDLD-formula. According to Proposition 3.25, there exists an EDL-formula ψ such that $\vdash_{EDLD} \varphi \leftrightarrow \psi$. According to Proposition 2.38 and Theorem 2.69, there exists a common prior model, i.e. an EDLD-model, $M = (W, \sim_i, \leq_i, V)_{i \in \mathcal{N}}$ and a world $w \in W$ such that $M, w \models \psi$. Since $\models_{EDLD} \varphi \leftrightarrow \psi$, it follows that $M, w \models \varphi$. Hence Λ_{EDLD} is weakly complete with respect to the class of all common prior models. □

It is now clear that the language of EDL can easily be extended with [!] without messing up completeness. This is because the reduction axioms allow us to reduce dynamic formulas to formulas of a static logic, for which we already have completeness. If one wants reduction axioms for $\Box_i$ that are sound on all multi-agent plausibility models, the language of EDL has to be enriched with a modality that captures the plausibility of an agent conditional on the knowledge of the entire group.

Adding $[!\varphi]$ or $[!j]$ to the language of EDL leads to similar problems. If we do this, we have to extend the language with more than just the dynamic modality in order to get reduction laws. If one wants reduction laws for $[!\varphi]$, one has to introduce a modality for conditional plausibility or belief, like Baltag and Smets (2006a, 2006b, 2008) do. If one wants reduction laws for $[!j]$, we need distributed knowledge not only for the entire group but also for subgroups of agents. Even though all of this can be done, we do not go into the details here as we want to keep the formal framework for modelling democratic deliberation as simple as possible.

Chapter 4

Preferences and their Justifications

This thesis aims at developing a formal model for debates that are ideal from the perspective of deliberative democracy, a political theory that places deliberation at the centre of political decision making. Deliberation is a process based on mutual respect in which participants justify their views and preferences to one another on the basis of reasons that everyone can accept. Justification is essential. Views or preferences that are not justifiable count as purely personal rather than as moral positions and do not deserve a place in the political realm. From the perspective of deliberative democracy, therefore, preferences are a secondary notion. The justifications – in which deliberators appeal to reasons or values that are of fundamental importance to them – come first.

This chapter focuses on preferences and their justifications. Section 4.1 introduces the theory of preference formation developed by Dietrich and List (2013a, 2013b). Their theory is interesting for our purpose, because it derives preferences over alternatives from the properties the agent considers to be motivationally salient. In addition, it treats preferences as dynamic rather than static. In Section 4.2, the formal models for preference formation are introduced. These models are a combination of the multi-agent plausibility frames presented in Chapter 2, which can model different epistemic doxastic attitudes, and Dietrich and List's framework. This allows us to define an agent's preferences over options in terms of his knowledge and/or belief, his motivational state or perspective and the properties that hold of the alternatives. Section 4.3 discusses different ways of defining preference in our framework and considers the relationship between preference, knowledge and belief. Section 4.4 introduces two special types of preference formation models, which play a central role in Chapter 6 where we formally investigate the claim that deliberation provides a solution to Arrow's impossibility theorem. In Section 4.5, we introduce a formal language for the preference formation models. By doing so, we obtain the logic of preference formation (LPF), of which we prove soundness and completeness.

4.1 Introduction to Dietrich and List's Setting

In standard rational and social choice theory, the preferences of an agent over a set of alternatives are assumed to be both fixed and exogenously given. That is, preferences do not change and their origins cannot be accounted for. They are a static and inexplicable feature of the agent. One of the key characteristics of deliberative democracy, however, is that it views preferences as dynamic rather than static. Moreover, it stresses the importance of the justification, and hence of the underlying reasons, of the agent's preferences. Therefore, the work of Dietrich and List (2013a, 2013b), who have recently developed a formal framework that is able to account for preference formation and alteration, is particularly interesting. This section introduces the basics of their framework. In the next section, we combine their framework with the multi-agent plausibility models from Chapter 2 and obtain the formal models for preference formation.

The central idea of Dietrich and List (2013a: 106) is that "an agent's preferences over the relevant fundamental objects depend on the reasons that motivate him or her and may vary with changes in them". Let $\mathcal{X}$ be a finite set of alternatives and let $x\mathbf{R}y$ denote the fact that the agent weakly prefers alternative x over alternative y. The preferences over these alternatives are determined by the motivationally salient properties or reasons. In order to explain this concept, we introduce a finite set $\mathbb{P}$ of properties. Intuitively, one can think of a property as a feature that an alternative either does or does not have. Thus, each property induces a partition of $\mathcal{X}$ into a set of alternatives that do and a set of alternatives that do not have that property. For example, we can divide the political parties in the Netherlands into the ones that are in favour of a European constitution and the ones that are not. The fact that a certain party is in favour of a European constitution, can be a reason for someone to vote for that party. Therefore, instead of seeing $\mathbb{P}$ as a set of properties, we can also think of it as a set of reasons. In that case, we can interpret property P applying to x as the fact that P can be used as a reason for justifying one's preference for alternative x.

According to Dietrich and List (2013b: 616), "in forming his or her preferences, an agent focuses on some, but not necessarily all, properties of the alternatives". The properties the agent focuses on are called the *motivationally salient* properties. Let $S \subseteq \mathbb{P}$ be the set of properties that are motivationally salient for the agent. This set depends both on the context and the psychological state of the agent. For instance, in determining which energy policies to fund, a member of the conservative party might not have given any thought to sustainability at all. Until he visits Greenland and the local residents show him the impact the climate change has on their lives. The motivational state of the agent changed due to a change in his psychological state. For an example of a change in an agent's motivational state due to a change in context, consider a politician who has to decide which anti-cancer drugs should be publically funded. In normal circumstances he may not care whether his decision affects his popularity among the electorate, but in election times he might. Let $\mathcal{S} \subseteq \mathcal{P}(\mathbb{P})$ be the set of all possible motivational states. Notice that $\mathcal{S}$ does not have to contain all subsets of $\mathbb{P}$, because it might well be the case that some property P is salient iff some other property P' is as well.

The preferences of the agent over the alternatives are determined by his motivational state and by his preferences over property packages. For any two alternatives $x, y \in \mathcal{X}$ and any motivational state $S \in \mathcal{S}$,

$$x\mathbf{R}^S y \text{ iff } \{P \in S \mid P \text{ holds of } x\} \succeq \{P \in S \mid P \text{ holds of } y\}$$

where $\succeq$ is a binary weighing relation on property packages. Notice that if the weighing relation is a linear order, then the induced preference relation over the alternatives is a total pre-order. The weighing relation over the property packages can be thought of as a preference relation of a more fundamental, stable kind. In other words, this weighing function represents the meta-level preferences of the agent.

To illustrate the idea of this framework, consider the following example:

Example. Suppose a local government has to decide whether to build a museum or a cinema in the city centre. The inhabitants of the city prefer a cinema, but the museum will create jobs for the local residents and attract tourists. For simplicity, suppose these are the only properties that matter. Hence $\mathbb{P} = \{P_1, P_2, P_3\}$ with:

- P_1: The alternative is in line with the will of the citizens.
- P_2: The alternative creates jobs.
- P_3: The alternative attracts tourists to the city.

Suppose the agent is a representative of the people's party with the following weighing relation: $\{P_2, P_3\} \succeq \{P_1\} \succeq \emptyset$. Suppose that, initially, the politician's sole concern is the will of the citizens. Thus, $S = \{P_1\}$ and, because the weighing relation is such that $\{P_1\} \succeq \emptyset$, he prefers the cinema. However, in a political debate someone argues that the museum benefits the local residents more than the cinema. Firstly, the museum creates jobs for the inhabitants of the citizens. Secondly, it attracts tourists which is good for local retailers, pubs and restaurants. The motivational state of our politician now becomes $S' = \{P_1, P_2, P_3\}$ and, by definition of the weighing relation, the representative of the

people's party now favours the museum. The meta-level preferences of the agent have not changed but his perspective, i.e. his motivational state, has. This caused him to change his preferences over the options.

4.2 Models for Preference Formation

The quintessence of democratic deliberation is that people justify their preferences to one another. Therefore, the formal framework for democratic deliberation has to be able to account for preference formation. Formally speaking, a democratic deliberation is a communicative situation in which a finite set $\mathcal{N}$ of agents is deliberating about a finite set $\mathcal{X}$ of alternatives. Essential to the deliberative process is that agents justify their preferences over these alternatives to one another. In other words, they have to explain where their preferences come from. Therefore, we introduce a finite set $\mathbb{P}$ of properties or reasons that the agents can appeal to in justifying their preferences. These properties or reasons are usually fundamental values that play a role in the decision at hand.

With these parameters in place, we can define the formal models for preference formation:

Definition 4.1. Let $\mathcal{N}$ be a finite set of agents, $\mathcal{X}$ a finite set of alternatives and $\mathbb{P}$ a finite set of properties or reasons. Let L denote the set of all linear orders on $\mathcal{P}(\mathbb{P})$.[1] A *model for preference formation* or a *preference formation model* is a tuple $M = (F, S_i, \succeq_i, r)_{i\in\mathcal{N}}$ where:

- $F = (W, \sim_i, \leq_i)_{i\in\mathcal{N}}$ is a multi-agent plausibility frame,
- $S_i : W \to \mathcal{P}(\mathbb{P})$ is an agent-dependent function that assigns to each world w some $S_i(w) \subseteq P$ called the *current motivational state* of agent i,
- $\succeq_i : W \to L$ is an agent-dependent function that assigns to each world w some $\succeq_i (w) \in L$ called the *meta-level preference order* of agent i,
- $r : W \times \mathcal{X} \to \mathcal{P}(\mathbb{P})$ is a function that assigns to each pair of a world w and an alternative x a set of properties or reasons $r(w, x) \subseteq P$,

such that the following hold:

1. For all $i \in \mathcal{N}$ and all $w, w' \in W : w \sim_i w'$ implies $S_i(w) = S_i(w')$.
2. For all $i \in \mathcal{N}$ and all $w, w' \in W : w \sim_i w'$ implies $\succeq_i (w) = \succeq_i (w')$.

We adopt the following notion: $\succeq_i^w := \succeq_i (w)$. If F is a common prior frame, M is called a *common prior model for preference formation.* If F is a standard frame, M is called a *standard model for preference formation.* If F is both standard and common prior, M is called a *standard common prior model for preference formation.*

The models for preference formation are variations of EDL-models. Recall that models for EDL consist of a multi-agent plausibility frame together with a valuation function. Models for preference formation also consist of a multi-agent plausibility frame. Thus, they allow us to model the epistemic and doxastic notions introduced in Chapter 2. However, instead of one valuation function, the preference formation models have three different kinds of "valuation functions".

The function $r : W \times \mathcal{X} \to \mathcal{P}(\mathbb{P})$ is a valuation function in the original sense of the word, as it represents ontic facts. To each combination of world and an alternative, r assigns a property package that contains precisely the properties that hold of the given alternative in the given world. In other words, $P \in r(w, x)$ means that in world w property P holds of alternative x. If one interprets the elements of P as reasons, $P \in r_i(w, x)$ mans that in world w it is possible to appeal to reason P in justifying one's preferences for alternative x. In this case, $r(w, x)$ can be thought of as the justification for x. Thus, depending on the interpretation of $\mathbb{P}$, we sometimes use the term justification instead of property package.

[1] Recall that a linear order is a binary relation that is antisymmetric, transitive and total.

The agent-dependent function $S_i : W \to \mathcal{P}(\mathbb{P})$ is a motivational salience function that assigns to each world a set of properties or reasons that, according to agent i, are essential in the decision at hand. Thus, $P \in S_i(w)$ means that property P is motivationally salient for agent i in determining his preferences at world w. As $S_i(w)$ represents the motivational state of agent i at world w, it can be seen as the perspective or viewpoint from which agent i approaches the issue at stake. Condition 1 says that the agent's motivational state is introspective. That is, if an agents considers a certain property to be motivationally salient, he knows that he does:

Proposition 4.2. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a model for preference formation, $i \in \mathcal{N}$ and $P \in \mathbb{P}$. Let $Q := \{w \in W \mid P \in S_i(w)\}$. Then $Q = K_i Q$.

Proof. The $\supseteq$-inclusion follows from the factivity of knowledge. For the $\subseteq$-inclusion, let $w \in Q$ be arbitrary and let $w' \in W$ be such that $w \sim_i w'$. By condition 1, it follows that $S_i(w) = S_i(w')$. From this, it follows that $w' \in Q$ and, hence, that $w \in K_i Q$. Thus, $Q \subseteq K_i Q$. □

In every world, each alternative is endowed with a property package and each agent with a motivational state. This allows us to define motivationally salient property packages:

Definition 4.3. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $x \in \mathcal{X}$. The *motivationally salient property package for agent i corresponding to x in world w* is defined as: $r_i(w, x) := r_i(w, x) \cap S_i(w)$.

Thus, the motivationally salient property package for agent i corresponding to x at world w contains precisely the properties that (i) hold of x at world w and (ii) agent i considers to be motivationally salient. An agent is indifferent between two alternatives if the motivationally salient property packages corresponding to them are the same:

Definition 4.4. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. Agent i is *indifferent* between alternatives x and y at world w iff $r_i(w, x) = r_i(w, y)$. Notation: $x \approx_i^w y$.

The agent-dependent function $\succeq_i$ assigns to each world a linear order on $\mathcal{P}(\mathbb{P})$. This linear order corresponds to Dietrich and List's weighing function and represents the meta-level preferences of the agent. In other words, $J \succeq_i^w J'$ with $J, J' \subseteq \mathbb{P}$, means that at world w agent i prefers property package J over J'. Condition 2 captures introspection of meta-level preferences. Thus, given two property packages, the agent knows which one he prefers:

Proposition 4.5. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a model for preference formation, $i \in \mathcal{N}$ and $J, J' \in \mathcal{P}(\mathbb{P})$. Let $Q := \{w \in W \mid J \succeq_i^w J'\}$. Then $Q = K_i Q$.

Proof. The proof is similar to that of Proposition 4.2. □

The fact that an agent's meta-level preferences are represented by a linear order on $\mathcal{P}(\mathbb{P})$ requires some discussion. We start by discussing some philosophically interesting constraints on the linear order on $\mathcal{P}(\mathbb{P})$. Afterwards, we answer the objection that it is unrealistic that the agent's meta-level preferences are represented by an order on the full powerset of $\mathbb{P}$.

In order not to loose any generality, we have not put any constraints on the linear order on $\mathcal{P}(\mathbb{P})$. However, two are worth considering. We start with the condition of separability:

Definition 4.6. A linear order $\succeq_i^w$ on $\mathcal{P}(\mathbb{P})$ satisfies *separability* if the following holds: For all $J, J', J^* \in \mathcal{P}(\mathbb{P})$ such that $J^* \cap J = \emptyset = J^* \cap J'$, $J \succeq_i^w J'$ iff $J \cup J^* \succeq_i^w J' \cup J^*$.

Philosophically, separability means that the agent evaluates each property or reason in $\mathbb{P}$ independently from the other properties. Separability does not hold in general. Consider, for instance, the following example in which the local government has to decide what kind of attraction to build in the city centre. Suppose the motivational state of a representative of the people's party consists of the will

of the citizens (P_1) and the attraction of tourists to the city (P_2). Suppose he prioritises the will of the citizens over tourist attraction, i.e. $\{P_1\} \succeq \{P_2\}$. In a debate, the politician is confronted with the costs of the different alternatives. Let P_3 denote the property of having high cost. It may well be the case that the politician prefers $\{P_2, P_3\} \succeq \{P_1, P_3\}$, because he understands that the city somehow needs to earn back the costs of the project. Separability is not satisfied, but we are still inclined to view the meta-level preferences of the agent as reasonable. However, as Dietrich and List (2013a: 120) argue, separability is an important property because only when it holds "any given reason can unambiguously be said to count 'in favour of' or 'against' the alternatives of which it is true. Without separability, the question of whether a reason counts for or against those alternatives is also dependent on which other reasons are present." Next, we consider the monotonicity condition:

Definition 4.7. A linear order $\succeq_i^w$ on $\mathcal{P}(\mathbb{P})$ is *monotonic* if the following holds: If $J, J' \in \mathcal{P}(\mathbb{P})$ are such that $J \supseteq J'$, then $J \succeq_i^w J'$.

Philosophically, the monotonicity condition says that all the properties in $\mathbb{P}$ are positive. From the perspective of deliberative democracy this condition makes sense, because agents have to "give one another reasons that are mutually acceptable and generally accessible" (Gutmann & Thompson, 2004: 28). A reason is mutually acceptable if it can, in principle, be accepted by other people. That is, although people might disagree on the relative importance of reasons, they cannot deny that the reasons others use to justify their preferences have intrinsic worth. For instance, suppose the government is debating about which energy policies to fund. A conservative party might believe that low energy prices should be the decisive factor, but should acknowledge the intrinsic worth of the green party's justification for solar energy with its appeal to sustainability. Thus, the mutual acceptability of reasons indeed captures the idea that the reasons in $\mathbb{P}$ are positive.

Lastly, we answer the objection that it is unrealistic that the agents have meta-level preferences over the full powerset of $\mathbb{P}$. This is unrealistic for at least two reasons. Firstly, powersets are exponentially large with respect to the set of properties $\mathbb{P}$ and it is unlikely that the agents have compared all subsets of $\mathbb{P}$. If 5 properties play a role in a certain decision, the agents probably do not have a complete order over all 32 different property packages. Secondly, the properties in $\mathbb{P}$ might be related. Some properties might be mutually exclusive and, thus, exclude each other a priori. Or, it might be the case that if a certain property P holds, P' does as well. For instance, suppose elections are coming up and everybody knows that parties that favour a more sustainable energy policy (P_1) are also in favour of more investments in public transport (P_2), as both benefit the environment. In this case, comparing property packages that contain P_1 but not P_2, does not make any sense. Both objections are just and amount to saying that each real life situation comes with an intuitive set $\mathcal{J} \subseteq \mathcal{P}(\mathbb{P})$, containing only property packages that are consistent and relevant, and an ordering on $\mathcal{J}$. However, from a formal point of view, we can always work with the full powerset. When modelling specific situations, we can extend an ordering on $\mathcal{J}$ to an ordering on $\mathcal{P}(\mathbb{P})$ by putting all the impossible or irrelevant packages at the bottom. For reasons of mathematical simplicity, we therefore choose to work with an ordering on the full $\mathcal{P}(\mathbb{P})$.

4.3 Preferences under Uncertainty

In the theory of preference formation of Dietrich and List (2013a, 2013b), people's preferences over alternatives are based on the meta-level preferences over motivationally salient property packages. Indeed, the meta-level preferences and the motivational state of an agent play an important role in determining his extrinsic preferences. But that is not all. Under incomplete information, people also base their preferences on the knowledge and beliefs they have about the alternatives. The models for preference formation allow for preference definitions that take into account both the agent's epistemic doxastic state and the properties that are motivationally salient for him. Section 4.3.1 discusses several ways of defining preference, all of which are inspired by standard notions from decision theory. In Section 4.3.2, we explore the relation between preference, knowledge and belief. Section 4.3.3 is an excursion that shows how Liu's (2011) account of preference formation, that defines extrinsic

preferences in terms of epistemic doxastic attitudes and the properties of the alternatives, can be incorporated into our framework.

4.3.1 Defining Preferences

In general, agents do not possess all the information about the state of the world whenever they have to make a decision between different alternatives. They form their preferences under uncertainty. This section considers five different ways of defining preference under uncertainty, all of which are based on well-known concepts from decision theory, game theory and economics: weak dominance, maximin, leximin, maximax and leximax.[2] For each type, we define a knowledge and a belief-based variant. Let $\mathcal{N}$ be a finite set of agents, $\mathcal{X}$ a finite set of alternatives and $\mathbb{P} = \{P_1, ..., P_k\}$ a finite set of properties or reasons. Let L be the set of all linear orders on $\mathcal{P}(\mathbb{P})$. Additionally, let $x\mathbf{R}_i^w y$ denote the fact that *agent i weakly prefers alternative x to y in world w.*

To begin with, we consider weak dominance. Knowledge-based weak dominance captures the idea that alternative x is preferred to alternative y if the motivationally salient property package corresponding to x is at least as good as the one corresponding to y in all epistemically possible worlds. That is, if the agent in question knows that the motivationally salient property package corresponding to x is just as good as the one corresponding to y. For belief-based weak dominance, the agent has to believe this is the case:

Definition 4.8. Let M be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $x, y \in X$. *Weak dominance preference* is defined as follows:

- **Knowledge-Based Weak Dominance**: $x\mathbf{R}_i^w y$ iff for all $w' \in W$ such that $w \sim_i w'$: $r_i(w', x) \succeq_i^{w'} r_i(w', y)$.
- **Belief-Based Weak Dominance**: $x\mathbf{R}_i^w y$ iff for all $w' \in W$ such that $w \sim_i w'$ there exists a w'' such that $w' \leq_i w''$ and for all w''' such that $w'' \leq_i w'''$ it holds that $r_i(w''', x) \succeq_i^{w'''} r_i(w''', y)$.

The definition of belief-based weak dominance becomes more clear when we consider standard models:

Proposition 4.9. Let M be a standard model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $x, y \in X$. Let $x\mathbf{R}_i^w y$ be defined as belief-based weak dominance. Then the following holds:

- $x\mathbf{R}_i^w y$ iff for all $v \in W$ such that $w \rightarrow_i v : r_i(v, x) \succeq_i^v r_i(v, y)$.

Proof. Left to the reader. □

Thus, in standard models, belief-based weak dominance captures the idea that an agent prefers x to y if, in all the most plausible worlds, the motivationally salient property package corresponding to x is at least as good as the one corresponding to y. That is, if the agent believes the motivationally salient property package corresponding to x is better than the one corresponding to y. Notice that, in standard models, the definitions of knowledge and belief-based weak dominance are similar. Where knowledge-based weak dominance compares the motivationally salient property packages in all epistemically possible worlds, belief-based weak dominance solely looks at the most plausible worlds.

Before we can define the other notions of preference, we need to define the set of all motivationally salient property packages corresponding to a certain alternative that are consistent with the agent's knowledge or belief:

[2] For a thorough discussion of these concepts, see Peterson (2009): 40-63.

Definition 4.10. Let M be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $x \in \mathcal{X}$. The *set of all motivationally salient property packages corresponding to x consistent with i's knowledge at w* is defined as $X_i^K(w) := \{J \subseteq \mathbb{P} \mid \exists w' \in W(w \sim_i w' \wedge J = r_i(w', x))\}$. The *set of all motivationally salient property packages corresponding to x consistent with i's beliefs at w* is defined as $X_i^B(w) := \{J \subseteq \mathbb{P} \mid \exists w' \in W(w \sim_i w' \wedge \forall w'' \in W(w' \leq_i w'' \Rightarrow \exists w''' \in W(w'' \leq_i w''' \wedge r_i(w''', x) = J)))\}$.

In the case of standard models, the definition of the set of motivationally salient property packages corresponding to a certain alternative consistent with the agent's knowledge and the one consistent with the agent's beliefs look similar:

Proposition 4.11. Let M be a standard model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $x \in \mathcal{X}$. Then $X_i^B(w) = \{J \subseteq \mathbb{P} \mid \exists w' \in W(w \rightarrow_i w' \wedge J = r_i(w', x))\}$.

Proof. Left to the reader. □

In order to define preferences, we need to show that the set of all motivationally salient property packages corresponding to a certain alternative that are consistent with the agent's knowledge or beliefs are non-empty:

Proposition 4.12. Let M be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $x \in \mathcal{X}$. The following hold:

1. $X_i^K(w) \neq \emptyset$.
2. $X_i^B(w) \neq \emptyset$.

Proof. Item 1 follows from the fact that $\sim_i$ is reflexive. For item 2, suppose for contradiction that $X_i^B(w) = \emptyset$. Notice that since $\mathbb{P}$ is finite, $\mathcal{P}(\mathbb{P})$ is as well. Let $J_0, ..., J_n$ be an enumeration of all possible motivationally salient property packages. Since $X_i^B(w) = \emptyset$, it follows that for all $m \leq n$: $J_m \notin X_i^B(w)$. That is, for all $m \leq n$ the following holds: for all $w' \in W$ such that $w \sim_i w'$ there exists a $w'' \in W$ such that $w' \leq_i w''$ and for all $w''' \in W$ such that $w'' \leq_i w'''$, it holds that $r_i(w''', x) \neq J_m$. Let this fact be denoted by $(*)$.

> Claim: For all $m \leq n$, there exists a $w_m \in W$ such that $w \sim_i w_m$ and for all $v \in W$ such that $w_k \leq_i v$, it holds that $r_i(v, x) \notin \{J_0, ..., J_m\}$.
> Proof: The proof is by induction on m.
>
> - Base Case: $m = 0$. Notice that $w \sim_i w$. Thus, according to $(*)$, there exists a $w'' \in W$ such that $w \leq_i w''$ and for all $w''' \in W$ such that $w'' \leq_i w'''$, it holds that $r_i(w''', x) \neq J_0$. Notice that $w \leq_i w''$ implies $w \sim_i w''$. Thus, in order to prove the base case, we can simply take $w_0 = w''$.
> - Inductive Case: Let $m < n$ and suppose the claim holds for all $k \leq m$. We now prove the claim for $m + 1$. Notice that $w \sim_i w_m$. Thus, according to $(*)$, there exists a $w'' \in W$ such that $w_m \leq_i w''$ and for all $w''' \in W$ such that $w'' \leq_i w'''$, it holds that $r_i(w''', x) \neq J_{m+1}$. Let $v \in W$ be such that $w'' \leq_i v$. By $(*)$, it follows that $r_i(v, x) \neq J_{m+1}$. Furthermore, notice that $w_m \leq_i w'' \leq_i v$ implies that $w_m \leq_i v$. The induction hypothesis now gives us that $r_i(v, x) \notin \{J_0, ..., J_m\}$. Combining these gives us that $r_i(v, x) \notin \{J_0, ..., J_{m+1}\}$.

We now apply this claim to $m = n$. That is, there exists a $w_n \in W$ such that $w \sim_i w_n$ and for all $v \in W$ such that $w_k \leq_i v$, it holds that $r_i(v, x) \notin \{J_0, ..., \mathcal{J}_n\} = \mathcal{P}(\mathbb{P})$. Let $v = w_n$. Since $w_n \leq_i w_n$, it follows that $r_i(w_n, x) \notin \mathcal{P}(\mathbb{P})$. This is a contradiction. Thus, $X_i^B(w) \neq \emptyset$. □

This proposition says that, for each alternative, there always exists a non-empty set of motivationally salient property packages corresponding to that alternative that are consistent with the agent's knowledge/beliefs.

With definition of the set of all motivationally salient property packages corresponding to a certain alternative that are consistent with the agent's knowledge/beliefs in place, we can define knowledge and belief-based maximin, leximin, maximax and leximax preferences.

The maximin and leximin preferences model the preference formation of a risk-averse or pessimistic agent, who focuses on the worst case scenario. In order to formalise this, we need to be able to define minimal elements:

Definition 4.13. Let M be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $\mathcal{J} \subseteq \mathcal{P}(\mathbb{P})$. The *minimal* $\succeq_i^w$*-element* of $\mathcal{J}$ is defined as $\min(\mathcal{J}) := \{J \in \mathcal{J} \mid \forall J' \in \mathcal{J}(J' \succeq_i^w J)\}$.

Notice that because $\succeq_i^w$ is a linear order, $\min(\mathcal{J})$ is unique. A risk-averse agent with maximin preference considers the worst possible outcome for each alternative and favours the alternative with the best worst outcome. Thus, alternative x is preferred to y if the worst motivationally salient property package corresponding to x consistent with the agent's knowledge/belief is better than the worst one corresponding to y:

Definition 4.14. Let M be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. *Maximin preference* is defined as follows:

- **Knowledge-Based Maximin**: $x\mathbf{R}_i^w y$ iff $\min(X_i^K(w)) \succeq_i^w \min(Y_i^K(w))$.
- **Belief-Based Maximin**: $x\mathbf{R}_i^w y$ iff $\min(X_i^B(w)) \succeq_i^w \min(Y_i^B(w))$.

Although the definition of maximin preference is a simple way to capture the preferences of pessimistic or risk-averse agent, it has been criticised. Consider the following simplified example in which a local government has to decide whether to build a museum or an entertainment park and wants to maximise the profit. Suppose there are two possible states of the world, w and w'. In w, the museum yields 1 million euros a year and the entertainment park 2 million. In w', the museum is good for 5 million euros and the entertainment park for 1 million. According to the definition of maximin preference, the government is indifferent between the two alternatives. This is counterintuitive, because the government cannot loose anything by building the museum. In order to solve this problem, maximin preference can be generalised to leximin. In order to do this, we define a sequence of minimal elements by recursion:

Definition 4.15. Let M be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $\mathcal{J} \subseteq \mathcal{P}(\mathbb{P})$. $\min_n(\mathcal{J})$ with $n \in \mathbb{N}$ is defined recursively:

- $\min_0(\mathcal{J}) := \min(\mathcal{J})$.
- Suppose $\min_{n-1}(\mathcal{J})$ is defined. $\min_n := \min(\mathcal{J} \setminus \{\min_0(\mathcal{J}), ..., \min_{n-1}(\mathcal{J})\})$.

Definition 4.16. Let M be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. *Leximin preference* is defined as follows:

- **Knowledge-Based Leximin**: Let $m := min\{|X_i^K(w)|, |Y_i^K(w)|\}$. $x\mathbf{R}_i^w y$ iff one of the following hold:
  1. There exists a $k \leq m$ such that
     (a) $\min_k(X_i^K(w)) \succeq_i^w \min_k(Y_i^K(w))$ and *not* $\min_k(Y_i^K(w)) \succeq_i^w \min_k(X_i^K(w))$

(b) and for all $l < k$: $\min_l(X_i^K(w)) \succeq_i^w \min_l(Y_i^K(w))$ and $\min_l(Y_i^K(w)) \succeq_i^w \min_l(X_i^K(w))$.[3]

2. For all $l \leq m$, it holds that

(a) $\min_l(X_i^K(w)) \succeq_i^w \min_l(Y_i^K(w))$ and $\min_l(Y_i^K(w)) \succeq_i^w \min_l(X_i^K(w))$

(b) and $|X_i^K(w)| \geq |Y_i^K(w)|$.

- **Belief-Based Leximin**: Let $m := min\{|X_i^B(w)|, |Y_i^B(w)|\}$. $x\mathbf{R}_i^w y$ iff one of the following hold:

1. There exists a $k \leq m$ such that

(a) $\min_k(X_i^B(w)) \succeq_i^w \min_k(Y_i^B(w))$ and *not* $\min_k(Y_i^B(w)) \succeq_i^w \min_k(X_i^B(w))$

(b) and for all $l < k$: $\min_l(X_i^B(w)) \succeq_i^w \min_l(Y_i^B(w))$ and $\min_l(Y_i^B(w)) \succeq_i^w \min_l(X_i^B(w))$.

2. For all $l \leq m$, it holds that

(a) $\min_l(X_i^B(w)) \succeq_i^w \min_l(Y_i^B(w))$ and $\min_l(Y_i^B(w)) \succeq_i^w \min_l(X_i^B(w))$

(b) and $|X_i^B(w)| \geq |Y_i^B(w)|$.

The definitions of knowledge and belief-based leximin might seem a bit strange at first sight. The first condition may look familiar from decision or game theory, whereas the second might not. Condition 1 says that alternative x is preferred to alternative y if there is a point at which x performs strictly better than y and before that point both alternatives do equally well. In standard decision theory, where one considers a fixed number of scenarios in all of which the pay-offs or outcomes for both x and y are defined, this suffices to define strict leximin preferences.[4] In our framework, however, we consider the set of all motivationally salient property packages corresponding to x consistent with the agent's knowledge/belief and similarly for y. In general, these sets do not have the same cardinality. It might be the case, for instance, that agent i has hard information at world w about the motivationally salient property package corresponding to y, whereas he does not have the same kind of information about x. Then $|X_K^i(w)| > 1$ and $|Y_K^i(w)| = 1$. Consider a variant of the above example in which the government has to decide between an entertainment park and a museum. The government knows the profit from the entertainment park is 1 million euros a year, whereas the profit from the museum might be either 1 or 5 million. In this case, preferring the museum is the rational thing to do. Condition 2 takes care of situations like these. It says that x is weakly preferred to y if both do equally well up to point m and there are no other outcomes for both or if both do equally well up to point m and there are better outcomes for x but not for y.

Where risk-averse or pessimistic agents consider the worst possible outcomes, optimistic agents focus on the best possible outcomes. The formal preference notions that model optimistic agents are maximax and leximax. In order to define these, we need the notion of maximal elements:

Definition 4.17. Let M be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $\mathcal{J} \subseteq \mathcal{P}(\mathbb{P})$. The *maximal* $\succeq_i^w$*-element* of $\mathcal{J}$ is defined as $\max(\mathcal{J}) := \{J \in \mathcal{J} \mid \forall J' \in \mathcal{J}(J \succeq_i^w J')\}$.

Notice that because $\succeq_i^w$ is a linear order, $\max(\mathcal{J})$ is unique. An agent with maximax preferences considers the best possible outcome for each alternative and favours the alternative with the best best outcome. Thus, alternative x is preferred to y if the best possible motivationally salient property package corresponding to x consistent with the agent's knowledge/belief is better than the one corresponding to y:

Definition 4.18. Let M be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. *Maximax preference* is defined as follows:

[3]Notice that if for some $l \leq m$ it is the case that $\min_l(X_i^K(w)) \succeq_i^w \min_l(Y_i^K(w))$ and $\min_l(Y_i^K(w)) \succeq_i^w \min_l(X_i^K(w))$, then $\min_l(X_i^K(w)) = \min_l(Y_i^K(w))$ because $\succeq_i^w$ is a linear order. However, we work with the above definition as this definition can also be used in cases where $\succeq_i^w$ is a total pre-order. The same comment applies to the definition of belief-based leximin preference.

[4]See, for instance, Petterson (2009): 44-46.

- **Knowledge-Based Maximax**: $x\mathbf{R}_i^w y$ iff $\max(X_i^K(w)) \succeq_i^w \max(Y_i^K(w))$.
- **Belief-Based Maximax**: $x\mathbf{R}_i^w y$ iff $\max(X_i^B(w)) \succeq_i^w \max(Y_i^B(w))$.

Just as the definition of maximin preference provides a simple way to model a pessimistic agent, maximax preference is a simple way to model optimistic agents. However, the objection to minimax preferences can *mutatis mutandis* also be raised against maximax preference. Therefore, we extend this definition to leximax preferences:

Definition 4.19. Let M be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $\mathcal{J} \subseteq \mathcal{P}(\mathbb{P})$. $\max_n(\mathcal{J})$ with $n \in \mathbb{N}$ is defined recursively:

- $\max_0(\mathcal{J}) := \max(\mathcal{J})$.
- Suppose $\max_{n-1}(\mathcal{J})$ is defined. $\max_n := \max(\mathcal{J} \setminus \{\max_0(\mathcal{J}), ..., \max_{n-1}(\mathcal{J})\})$.

Definition 4.20. Let M be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. *Leximax preference* is defined as follows:

- **Knowledge-Based Leximax**: Let $m := min\{|X_i^K(w)|, |Y_i^K(w)|\}$. $x\mathbf{R}_i^w y$ iff one of the following hold:
  1. There exists a $k \leq m$ such that
     (a) $\max_k(X_i^K(w)) \succeq_i^w \max_k(Y_i^K(w))$ and *not* $\max_k(Y_i^K(w)) \succeq_i^w \max_k(X_i^K(w))$
     (b) and for all $l < k$: $\max_l(X_i^K(w)) \succeq_i^w \max_l(Y_i^K(w))$ and $\max_l(Y_i^K(w)) \succeq_i^w \max_l(X_i^K(w))$.[5]
  2. For all $l \leq m$, it holds that
     (a) $\max_l(X_i^K(w)) \succeq_i^w \max_l(Y_i^K(w))$ and $\max_l(Y_i^K(w)) \succeq_i^w \max_l(X_i^K(w))$
     (b) and $|X_i^K(w)| \leq |Y_i^K(w)|$.
- **Belief-Based Leximax**: Let $m := min\{|X_i^B(w)|, |Y_i^B(w)|\}$. $x\mathbf{R}_i^w y$ iff one of the following hold:
  1. There exists a $k \leq m$ such that
     (a) $\max_k(X_i^B(w)) \succeq_i^w \max_k(Y_i^B(w))$ and *not* $\max_k(Y_i^B(w)) \succeq_i^w \max_k(X_i^B(w))$
     (b) and for all $l < k$: $\max_l(X_i^B(w)) \succeq_i^w \max_l(Y_i^B(w))$ and $\max_l(Y_i^B(w)) \succeq_i^w \max_l(X_i^B(w))$.
  2. For all $l \leq m$, it holds that
     (a) $\max_l(X_i^B(w)) \succeq_i^w \max_l(Y_i^B(w))$ and $\max_l(Y_i^B(w)) \succeq_i^w \max_l(X_i^B(w))$
     (b) and $|X_i^B(w)| \leq |Y_i^B(w)|$.

Notice that Condition 2.b in Definition 4.16 of leximin preferences uses $\geq$, whereas this condition uses $\leq$ in Definition 4.20 of leximax preferences. This deserves some explanation. For leximin preferences, the motivationally salient property packages that correspond to the alternatives and that are consistent with the agent's knowledge/belief are ordered from worst to best. Therefore, if two alternatives x and y perform equally well up to some point m and there are other outcomes for x but not for y, x should be preferred because the remaining outcomes are better than the ones considered thusfar. However, for leximax preferences, it is the other way around as the motivationally salient property packages that correspond to the alternatives and that are consistent with the agent's knowledge/belief are ordered from best to worst.

[5]Notice that if for some $l \leq m$ it is the case that $\max_l(X_i^K(w)) \succeq_i^w \max_l(Y_i^K(w))$ and $\max_l(Y_i^K(w)) \succeq_i^w \max_l(X_i^K(w))$, then $\max_l(X_i^K(w)) = \max_l(Y_i^K(w))$ because $\succeq_i^w$ is a linear order. However, we work with the above definition as this definition can also be used in cases where $\succeq_i^w$ is a total pre-order. The same comment applies to the definition of belief-based leximax preference.

With these different notions of preferences in place, we can show that weak dominance preference induces a pre-order on the alternatives in $\mathcal{X}$, whereas the other notions give rise to a total-pre-order:

Proposition 4.21. Let M be a model for preference formation, $w \in W$ and $i \in \mathcal{N}$. The following hold:

1. If $\mathbf{R}_i^w$ is defined as knowledge or belief-based weak dominance, $\mathbf{R}_i^w$ induces a pre-order on $\mathcal{X}$.
2. If $\mathbf{R}_i^w$ is defined as knowledge or belief-based maximin, $\mathbf{R}_i^w$ induces a total pre-order on $\mathcal{X}$.
3. If $\mathbf{R}_i^w$ is defined as knowledge or belief-based leximin, $\mathbf{R}_i^w$ induces a total pre-order on $\mathcal{X}$.
4. If $\mathbf{R}_i^w$ is defined as knowledge or belief-based maximax, $\mathbf{R}_i^w$ induces a total pre-order on $\mathcal{X}$.
5. If $\mathbf{R}_i^w$ is defined as knowledge or belief-based leximax, $\mathbf{R}_i^w$ induces a total pre-order on $\mathcal{X}$.

Proof. Left to the reader. □

Instead of a proof, we give the intuition behind this proposition. Knowledge-based weak dominance does not in general induce a total pre-order on the alternatives, because it may well be the case that alternative x outperforms y in one epistemically possible world but not in another. In the case of standard models, we can use the same argument for belief-based weak dominance: x might outperform y in one but not all most plausible worlds. Since belief-based weak dominance does not give rise to a total pre-order in the special case of standard models for preference formation, it also will not in the general case.

The other notions of preferences do induce a total pre-order, because in determining his preferences the agent "picks out" – either once in the case of maximin or maximax preferences or at each level in the case of leximin and leximax preferences – one motivationally salient property package corresponding to x and one corresponding to y and compares these. This is always possible, as $\succeq_i^w$ is a linear order.

4.3.2 Preferences, Knowledge and Belief

As Liu (2011: 102) points out, "preference describes a state of mind in the same way as belief does". If an agent prefers alternative x to y, he believes, for some reason or another, that x is better than y. Therefore, it should be the case that an agent prefers x to y iff he knows and believes that he does. And that he does not prefer x to y iff he knows and believes that he does not. In other words, agents should be fully introspective with respect to their preferences. This section shows that this is indeed the case. In order to prove this, the following lemma is useful:

Lemma 4.22. Let M be a model for preference formation, $i \in \mathcal{N}$ and $x \in X$. Let $w, v \in W$ be such that $w \sim_i v$. Then the following hold:

1. $X_i^K(w) = X_i^K(v)$
2. $X_i^B(w) = X_i^B(v)$

Proof. For the $\subseteq$-inclusion of item 1, let $J \in X_i^K(w)$. That is, there exists a $w' \in W$ such that $w \sim_i w'$ and $J = r_i(w', x)$. As $\sim_i$ is an equivalence relation, $w \sim_i v$ and $w \sim_i w'$ imply $v \sim_i w'$. Hence, $J \in X_i^K(v)$. The proof of $\supseteq$-inclusion is the same. For the $\subseteq$-inclusion of item 2, let $J \in X_i^B(w)$. That is, there exists a $w' \in W$ such that $w \sim_i w'$ and for all $w'' \in W$ such that $w' \leq_i w''$, there exists a $w''' \in W$ such that $w'' \leq_i w'''$ and $r_i(w''', x) = J$. Notice that $w \sim_i v$ and $w \sim_i w'$ imply $v \sim_i w'$. Thus, there exists a $w' \in W$ such that $v \sim_i w'$ and for all $w'' \in W$ such that $w' \leq_i w''$, there exists a $w''' \in W$ such that $w'' \leq_i w'''$ and $r_i(w''', x) = J$. Thus, $J \in X_i^B(v)$. The proof of the $\supseteq$-inclusion is similar. □

The following results show that preferences can be thought of as epistemic doxastic attitudes:

Theorem 4.23. Let M be a model for preference formation, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. Let $\mathbf{R}_i^w$ be defined as knowledge or belief-based weak dominance, maximin, leximin, maximax or leximax and let $Q := \{w \in W \mid x\mathbf{R}_i^w y\}$. Then $Q = K_iQ = B_iQ$.

Proof. For starters, suppose that $\mathbf{R}_i^w$ is defined as knowledge-based weak dominance. We prove the following three inclusions:

1. $Q \subseteq K_iQ$. Suppose $w \in Q$. That is, for all $w' \in W$ such that $w \sim_i w'$ it holds that $r_i(w', x) \succeq_i^{w'} r_i(w', y)$. In order to show that $w \in K_iQ$, we need to show that for all $v \in W$ such that $w \sim_i v$, it holds that $v \in Q$. Let $v \in W$ be such that $w \sim_i v$. Let $v' \in W$ be such that $v \sim_i v'$. Notice that $w \sim_i v \sim_i v'$ implies that $w \sim_i v'$. Since $w \in Q$, it holds that $r_i(v', x) \succeq_i^{v'} r_i(v', y)$. Consequently, $v \in Q$ and $w \in K_iQ$.

2. $K_iQ \subseteq B_iQ$. This follows from the fact that knowledge implies belief.

3. $B_iQ \subseteq Q$. According to Observation 2.10, this means that for all $w' \in W$ such that $w \sim_i w'$ there exists a $w'' \in W$ such that $w' \leq_i w''$ and for all $w''' \in W$ such that $w'' \leq_i w'''$, it holds that $w''' \in Q$. $(*)$ In order to show that $w \in Q$, we need to show that for all $v \in W$ such that $w \sim_i v$ it holds that $r_i(v, x) \succeq_i^v r_i(v, y)$. Let $v \in W$ be such that $w \sim_i v$. By $(*)$, there exists a $v' \in W$ such that $v \leq_i v'$ and for all $v'' \in W$ such that $v' \leq_i v''$, it holds that $v'' \in Q$. In particular, the reflexivity of $\leq_i$ gives us that $v' \in Q$. That is, for all $u \in W$ such that $v' \sim_i u$ it holds that $r_i(u, x) \succeq_i^u r_i(u, y)$. Using the fact that $\leq_i \subseteq \sim_i$ and that $\sim_i$ is an equivalence relation, we can conclude from $v \leq_i v'$ that $v' \sim_i v$. Consequently, $r_i(v, x) \succeq_i^v r_i(v, y)$. Thus, $w \in Q$.

In conclusion, from the fact that $Q \subseteq K_iQ \subseteq B_iQ \subseteq Q$, it follows that $Q = K_iQ = B_iQ$.

Now, suppose that $\mathbf{R}_i^w$ is defined as belief-based weak dominance. Again, we prove the following three inclusions:

1. $Q \subseteq K_iQ$. Suppose $w \in Q$. That is, for all $w' \in W$ such that $w \sim_i w'$ there exists a $w'' \in W$ such that $w' \leq_i w''$ and for all $w''' \in W$ such that $w'' \leq_i w'''$, it holds that $r_i(w''', x) \succeq_i^{w'''} r_i(w''', y)$. In order to show that $w \in K_iQ$, we need to show that for all $v \in W$ such that $w \sim_i v$, it holds that $v \in Q$. Let $v \in W$ be such that $w \sim_i v$. By definition of belief-based weak dominance, it remains to be shown that for all $v' \in W$ such that $v \sim_i v'$ there exists a $v'' \in W$ such that $v' \leq_i v''$ and for all $v''' \in W$ such that $v'' \leq_i v'''$, it holds that $r_i(v''', x) \succeq_i^{v'''} r_i(v''', y)$. This can easily be proven by using the fact that $w \in Q$ and that $\sim_i$ is an equivalence relation.

2. $K_iQ \subseteq B_iQ$. This follows from the fact that knowledge implies belief.

3. $B_iQ \subseteq Q$. Suppose $w \in B_iQ$. That is, for all $w' \in W$ such that $w \sim_i w'$ there exists a $w'' \in W$ such that $w' \leq_i w''$ and for all $w''' \in W$ such that $w'' \leq_i w'''$, it holds that $w''' \in Q$. $(*)$ In order to show that $w \in Q$, we need to show that for all $v \in W$ such that $w \sim_i v$ there exists a $v' \in W$ such that $v \leq_i v'$ and for all $v'' \in W$ such that $v' \leq_i v''$, it holds that $r_i(v'', x) \succeq_i^{v''} r_i(v'', y)$. Let $v \in W$ be such that $w \sim_i v$. From $(*)$, it follows that there exists a $u \in W$ such that $v \leq_i u$ and for all $u' \in W$ such that $u \leq_i u'$, it holds that $u' \in Q$. Due to the reflexivity of $\leq_i$, it follows that $u \in Q$. By item 1, it follows that $u \in K_iQ$. Furthermore, $w \sim_i v \leq_i u$ implies that $w \sim_i u$. Combining this with the fact that $u \in K_iQ$ gives us that $w \in Q$.

From these inclusions, we can conclude that $Q = K_iQ = B_iQ$.

For the knowledge and belief-based version of maximin, leximin, maximax and leximax, the fact that $Q = K_iQ = B_iQ$ follows immediately from the definitions and the fact that for all $w, w' \in W$ such that $w \sim_i w'$ it holds (i) that $\succeq_i^w = \succeq_i^{w'}$ and (ii) that $X_i^K(w) = X_i^K(w')$ and $X_i^B(w) = X_i^B(w')$, as was shown in Lemma 4.22. □

Corollary 4.24. Let M be a model for preference formation, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. Let $\mathbf{R}_i^w$ be defined as knowledge or belief-based weak dominance, maximin, leximin, maximax or leximax and let $Q' := \{w \in W \mid \neg(x\mathbf{R}_i^w y)\}$. Then $Q' = K_iQ' = B_iQ'$.

Proof. Let $\mathbf{R}_i^w$ be defined as knowledge or belief-based weak dominance, maximin, leximin, maximax or leximax preference. Notice that for all $w \in W$, it is the case that $x\mathbf{R}_i^w y$ or $\neg(x\mathbf{R}_i^w y)$. Hence $Q' = \overline{Q}$, where Q is defined as in Theorem 4.23. Therefore, we need to show that $\overline{Q} = K_i\overline{Q} = B_i\overline{Q}$. Theorem 4.23 gives us that $Q = K_iQ = B_iQ$ and, consequently, that $\overline{Q} = \overline{K_iQ} = \overline{B_iQ}$. Therefore, it suffices to show the following:

1. $K_i\overline{Q} = \overline{K_iQ}$
2. $B_i\overline{Q} = \overline{B_iQ}$

For item 1, notice that $K_i\overline{Q} = \{w \in W \mid \forall w' \in W(w \sim_i w' \Rightarrow w' \notin Q)\}$ and that $\overline{K_iQ} = \{w \in W \mid \exists w' \in W(w \sim_i w' \wedge w' \notin Q\}$. For the $\subseteq$-inclusion, suppose that $w \in K_i\overline{Q}$. Due to the reflexivity of $\sim_i$, we have $w \sim_i w$. Since $w \in K_i\overline{Q}$, it follows that $w \notin Q$. Thus, $w \in \overline{K_iQ}$. For the $\supseteq$-inclusion, suppose that $w \in \overline{K_iQ}$. That is, there exists a $w^* \in W$ such that $w \sim_i w^*$ and $w^* \notin Q$. Let $w' \in W$ be such that $w \sim_i w'$. Suppose, for contradiction, that $w' \in Q$. From the fact that $w \sim_i w'$ and $w \sim_i w^*$, it follows that $w' \sim_i w^*$. From Theorem 4.23, we know that $Q = K_iQ$ and, thus, that $w^* \in Q$. Contradiction. Thus, $w' \notin Q$. Since for all $w' \in W$ such that $w \sim_i w'$ it holds that $w' \notin Q$, it is the case that $w \in K_i\overline{Q}$.
For item 2, notice the following:

- $B_i\overline{Q} = \{w \in W \mid \forall w' \in W(w \sim_i w' \Rightarrow \exists w'' \in W(w' \leq_i w'' \wedge \forall w''' \in W(w'' \leq_i w''' \Rightarrow w''' \notin Q)))\}$
- $\overline{B_iQ} = \{w \in W \mid \exists v \in W(w \sim_i v \wedge \forall v' \in W(v \leq_i v' \Rightarrow \exists v'' \in W(v' \leq_i v'' \wedge v'' \notin Q)))\}$

For the $\subseteq$-inclusion, let $w \in B_i\overline{Q}$. Since $\sim_i$ is reflexive, the definition of $B_i\overline{Q}$ gives us that there exists a $w'' \in W$ such that $w \leq_i w''$ and for all $w''' \in W$ such that $w'' \leq_i w'''$, it holds that $w''' \notin Q$. $(*)$ In order to show that $w \in \overline{B_iQ}$, let $v = w''$. Now let $v' \in W$ be such that $v \leq_i v'$. It remains to be shown that there exists a $v'' \in W$ such that $v' \leq_i v''$ and $v'' \notin Q$. Take $v'' = v'$. The reflexivity of $\leq_i$ gives us that $v' \leq_i v'$. Since $w'' = v \leq_i v'$, $(*)$ gives us that $v' \notin Q$. Thus, $w \in \overline{B_iQ}$.

For the $\supseteq$-inclusion, let $w \in \overline{B_iQ}$. That is, there exists a $v \in W$ such that $w \sim_i v$ and for all $v' \in W$ such that $v \leq_i v'$ there exists a $v'' \in W$ such that $v' \leq_i v''$ and $v'' \notin Q$. Notice that since $v \leq_i v$, it follows that there exists a $v^* \in W$ such that $v \leq_i v^*$ and $v^* \notin Q$. Suppose, for contradiction, that $w \notin B_i\overline{Q}$. That is, there exists a $w' \in W$ such that $w \sim_i w'$ and for all $w'' \in W$ such that $w' \leq_i w''$ there exists a $w''' \in W$ such that $w'' \leq_i w'''$ and $w''' \in Q$. Since $w' \leq_i w'$, this means that there exists a $w^* \in W$ such that $w' \leq_i w^*$ and $w^* \in Q$. From Theorem 4.23, we know that $Q = K_iQ$. Consequently, $w^* \in K_iQ$. Moreover, since $\leq_i \subseteq \sim_i$ and $\sim_i$ is an equivalence relation, we can conclude from $w \sim_i v \leq_i v^*$ and $w \sim_i w' \leq_i w^*$ that $w^* \sim_i v^*$. Since $w^* \in K_iQ$, it follows that $v^* \in Q$. Contradiction. Thus, it must be the case that $w \in B_i\overline{Q}$.[6] □

4.3.3 Incorporating Liu's Framework

In *Reasoning about Preference Dynamics*, Liu (2011: 99-113) defines preference in terms of the properties of the alternatives and the doxastic state of the agent. More precisely, she defines them directly in terms of the agent's beliefs about the properties of the alternatives. This section shows how her account can be incorporated into our framework.

Let $\mathcal{N}$ be a finite set of agents, $\mathcal{X}$ a finite set of alternatives and $\mathbb{P}$ a finite set of properties. Liu (2011) endowes each agent with a priority base:

Definition 4.25. For all $i \in \mathcal{N}$, the *priority base* of agent i is a linear order $\gg_i$ on $\mathbb{P}$.

The priority base ranks properties in order of priority. In Liu's framework, the meta-level preferences of the agent can be derived from the priority base by interpreting it lexicographically. That is, given two alternatives, the agent first considers the property that has top priority. If that property holds

[6] In the proof of this corollary, we have not used the preference definitions. Therefore, we can conclude that for all $Q \subseteq W$ such that $Q = K_iQ = B_iQ$, it is the case that $\overline{Q} = K_i\overline{Q} = B_i\overline{Q}$.

of one alternative and not of the other, the alternative of which it holds is preferred. If the property with top priority holds of both or of neither, the agent moves on to the second property, and so forth. Any linear order on $\mathbb{P}$, and hence every priority base, generates a linear order on $\mathcal{P}(\mathbb{P})$:

Definition 4.26. Let $\mathbb{P}$ be a finite set of properties and let $\gg$ be a linear order on $\mathbb{P}$. The *lexicographic order on $\mathcal{P}(\mathbb{P})$ generated by $\gg$* is defined as follows: For all $J, J' \in \mathcal{P}(\mathbb{P}) : J \succeq J'$ iff there exists a $P \in \mathbb{P}$ such that

1. $P \in J$ and $P \notin J'$
2. and for all P' such that $P' \gg P$: $P \in J$ iff $P \in J'$.

A linear order $\succeq$ on $\mathcal{P}(\mathbb{P})$ is said to be *lexicographically generable* if there exists a linear order $\gg$ on $\mathbb{P}$ such that the lexicographic order on $\mathcal{P}(\mathbb{P})$ generated by $\gg$ is $\succeq$.

Notice that if $\gg$ is a linear order on $\mathbb{P}$, the lexicographic order based on $\gg$ is also a linear order. Moreover, notice that if a linear order $\succeq$ is lexicographically generable, there is a unique order $\gg$ on $\mathbb{P}$ that generates it. Recall that in Liu's framework, the lexicographic order generated by the agent's priority base represents his meta-level preferences. With the above definition in place, it is possible to define Liu-models:

Definition 4.27. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a model for preference formation. M is a *Liu-model* iff the following hold:

1. For all $w \in W$ and all $i \in \mathcal{N} : S_i(w) = \mathbb{P}$.
2. For all $w \in W$ and all $i \in \mathcal{N}$, $\succeq_i^w$ is the lexicographically generable.

The linear order on $\mathbb{P}$ that lexicographically generates $\succeq_i^w$ is called *the priority base of agent i at world w* and is denoted by $\gg_i^w$.

Thus, a Liu-model is a model for preference formation that satisfies two additional properties. Firstly, the motivational state of each agent in each world equals $\mathbb{P}$, as Liu (2011) does not distinguish between properties that are and properties that are not motivationally salient for the agent. Secondly, the meta-level preferences of the agents can be thought of as obtained from a priority base.

Liu (2011: 100-101) introduces three different definitions of preference, all of which are based on the agent's beliefs and the properties in his priority base: decisive preference, conservative preference and deliberate preference. Before we formally define these in our framework, we introduce the following notation:

Definition 4.28. Let M be a model for preference formation, $P \in \mathbb{P}$ and $x \in \mathcal{X}$. We define the set $P(x) := \{w \in W \mid P \in r(w, x)\}$.

For starters, we consider Liu's notion of decisive preference. In order to determine whether alternative x is preferred to alternative y, the agent runs through his priority base. As soon has he encounters a property of which he believes it holds of one and he does not believe that it holds of the other, he prefers the former alternative:

Definition 4.29. Let M be a Liu-model, $w \in W$, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. *Decisive preference* is defined as follows: $x\mathbf{R}_i^w y$ iff $\{P \in \mathbb{P} \mid w \in B_i P(x)\} \succeq_i^w \{P \in \mathbb{P} \mid w \in B_i P(y)\}$.

Notice that in a Liu-model, $\succeq_i^w$ is obtained from some priority base $\gg_i^w$. Thus, if an agent decisively prefers alternative x to alternative y, this indeed means that (i) there exists a property $P \in \mathbb{P}$ such that $w \in B_i P(x)$ and $w \in \neg B_i P(y)$ and that (ii) for all $P' \in \mathbb{P}$ such that $P' \gg_i^w P$ it is the case that $w \in B_i P'(x)$ iff $w \in B_i P'(y)$.

Conservative preference requires something stronger. Given two alternatives x and y, the agent again forms his preference by running through his priority base. Let P be the property with top priority. If the agent believes P holds of x and he believes P does not hold of y (or vice versa), he

prefers x to y (or vice versa). If this is not the case, it depends on his information whether he moves on to the next property. If it is the case that $w \in B_i P(x)$ iff $w \in B_i P(y)$ and that $w \in B_i \neg P(x)$ iff $w \in B_i \neg P(y)$, he does. Otherwise, he does not. In other words, the agent only moves on to the next property if he actually has beliefs about whether or not P holds of the relevant alternatives. If for one of the alternatives, he does not believe that P holds and he does not believe that P does not hold, he remains undecided:

Definition 4.30. Let M be a Liu-model, $w \in W$, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. Let $\gg_i^w$ be the linear order that generates $\succeq_i^w$. *Conservative preference* is defined as follows: $x\mathbf{R}_i^w y$ iff there exists a $P \in \mathbb{P}$ such that

1. $w \in B_i P(x)$ and $w \in B_i \neg P(y)$
2. and for all $P' \in \mathbb{P}$ such that $P' \gg_i^w P$:
   - $w \in B_i P'(x)$ iff $w \in B_i P'(y)$
   - and $w \in B_i \neg P'(x)$ iff $w \in B_i \neg P'(y)$.

Lastly, we consider deliberate preference. An agent has deliberate preference for x over y iff he believes that x is at least as good as y:

Definition 4.31. Let M be a Liu-model, $w \in W$, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. *Deliberate preference* is defined as follows: $x\mathbf{R}_i^w y$ iff for all $w' \in W$ such that $w \sim_i w'$ there exists a $w'' \in W$ such that $w' \leq_i w''$ and for all $w''' \in W$ such that $w'' \leq_i w'''$, it holds that $r(w''', x) \succeq_i^{w'''} r(w''', y)$.

Deliberate preference corresponds to our notion of belief-based weak dominance, because in Liu-models the motivational state of each agent in each world equals $\mathbb{P}$, which implies that for all $i \in \mathcal{N}$, all $x, y \in \mathcal{X}$ and all $w \in W$: $r_i(w, x) = r(w, x)$:

Observation 4.32. Liu's deliberate preference is belief-based weak dominance preference.

We end this section by considering some properties of Liu's preference notions. For starters, it is the case that conservative and deliberate preferences induce a pre-order on the alternatives, whereas decisive preference gives rise to a total pre-order (hence its name):

Proposition 4.33. Let M be a Liu-model, $w \in W$ and $i \in \mathcal{N}$. The following hold:

1. If $\mathbf{R}_i^w$ is defined as decisive preference, $\mathbf{R}_i^w$ induces a total pre-order on $\mathcal{X}$.
2. If $\mathbf{R}_i^w$ is defined as conservative preference, $\mathbf{R}_i^w$ induces a pre-order on $\mathcal{X}$.
3. If $\mathbf{R}_i^w$ is defined as deliberate preference, $\mathbf{R}_i^w$ induces a pre-order on $\mathcal{X}$.

Proof. Left to the reader. □

Liu (2011) stresses that preferences are epistemic doxastic attitudes. If an agent prefers one alternative to another, he believes that that alternative is better. The next theorem shows that we indeed have full introspection for the preference definitions considered by Liu:

Theorem 4.34. Let M be a Liu-model, $w \in W$, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. Let $\mathbf{R}_i^w$ be defined as decisive, conservative or deliberate preference. Let $Q := \{w \in W \mid x\mathbf{R}_i^w y\}$. Then $Q = K_i Q = B_i Q$.

Proof. For starters, suppose that $\mathbf{R}_i^w$ is defined as decisive preference. We prove the following three inclusions:

1. $Q \subseteq K_iQ$. Suppose $w \in Q$. That is, $\{P \in \mathbb{P} \mid w \in B_iP(x)\} \succeq_i^w \{P \in \mathbb{P} \mid w \in B_iP(y)\}$. Let $w' \in W$ be such that $w \sim_i w'$. As Proposition 2.11 shows, if an agent believes something he knows that he does. Consequently, for all $P \in \mathbb{P}$ and all $z \in \mathcal{X}$ it is the case that $B_iP(z) = K_iB_iP(z)$. From the fact that $w \sim_i w'$, it follows for all $P \in \mathbb{P}$ and all $z \in X$ that $\{P \in \mathbb{P} \mid w \in B_iP(z)\} = \{P \in \mathbb{P} \mid w' \in B_iP(z)\}$. Moreover, $w \sim_i w'$ implies that $\succeq_i^w = \succeq_i^{w'}$. Therefore, $w' \in Q$ and $w \in K_iQ$.

2. $K_iQ \subseteq B_iQ$. This follows from the fact that knowledge implies belief.

3. $B_iQ \subseteq Q$. Suppose $w \in B_iQ$. By Observation 2.10, this means that for all $w' \in W$ such that $w \sim_i w'$ there exists a $w'' \in W$ such that $w' \leq_i w''$ and for all $w''' \in W$ such that $w'' \leq_i w'''$, it holds that $w''' \in Q$. Due to the reflexivity of $\sim_i$, it follows that there exists a $w'' \in W$ such that $w \leq_i w''$ and for all $w''' \in W$ such that $w'' \leq_i w'''$, it follows that $w''' \in Q$. The reflexivity of $\leq_i$ gives us that $w'' \in Q$. That is, $x\mathbf{R}_i^{w''}y$ or, equivalently, $\{P \in \mathbb{P} \mid w'' \in B_iP(x)\} \succeq_i^{w''} \{P \in \mathbb{P} \mid w'' \in B_iP(y)\}$. Notice that $w \leq_i w''$ implies that $w \sim_i w''$. Consequently, $\succeq_i^{w''} = \succeq_i^w$. Furthermore, using the same reasoning as in item 1, we can show for all $P \in \mathbb{P}$ and all $z \in X$ that $\{P \in \mathbb{P} \mid w \in B_iP(z)\} = \{P \in \mathbb{P} \mid w'' \in B_iP(z)\}$. Hence $w \in Q$.

In conclusion, $Q \subseteq K_iQ \subseteq B_iQ \subseteq Q$, i.e. $Q = K_iQ = B_iQ$.

For conservative preference, the proof is a bit more involved but the its essence lies again in the fact that the agents know what they believe. In particular, for all $P \in \mathbb{P}$, all $z \in \mathcal{X}$ and all $w, w' \in W$ such that $w \sim_i w'$, it holds that $\{P \in \mathbb{P} \mid w \in B_iP(z)\} = \{P \in \mathbb{P} \mid w' \in B_iP(z)\}$ and $\{P \in \mathbb{P} \mid w \in B_i\neg P(z)\} = \{P \in \mathbb{P} \mid w' \in B_i\neg P(z)\}$. For deliberate preference, the theorem follows immediately from Observation 4.32 and Theorem 4.23. □

The next result follows as a corollary:

Corollary 4.35. Let M be a Liu-model, $w \in W$, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. Let $\mathbf{R}_i^w$ be defined as decisive, conservative or deliberate preference. Let $Q' := \{w \in W \mid \neg(x\mathbf{R}_i^w y)\}$. Then $Q' = K_iQ' = B_iQ'$.

Proof. The proof is exactly the same as that of Corollary 4.24. □

4.4 Special Kinds of Models for Preference Formation

This section introduces two special kinds of models for preference formation: sufficient information models and opinionated models. Both types play a crucial role in Chapter 6, where we study the relation between deliberation and single-peakedness. However, as these models are interesting in their own right, they are already introduced at this stage.

Firstly, in sufficient information models, the group of agents potentially possesses all the relevant factual information about the world. That is, for each property that is motivationally salient for at least one of the group members and for each alternative, it is distributed knowledge whether that property does or does not hold of that alternative:

Definition 4.36. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i\in\mathcal{N}}$ be a model for preference formation. M is a *sufficient information model* iff the following holds:

- For all $x \in \mathcal{X}$, all $P \in \mathbb{P}$ and all $w, w' \in W$: If $w \sim_{\mathcal{N}} w'$ and $P \in \bigcup_{i\in\mathcal{N}} S_iP(w)$, then $P \in r(w, x)$ iff $P \in r(w', x)$.

Recall that $\sim_{\mathcal{N}} := \bigcap_{i\in\mathcal{N}} \sim_i$ is the accessibility relation for distributed knowledge. Given a property P that is motivationally salient for one at least one of the group members, an alternative x and a designated actual world w, the condition on sufficient information models expresses the fact that if P holds of x in the actual world, it does so in all the worlds that none of the agents can distinguish from the actual world on the basis of his hard information. In other words, it is distributed knowledge

whether or not P holds of x. Since this is the case for all motivationally salient properties and all alternatives, we can say that the group of agents together possesses sufficient information to determine which relevant ontic facts hold in the actual world.

Secondly, we consider opinionated models. Opinionated models are standard models in which the agents always have an opinion about the properties that are motivationally salient for them:

Definition 4.37. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a model for preference formation. For all $w, w' \in W$, let $w <_i w' := w \leq_i w' \wedge \neg(w' \leq_i w)$. M is an *opinionated model* iff M is standard and the following holds:[7]

- For all $i \in \mathcal{N}$, all $x \in X$, all $P \in \mathbb{P}$ and all $w, w' \in W$: If $w \sim_i w'$, $P \in S_i(w)$, $P \in r(w, x)$ and $P \notin r(w', x)$, then $w <_i w'$ or $w' <_i w$.

The condition on opinionated models says the following: If two worlds that are both epistemically possible for an agent differ in a property that is motivationally salient for him, then one world is strictly more plausible than the other. Recall that, in standard models, belief is defined as truth in all the most plausible worlds. Therefore, this implies that for any property P in the motivational state of the agent and for any alternative $x \in X$, the agent either believes that P holds of x or he believes that P does not hold of x. That is, the agent is opinionated:

Lemma 4.38. Let M be an opinionated model, $w \in W$, $i \in \mathcal{N}$, $x \in X$ and $P \in \mathbb{P}$. If $P \in S_i(w)$, then $w \in B_i P(x)$ or $w \in B_i \neg P(x)$.

Proof. Suppose that $P \in S_i(w)$. Suppose, for contradiction, that $w \notin B_i P(x)$ and $w \notin B_i \neg P(x)$. Using Definition 2.18 and Theorem 2.19, the first conjunct says that there exists a $w' \in W$ such that $w \to_i w'$ and $P \notin r(w, x)$. Similarly, the second conjunct says that there exists a $w'' \in W$ such that $w \to_i w''$ and $P \in r(w'', x)$. As $\to_i \subseteq \sim_i$ and $\sim_i$ is an equivalence relation, it follows that $w' \sim_i w''$. As M is opinionated, it must be the case that $w' <_i w''$ or $w'' <_i w'$. Suppose $w' <_i w''$. Then it follows from $w' \leq_i w''$ and $w' \in \mathrm{Max}(\leq_i)$ that $w'' \leq_i w'$, which is a contradiction. Similarly, the case where $w'' <_i w'$ leads to a contradiction. Therefore, it must be the case that $w \in B_i P(x)$ or $w \in B_i \neg P(x)$. □

In other words, the motivationally salient property package corresponding to a given alternative is the same in all most plausible worlds:

Lemma 4.39. Let M be an opinionated model, $w \in W$, $i \in \mathcal{N}$, $x \in X$ and $P \in \mathbb{P}$. For all $w', w'' \in W$ such that $w \to_i w'$ and $w \to_i w''$, it holds that $r_i(w', x) = r_i(w'', x)$.

Proof. Let $w', w'' \in W$ be such that $w \to_i w'$ and $w \to_i w''$. Notice that because $\to_i \subseteq \sim_i$, it follows that $w \sim_i w'$ and $w \sim_i w''$. Condition 1 on preference formation models gives us that $S_i(w) = S_i(w') = S_i(w'')$. For the $\subseteq$-inclusion, let $P \in r_i(w', x)$. By definition of motivationally salient property packages, $P \in S_i(w') \cap r(w', x)$. Consequently, $P \in S_i(w)$. By Lemma 4.38, it follows that $w \in B_i P$ or $w \in B_i \neg P$. The latter leads to a contradiction, because w' is a maximal world and $P \in r(w', x)$. Therefore, it must be the case that $w \in B_i P$. As w'' is a maximal world, it follows that $P \in r(w'', x)$ and, thus, that $P \in S_i(w'') \cap r(w'', x) = r_i(w'', x)$. In conclusion, $r_i(w', x) \subseteq r_i(w'', x)$. The proof of the $\supseteq$-inclusion is similar. □

As a consequence, in opinionated models, all the belief-based notions of preference collapse and induce a total pre-order on the alternatives:

[7] The assumption of standardness is solely made for reasons of mathematical simplicity. A similar condition can be defined for preference formation models in general.

Theorem 4.40. Let M be an opinionated model, $w \in W$, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. Then the following hold:

1. All the notions of belief-based preference collapse.
2. Let $\mathbf{R}_i^w$ be defined as belief-based preference. Then $\mathbf{R}_i^w$ induces a total pre-order on the alternatives in $\mathcal{X}$.

Proof. For item 1, let $w^* \in W$ be such that $w \rightarrow_i w^*$. Notice that such a w^* exists, since the plausibility relation $\leq_i$ is converse well-founded. Firstly, according to Proposition 4.9, the definition of belief-based weak dominance in standard models amounts to: $x\mathbf{R}_i^w y$ iff for all $w' \in W$ such that $w \rightarrow_i w' : r_i(w', x) \succeq_i^{w'} r_i(w', y)$. By Lemma 4.39 and the fact that $\succeq_i^{w^*} = \succeq_i^w$, this amounts to:

$$x\mathbf{R}_i^w y \text{ iff } r_i(w^*, x) \succeq_i^{w^*} r_i(w^*, y) \text{ iff } r_i(w^*, x) \succeq_i^w r_i(w^*, y)$$

For belief-based maximin, leximin, maximax and leximax, recall that in standard models $X_i^B(w) = \{J \subseteq \mathbb{P} \mid \exists w' \in W(w \rightarrow_i w' \wedge J = r_i(w', x))\}$, as Proposition 4.11 shows. By Lemma 4.39, it follows that $X_i^B(w) = \{r_i(w^*, x)\}$. And similarly, $Y_i^B(w) = \{r_i(w^*, y)\}$. As both the set $X_i^B(w)$ and the set $Y_i^B(w)$ consist of precisely 1 element, the definitions of belief-based maximin, leximin, maximax and leximax all collapse into:

$$x\mathbf{R}_i^w y \text{ iff } r_i(w^*, x) \succeq_i^w r_i(w^*, y)$$

Thus, all the definitions of belief-based preference collapse. Consequently, item 2 follows immediately from Proposition 4.21.1. □

4.5 The Logic of Preference Formation

This section is about the logic of preference formation (LPF), which is obtained by introducing a language for the preference formation models. The syntax and semantics of LPF are defined in Section 4.5.1. Section 4.5.2 proves soundness and completeness. In Section 4.5.3, we show that the language of LPF is expressive enough to encode the preference definitions from Section 4.3.1.

4.5.1 Syntax and Semantics of LPF

The language of LPF is a variant of the language of EDL. Instead of having one type of propositional letters, the language of LPF has three different kinds:

Definition 4.41. Let $\mathcal{N}$ be a finite set of agents, $\mathcal{X}$ a finite set of alternatives and $\mathbb{P}$ a finite set of properties or reasons. The set $\mathcal{L}_{LPF}$ of formulas of φ of LPF is defined recursively:

$$\varphi ::= Px \mid J \succeq_i J' \mid S_i P \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid \Box_i\varphi \mid D_K\varphi \mid D_\Box\varphi$$

where $x \in X$, $P \in \mathbb{P}$, $i \in \mathcal{N}$ and $J, J' \subseteq \mathbb{P}$. We define $\bot := Px \wedge \neg Px$ and $\top := \neg\bot$. The Boolean connectives $\vee$ and $\rightarrow$ are defined in terms of $\neg$ and $\wedge$ in the standard manner. The duals of the modal operators are defined in the following way: $\hat{K}_i := \neg K_i \neg$ and $\Diamond_i := \neg\Box_i\neg$. Belief is defined as $B_i := K_i\Diamond_i\Box_i$ and distributed belief as $D_B := D_K D_\Diamond D_\Box$, where $D_\Diamond := \neg D_\Box \neg$.

The intended meaning of the three different types of propositional letters is as expected. That is, the formula Px means that property P holds of alternative x. Furthermore, the formula $S_i P$ captures the fact that property P is motivationally salient for agent i. Finally, $J \succeq_i J'$ means that agent i prefers

property package J over property package J'.

As anticipated, the models for LPF are preference formation models:

Definition 4.42. A tuple $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ is a model for LPF iff it is a model for preference formation.

The truth conditions for the formulas of LPF are as follows:

Definition 4.43. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a model for LPF, $w \in W$ and $i \in \mathcal{N}$. The *satisfaction* relation $\models$ between pairs (M, w) and formulas $\varphi \in \mathcal{L}_{LPF}$ is defined as in Definition 2.31, where the truth condition for propositional letters is replaced by the following:

- $M, w \models Px$ iff $P \in r(w, x)$
- $M, w \models J \succeq_i J'$ iff $J \succeq_i^w J'$
- $M, w \models S_i P$ iff $P \in S_i(w)$

A formula φ is *valid* if for all $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ and all $w \in W$, we have $M, w \models \varphi$. We denote validity by $\models \varphi$.

4.5.2 Soundness and Completeness of LPF

The proof system of LPF is the proof system of EDL extended with some axioms about the new types of propositional letters. In particular, these axioms ensure introspection with respect to the agents' motivational states and meta-level preferences and they guarantee that the agents' meta-level preference relations are linear orders:

Definition 4.44. The proof system of LPF (notation: Λ_{LPF}) is the proof system Λ_{EDL} extended with the following axioms for all $i \in \mathcal{N}$ and all $P \in \mathbb{P}$:

- Introspection of salience: $S_i P \to K_i S_i P$
- Introspection of meta-level preferences: $J \succeq_i J' \to K_i(J \succeq_i J')$
- The relation $\succeq_i$ induces a linear order on $\mathcal{P}(\mathbb{P})$:
  - Antisymmetry: For all $J, J' \subseteq \mathbb{P}$ such that $J \neq J' : \neg(J \succeq_i J' \wedge J' \succeq_i J)$.
  - Transitivity: For all $J, J', J'' \subseteq \mathbb{P} : J \succeq_i J' \wedge J' \succeq_i J'' \to J \succeq_i J''$.
  - Totality: For all $J, J' \subseteq \mathbb{P} : J \succeq_i J' \vee J' \succeq_i J$.

As the models for preference formation are variants of multi-agent plausibility models, the soundness of LPF follows from the soundness of EDL together with the soundness of the above axioms:

Theorem 4.45. Λ_{LPF} is sound with respect to the class of all models for preference formation.

Proof. As soundness can be proven by induction on the length of the proof, it suffices to show that each axiom is sound and that the inference rules preserve truth. Theorem 2.35 shows that the axioms and the inference rules of EDL are valid on all multi-agent plausibility frames. Models for preference formation consist of multi-agent plausibility frames. Therefore, in order to show that Λ_{LPF} is sound, it suffices to show the soundness of the axioms that are not included in Λ_{EDL}.

The soundness of $S_i P \to K_i S_i P$ and $J \succeq_i J' \to K_i(J \succeq_i J')$ follow immediately from Propositions 4.2 and 4.5 respectively. The soundness of the axioms expressing antisymmetry, transitivity and totality follows from the fact that for all $i \in \mathcal{N}$ and all $w \in W$, $\succeq_i^w$ is a linear order. □

Since Λ_{LPF} is sound with respect to the class of all preference formation models, we get the following result as a corollary:

Corollary 4.46. Λ_{LPF} is sound with respect to the class of all common prior models for preference formation.

In order to show that LPF is complete, we modify the completeness proof for EDL:

Theorem 4.47. Λ_{LPF} is weakly complete with respect to the class of all common prior models for preference formation.

Proof. According to Proposition 2.38, it suffices to show that every consistent LPF-formula φ is satisfiable on some common prior model for preference formation. The completeness proof for Λ_{LPF} is a variant of the completeness proof for Λ_{EDL}. Therefore, we only indicate how the completeness proof of EDL has to be changed. The reader may check the details. Let φ be a consistent LPF-formula.

In step 1 of the completeness proof for EDL, we constructed a canonical pseudo-model. Similarly, one can define pseudo-models $K = (W, \sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}}, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ for preference formation by adding the following constraints to Definition 2.39 of pseudo-models:

1. For all $i \in \mathcal{N}$ and all $w, w' \in W : w \sim_i w'$ implies $S_i(w) = S_i(w')$.
2. For all $i \in \mathcal{N}$ and all $w, w' \in W : w \sim_i w'$ implies $\succeq_i (w) = \succeq_i (w')$.

One can now define the canonical pseudo-model $K^\Omega = (W^\Omega, \sim_i^\Omega, \leq_i^\Omega, \sim_{\mathcal{N}}^\Omega, \leq_{\mathcal{N}}^\Omega, S_i^\Omega, \succeq_i^\Omega, r^\Omega)_{i \in \mathcal{N}}$ for preference formation in the same way as the canonical pseudo-model for EDL (Definition 2.47), where the definition of V^Ω should be replaced by the following:

- For all $i \in \mathcal{N}$, $S_i^\Omega : W \to \mathcal{P}(\mathbb{P})$ is a motivational salience function defined by: $S_i^\Omega(w) := \{P \in \mathbb{P} \mid S_i P \in w\}$.
- For all $i \in \mathcal{N}$, $\succeq_i^\Omega$: $W \to L$ is a meta-level preference function defined by: $\succeq_i^\Omega (w) := \{(J, J') \in \mathcal{P}(\mathbb{P})^2 \mid (J \succeq_i J') \in w\}$.
- $r^\Omega : W \times \mathcal{X} \to P$ is a valuation function defined by: $r^\Omega(w, x) := \{P \in \mathbb{P} \mid Px \in w\}$.

One can show that the canonical pseudo-model for preference formation is indeed a pseudo-model for preference formation. The "normal constraints" on pseudo-models follow from Proposition 2.48. The axiom $S_i P \to K_i S_i P$ ensures that condition 1 holds and the axiom $J \succeq_i J' \to K_i (J \succeq_i J')$ that condition 2 holds. The fact that $\succeq_i^\Omega$ assigs a linear order on $\mathcal{P}(\mathbb{P})$ to each world follows from the antisymmetry, transitivity and totality axioms of Definition 4.44. Furthermore, the definitions of S_i^Ω, $\succeq_i^\Omega$ and r^Ω ensure that the Truth Lemma still holds. That is, for all $w \in W^\Omega$ and all formulas φ of LPF, it holds that $M^\Omega, w \models \varphi$ iff $\varphi \in w$.

Recall that φ is a consistent LPF-formula. According to Lindenbaum's Lemma, $\{\varphi\}$ can be extended to a maximal consistent set Φ. By definition of the canonical pseudo-model for preference formation, there exists a world $w^* \in W^\Omega$ such that $w^* = \Phi$. By the Truth Lemma, it follows that $K^\Omega, w^* \models \varphi$.

In step 2 of the completeness proof for EDL, we unravelled the canonical pseudo-model around some world $w \in W^\Omega$. Recall that the resulting structure was a labelled tree and that the worlds in this tree consisted of finite histories $\bar{h}$ from world w to some other world in the canonical pseudo-model. Similarly, one can define the unravelling $\vec{K} = (\vec{W}, R_{\sim_i}, R_{\leq_i}, R_{\sim_{\mathcal{N}}}, R_{\leq_{\mathcal{N}}}, \vec{S}, \vec{\succeq_i}, \vec{r})_{i \in \mathcal{N}}$ of the canonical pseudo-model for preference formation around world w^* as in Definition 2.55, where the definition of $\vec{V}$ should be replaced by the following:

- For all $i \in \mathcal{N}$, $\vec{S}_i : \vec{W} \to \mathcal{P}(\mathbb{P})$ is a motivational salience function defined by: $\vec{S}_i(\bar{h}) := S_i^\Omega(last(\bar{h}))$.
- For all $i \in \mathcal{N}$, $\vec{\succeq_i} : \vec{W} \to L$ is a meta-level preference function defined by: $\vec{\succeq_i}(\bar{h}) := \succeq_i^\Omega (last(\bar{h}))$.
- $\vec{r} : \vec{W} \times \mathcal{X} \to \mathcal{P}(\mathbb{P})$ is a valuation function defined by: $\vec{r}(\bar{h}, x) := r^\Omega(last(\bar{h}), x)$.

In step 3 of the completeness proof for EDL, we redefined the relations of the unravelling of the canonical pseudo-model in such a way that we obtained common prior model. Similarly to Definition 2.63, one can redefine the relations of the unravelled structure $\vec{K}$ in such a way that one gets a common prior model for preference formation $M = (\vec{W}, \sim_i, \leq_i, \sim_{\mathcal{N}}, \leq_{\mathcal{N}}, \vec{S}_i, \vec{\succeq}_i, \vec{r})_{i \in \mathcal{N}}$. In order to show that M is a common prior model for preference formation, one needs to do two things. Firstly, one can imitate the proofs of Lemma's 2.65 and 2.66 to show that the relations satisfy the desired properties. Secondly, one has to shows that conditions 1 and 2 on preference formation models hold. Using the proof strategy of Lemma 2.67, one can show that for all $\bar{h}, \bar{h}' \in \vec{W}$ it holds that $\bar{h} \sim_i \bar{h}'$ implies $last(\bar{h}) \sim_i^{\Omega} last(\bar{h}')$. The conditions now follow immediately from the definitions of $\vec{S}_i$ and $\vec{\succeq}_i$. One can define a mapping $f : M \to K^{\Omega}$ such that for all $\bar{h} \in \vec{W} : f(\bar{h}) = last(\bar{h}) \in W^{\Omega}$ and show, in the same way as in Lemma 2.67, that this is a bounded morphism. Consequently, M and K^{Ω} satisfy the same LPF-formulas. In particular, $M, (w^*) \models \varphi$.

In conclusion, the consistent LPF-formula φ is satisfiable on a common prior model for preference formation. Since φ was arbitrarily chosen, Λ_{LPF} is weakly complete with respect to the class of all common prior models for preference formation. □

Every common prior model for preference formation is also a model preference formation. Since the language of LPF is not able to distinguish between the two, we get the following result as a corollary:

Corollary 4.48. Λ_{LPF} is weakly complete with respect to the class of all models for preference formation.

4.5.3 Encoding Preferences in the Language of LPF

This section shows that the language of LPF is expressive enough to encode the different preference definitions discussed in Section 4.3.1.

Firstly, recall that knowledge weak dominance captures the idea that an agent prefers alternative x over alternative y iff the motivationally salient property package corresponding to x weakly dominates the one corresponding to y in all epistemically possible worlds. That is, the agent prefers x to y if he knows that the motivationally salient property package corresponding to x is better than the one corresponding to y. For belief-based weak dominance, he has to believe this is the case. These definitions can be encoded syntactically as follows:

Definition 4.49. Let $i \in \mathcal{N}$ and $x, y \in X$. *Weak dominance preference* is defined as follows:

- **Knowledge-Based Weak Dominance**: $Pref_{i,K}^{wd}(x,y) := K_i(\bigwedge_{J \subseteq \mathbb{P}} \bigwedge_{J' \subseteq \mathbb{P}} (\bigwedge_{P \in J} (S_i P \wedge Px) \wedge \bigwedge_{P \notin J} \neg(S_i P \wedge Px)) \wedge (\bigwedge_{P \in J'} (S_i P \wedge Py) \wedge \bigwedge_{P \notin J'} \neg(S_i P \wedge Py)) \rightarrow J \succeq_i J')$.
- **Belief-Based Weak Dominance**: $Pref_{i,B}^{wd} := B_i(\bigwedge_{J \subseteq \mathbb{P}} \bigwedge_{J' \subseteq \mathbb{P}} (\bigwedge_{P \in J} (S_i P \wedge Px) \wedge \bigwedge_{P \notin J} \neg(S_i P \wedge Px)) \wedge (\bigwedge_{P \in J'} (S_i P \wedge Py) \wedge \bigwedge_{P \notin J'} \neg(S_i P \wedge Py)) \rightarrow J \succeq_i J')$.

Notice that the definition of knowledge-based weak dominance starts by quantifying over all epistemically possible worlds. In order for $Pref_{i,K}^{wd}(x,y)$ to hold, the conjunction of implications that follows must be true in all these worlds. This conjunction of implications expresses that if J is the motivationally salient property package corresponding to x and J' the one corresponding to y, then the meta-level preferences of the agent are such that J is weakly preferred to J'. The definition of belief-based weak dominance is similar, but the modality K_i has been replaced by B_i.

Knowledge and belief-based maximin, leximin, maximax and leximax preferences consider motivationally salient property packages that are consistent with the agent's knowledge or belief. In order to syntactically define these preferences, we introduce abbreviations expressing the fact that a set

$J \subseteq \mathbb{P}$ is a motivationally salient property package corresponding to some alternative $x \in \mathcal{X}$ consistent with the agent's knowledge/belief:

Definition 4.50. Let $i \in \mathcal{N}$, $x \in X$ and $J \subseteq \mathbb{P}$. Additionally, let $\hat{K}_i := \neg K_i \neg$ and $\hat{B}_i := \neg B_i \neg$. We introduce the abbreviations $X_i^K(J)$ and $X_i^B(J)$, which intuitively denote that J is a motivationally salient property package that corresponds to x and is consistent with i's knowledge/belief respectively:

- $X_i^K(J) := \hat{K}_i(\bigwedge_{P \in J}(S_iP \wedge Px) \wedge \bigwedge_{P \notin J} \neg(S_iP \wedge Px))$
- $X_i^B(J) := \hat{B}_i(\bigwedge_{P \in J}(S_iP \wedge Px) \wedge \bigwedge_{P \notin J} \neg(S_iP \wedge Px))$

The definition of $X_i^K(J)$ say thats there exists an epistemically possible world such that in that world a property belongs to J iff it is motivationally salient and holds of x. That is, J is a motivationally salient property package consistent with the agent's knowledge. Similarly, $X_i^B(J)$ says that J is a motivationally salient property package consistent with the agent's belief. Notice that this definition is correct, because $X_i^B(w) := \{J \subseteq \mathbb{P} \mid \exists w' \in W(w \sim_i w' \wedge \forall w'' \in W(w' \leq_i w'' \Rightarrow \exists w''' \in W(w'' \leq_i w''' \wedge r_i(w''', x) = J)))\}$ and $\hat{B}_i = \neg B_i \neg = \neg K_i \Diamond_i \Box_i \neg = \hat{K}_i \Box_i \Diamond_i$.

For pessimistic agents, one can use maximin and leximin preferences. An agent with maximin preference considers the worst motivationally salient property packages corresponding to x consistent with his knowledge/belief and the worst one corresponding to y, and compares these in determining whether he prefers x to y. An agent with leximin preference does the same, but, whenever these property packages are equally bad, he moves on to the second worst package and so on. Thus, in order to syntactically define maximin and leximin preferences, we need to be able to express the fact that an alternative is the worst motivationally salient property package consistent with the agent's knowledge or belief, the second worst, et cetera:

Definition 4.51. Let $i \in \mathcal{N}$, $x \in X$, $J \subseteq \mathbb{P}$ and $m \in \mathbb{N}$. We recursively define the abbreviations $\min_m^{x,i,K}(J)$ and $\min_m^{x,i,B}(J)$, which intuitively denote the fact that J is the m-th worst motivationally salient property packages corresponding to x consistent with agent i's knowledge/belief.
For the knowledge-based variant, we recursively define $\min_m^{x,i,K}(J)$ as follows:

- $\min_0^{x,i,K}(J) := X_i^K(J) \wedge \bigwedge_{J' \subseteq \mathbb{P}}(X_i^K(J') \rightarrow J' \succeq_i J)$
- Suppose $\min_j^{x,i,K}$ is defined for all $j \leq m$.
  $\min_{m+1}^{x,i,K}(J) := X_i^K(J) \wedge \bigwedge_{J' \subseteq \mathbb{P}}((X_i^K(J') \wedge \neg \bigvee_{j \leq m} \min_j^{x,i,K}(J')) \rightarrow J' \succeq_i J)$

For the belief-based variant, we recursively define $\min_m^{x,i,B}(J)$ as follows:

- $\min_0^{x,i,B}(J) := X_i^B(J) \wedge \bigwedge_{J' \subseteq \mathbb{P}}(X_i^B(J') \rightarrow J' \succeq_i J)$
- Suppose $\min_j^{x,i,B}$ is defined for all $j \leq m$.
  $\min_{m+1}^{x,i,B}(J) := X_i^B(J) \wedge \bigwedge_{J' \subseteq \mathbb{P}}((X_i^B(J') \wedge \neg \bigvee_{j \leq m} \min_j^{x,i,B}(J')) \rightarrow J' \succeq_i J)$

These definitions consider the set of all motivationally salient property packages corresponding to x consistent with the agent's knowledge/belief and say that a package J in that set is the m-th minimal, if the meta-level preferences of the agent are such that all other packages in that set that are not j-th minimal with $j < m$, are at least as good as J.

With these abbreviations in place, it is possible to define knowledge and belief-based maximin and leximin preferences:

Definition 4.52. Let $i \in \mathcal{N}$ and $x, y \in X$. *Maximin preference* is defined as follows:

- **Knowledge-Based Maximin**:
  $Pref_{i,K}^{mmin}(x,y) := \bigwedge_{J\subseteq\mathbb{P}} \bigwedge_{J'\subseteq\mathbb{P}} (\min_0^{x,i,K}(J) \wedge \min_0^{y,i,K}(J') \rightarrow J \succeq_i J')$
- **Belief-Based Weak Maximin**:
  $Pref_{i,B}^{mmin}(x,y) := \bigwedge_{J\subseteq\mathbb{P}} \bigwedge_{J'\subseteq\mathbb{P}} (\min_0^{x,i,B}(J) \wedge \min_0^{y,i,B}(J') \rightarrow J \succeq_i J')$

Definition 4.53. Let $i \in \mathcal{N}$ and $x, y \in X$. Let $\mathbb{P} := \{P_1, ..., P_k\}$. *Leximin preference* is defined as follows:

- **Knowledge-Based Leximin**: For knowledge-based leximin, we recursively define $Pref_i^m(x,y)$ for all $m \leq 2^k$ as follows:
  - $Pref_i^0(x,y) := \bigwedge_{J\subseteq\mathbb{P}} \bigwedge_{J'\subseteq\mathbb{P}} (\min_0^{x,i,K}(J) \wedge \min_0^{y,i,K}(J') \rightarrow J \succeq_i J')$
  - Suppose $Pref_i^j(x,y)$ is defined for all $j \leq m$.
    $Pref_i^{m+1}(x,y) := (Pref_i^m(x,y) \wedge \neg Pref_i^m(y,x)) \vee (Pref_i^m(x,y) \wedge Pref_i^m(y,x) \ \wedge$
    - $(\bigvee_{J\subseteq\mathbb{P}} \bigvee_{J'\subseteq\mathbb{P}} (\min_{m+1}^{x,i,K}(J) \wedge \min_{m+1}^{y,i,K}(J') \wedge J \succeq_i J') \ \vee$
    - $(\bigvee_{J\subseteq\mathbb{P}} \min_{m+1}^{x,i,K}(J) \wedge \neg \bigvee_{J'\subseteq\mathbb{P}} \min_{m+1}^{y,i,K}(J')))$

  Knowledge-based leximin is defined as $Pref_{i,K}^{lmin}(x,y) := Pref_i^{2^k}(x,y)$.
- **Belief-Based Leximin**: For belief-based leximin, we recursively define $Pref_i^m(x,y)$ for all $m \leq 2^k$ as follows:
  - $Pref_i^0(x,y) := \bigwedge_{J\subseteq\mathbb{P}} \bigwedge_{J'\subseteq\mathbb{P}} (\min_0^{x,i,B}(J) \wedge \min_0^{y,i,B}(J') \rightarrow J \succeq_i J')$
  - Suppose $Pref_i^j(x,y)$ is defined for all $j \leq m$.
    $Pref_i^{m+1}(x,y) := (Pref_i^m(x,y) \wedge \neg Pref_i^m(y,x)) \vee (Pref_i^m(x,y) \wedge Pref_i^m(y,x) \ \wedge$
    - $(\bigvee_{J\subseteq\mathbb{P}} \bigvee_{J'\subseteq\mathbb{P}} (\min_{m+1}^{x,i,B}(J) \wedge \min_{m+1}^{y,i,B}(J') \wedge J \succeq_i J') \ \vee$
    - $(\bigvee_{J\subseteq\mathbb{P}} \min_{m+1}^{x,i,B}(J) \wedge \neg \bigvee_{J'\subseteq\mathbb{P}} \min_{m+1}^{y,i,B}(J')))$

  Belief-based leximin is defined as $Pref_{i,B}^{lmin}(x,y) := Pref_i^{2^k}(x,y)$.

Notice that for knowledge and belief-based maximin, the syntactic definitions pick out the worst motivationally salient property package corresponding to x consistent with the agent's knowledge/belief and the worst one corresponding to y and compares these. The definitions of leximin preference are rather involved and recursively defined. The base case corresponds to normal maximin preference. At step $m+1$, the agent prefers alternative x to alternative y in two cases. Firstly, he does if he strictly preferred x to y at an earlier step. Secondly, he does if x and y are equally good up to step m and step $m+1$ does not tilt the balance in favour of y. The balance is not tilted in favour of y if (i) the $m+1$-th minimal motivionally salient property package corresponding to x consistent with the agent's knowledge/belief is at least as good as the one corresponding to y or (ii) if such a package exists for x and not for y. The last clause means that x and y perform equally well until step m, but that there exists better options for x whereas there do not exist better options for y.[8] The maximum number of motivationally salient property packages that correspond to some alternative and are consistent with the agent's knowledge/belief equals $|\mathcal{P}(P)| = 2^k$. Therefore, our recursive definition goes up to this stage.

[8] For a thorough explanation of this situation, see Section 4.3.1.

Optimistic agents can be modelled with maximax and leximax preferences. An agent with maximax preferences considers the best motivationally salient property packages that correspond to the alternatives in question and that are consistent with his knowledge/belief and compares these. Leximax preference is the optimistic variant of leximin, i.e. if the best property packages for two alternatives are the same, the agent moves to the second best, and so forth. In order to define maximax and leximax preference syntactically, we introduce abbreviations denoting the fact that a set $J \subseteq \mathbb{P}$ is the the m-th best motivationally salient property package corresponding to some alternative $x \in X$ consistent with the agent's knowledge/belief:

Definition 4.54. Let $i \in \mathcal{N}$, $x \in X$, $J \subseteq \mathbb{P}$ and $m \in \mathbb{N}$. We recursively define the abbreviations $\max_m^{x,i,K}(J)$ and $\max_m^{x,i,B}(J)$, which intuitively denote the fact that J is the m-th best motivationally salient property package corresponding to x consistent with the agent's knowledge/belief.
For the knowledge-based variant, we recursively define $\max_m^{x,i,K}(J)$ as follows:

- $\max_0^{x,i,K}(J) := X_i^K(J) \wedge \bigwedge_{J' \subseteq \mathbb{P}} (X_i^K(J') \rightarrow J \succeq_i J')$
- Suppose $\max_j^{x,i,K}$ is defined for all $j \leq m$.
  $\max_{m+1}^{x,i,K}(J) := X_i^K(J) \wedge \bigwedge_{J' \subseteq \mathbb{P}} ((X_i^K(J') \wedge \neg \bigvee_{j \leq m} \max_j^{x,i,K}(J')) \rightarrow J \succeq_i J')$

For the belief-based variant, we recursively define $\min_m^{x,i,B}(J)$ as follows:

- $\max_0^{x,i,B}(J) := X_i^B(J) \wedge \bigwedge_{J' \subseteq \mathbb{P}} (X_i^B(J') \rightarrow J \succeq_i J')$
- Suppose $\max_j^{x,i,B}$ is defined for all $j \leq m$.
  $\max_{m+1}^{x,i,B}(J) := X_i^B(J) \wedge \bigwedge_{J' \subseteq \mathbb{P}} ((X_i^B(J') \wedge \neg \bigvee_{j \leq m} \max_j^{x,i,B}(J')) \rightarrow J \succeq_i J')$

With these abbreviations in place, we can define knowledge and belief-based maximax and leximax preference in a similar way as maximin and leximin preference. Notice that if two alternatives x and y perform equally well up to some point and there are other options for x and not for y, then x is preferred in the case of leximin but y in the case of leximax preference. This is because leximin preference orders the outcomes from worst to best, whereas leximax orders them from best to worst.[9] Hence, we get the following definitions:

Definition 4.55. Let $i \in \mathcal{N}$ and $x, y \in X$. *Maximax preference* is defined as follows:

- **Knowledge-Based Maximax**:
  $Pref_{i,K}^{mmax}(x, y) := \bigwedge_{J \subseteq \mathbb{P}} \bigwedge_{J' \subseteq \mathbb{P}} (\max_0^{x,i,K}(J) \wedge \max_0^{y,i,K}(J') \rightarrow J \succeq_i J')$
- **Belief-Based Weak Maximax**:
  $Pref_{i,B}^{mmax}(x, y) := \bigwedge_{J \subseteq \mathbb{P}} \bigwedge_{J' \subseteq \mathbb{P}} (\max_0^{x,i,B}(J) \wedge \max_0^{y,i,B}(J') \rightarrow J \succeq_i J')$

Definition 4.56. Let $i \in \mathcal{N}$ and $x, y \in X$. Let $\mathbb{P} := \{P_1, ..., P_k\}$. *Leximax preference* is defined as follows:

- **Knowledge-Based Leximax**: For knowledge-based leximax, we recursively define $Pref_i^m(x, y)$ for all $m \leq 2^k$ as follows:
  - $Pref_i^0(x, y) := \bigwedge_{J \subseteq \mathbb{P}} \bigwedge_{J' \subseteq \mathbb{P}} (\max_0^{x,i,K}(J) \wedge \max_0^{y,i,K}(J') \rightarrow J \succeq_i J')$
  - Suppose $Pref_i^j(x, y)$ is defined for all $j \leq m$.
    $Pref_i^{m+1}(x, y) := (Pref_i^m(x, y) \wedge \neg Pref_i^m(y, x)) \vee (Pref_i^m(x, y) \wedge Pref_i^m(y, x) \ \wedge$

[9] For a thorough explanation of this situation, see Section 4.3.1.

$*$ $(\bigvee_{J\subseteq\mathbb{P}}\bigvee_{J'\subseteq\mathbb{P}}(\max_{m+1}^{x,i,K}(J)\wedge\max_{m+1}^{y,i,K}(J')\wedge J\succeq_i J')\vee$

$*$ $(\neg\bigvee_{J\subseteq\mathbb{P}}\max_{m+1}^{x,i,K}(J)\wedge\bigvee_{J'\subseteq\mathbb{P}}\max_{m+1}^{y,i,K}(J')))$

Knowledge-based leximax is defined as $Pref_{i,K}^{lmax}(x,y) := Pref_i^{2^k}(x,y)$.

- **Belief-Based Leximax**: For belief-based leximax, we recursively define $Pref_i^m(x,y)$ for all $m \leq 2^k$ as follows:
  - $Pref_i^0(x,y) := \bigwedge_{J\subseteq\mathbb{P}}\bigwedge_{J'\subseteq\mathbb{P}}(\max_0^{x,i,B}(J)\wedge\max_0^{y,i,B}(J')\rightarrow J\succeq_i J')$
  - Suppose $Pref_i^j(x,y)$ is defined for all $j \leq m$.
    $Pref_i^{m+1}(x,y) := (Pref_i^m(x,y)\wedge\neg Pref_i^m(y,x))\vee(Pref_i^m(x,y)\wedge Pref_i^m(y,x)\ \wedge$

    $*$ $(\bigvee_{J\subseteq\mathbb{P}}\bigvee_{J'\subseteq\mathbb{P}}(\max_{m+1}^{x,i,B}(J)\wedge\max_{m+1}^{y,i,B}(J')\wedge J\succeq_i J')\ \vee$

    $*$ $(\neg\bigvee_{J\subseteq\mathbb{P}}\max_{m+1}^{x,i,B}(J)\wedge\bigvee_{J'\subseteq\mathbb{P}}\max_{m+1}^{y,i,B}(J')))$

  Belief-based leximax is defined as $Pref_{i,B}^{lmax}(x,y) := Pref_i^{2^k}(x,y)$.

The following theorem shows that the above syntactic preference definitions are correct:

Theorem 4.57. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i\in\mathcal{N}}$ be a model for preference formation, $w \in W$, $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. Let $Pref_{i,\alpha(i)}^{\pi(i)}(x,y)$ be such that $\alpha(i) \in \{B, K\}$ and $\pi(i) \in \{wd, mmin, lmin, mmax, lmax\}$ and let $x\mathbf{R}_i^w y$ be the corresponding semantic notion. The following holds: $M, w \models Pref_{i,\alpha(i)}^{\pi(i)}(x,y)$ iff $x\mathbf{R}_i^w y$.

Proof. Left to the reader. □

In conclusion, Theorem 4.57 shows that the language of LPF is expressive enough to encode the different ways of modelling preference discussed in Section 4.3.1. As a consequence, we can enrich the language of LPF with preference predicates without losing completeness if we assume that the preference type of an agent is fixed throughout the model, i.e. if in all possible worlds the agent determines his preferences in the same way.

More precisely, one can add an attitude function $\alpha : \mathcal{N} \rightarrow \{B, K\}$ and a procedure function $\pi : \mathcal{N} \rightarrow \{wd, mmin, lmin, mmax, lmax\}$ to the parameters of the logic, where $\alpha(i)$ says whether i's preferences are knowledge or belief-based and $\pi(i)$ denotes the procedure used by i to form his preferences. For all $i \in \mathcal{N}$ and all $x, y \in \mathcal{X}$, one can then add the formula $Pref_{i,\alpha(i)}^{\pi(i)}(x,y)$ to the language of LPF. The truth condition of this type of formulas is as follows:

$$M, w \models Pref_{i,\alpha(i)}^{\pi(i)}(x,y) \text{ iff } x\mathbf{R}_i^w y$$

where $\mathbf{R}_i^w$ is the semantic notion corresponding agent i's preference type. Soundness and completeness are ensured by extending the proof system of LPF with the following axiom for all $i \in \mathcal{N}$ and all $x, y \in \mathcal{X}$:

$$Pref_{i,\alpha(i)}^{\pi(i)}(x,y) \leftrightarrow [\text{definition}]$$

where [definition] is the syntactic preference definition corresponding to agent i's preference type.

For instance, suppose agent i has knowledge-based weak dominance preference. Then we get the following truth condition: $M, w \models Pref_{i,K}^{wd}(x,y)$ iff for all $w' \in W$ such that $w \sim_i w'$ it holds that

$r_i(w', x) \succeq_i^{w'} r_i(w', y)$. To the proof system of LPF, we add the the following axiom: $Pref_{i,K}^{wd}(x, y) \leftrightarrow K_i(\bigwedge_{J \subseteq \mathbb{P}} \bigwedge_{J' \subseteq \mathbb{P}} (\bigwedge_{P \in J} (S_i P \wedge Px) \wedge \bigwedge_{P \notin J} \neg(S_i P \wedge Px)) \wedge (\bigwedge_{P \in J'} (S_i P \wedge Py) \wedge \bigwedge_{P \notin J'} \neg(S_i P \wedge Py)) \rightarrow J \succeq_i J')$.

As mentioned, by doing such a thing for all the agents and all combinations of alternatives, we obtain a sound and complete logic. However, since preferences can be defined as abbreviations, we have decided, for reasons of mathematical simplicity, not to add preference predicates to the language of LPF.

Chapter 5

Democratic Deliberation

Although deliberative democracy is an umbrella term for many different political and philosophical theories, the essence of all of them is that deliberation plays a central role in political decision making. This chapter is about modelling this most crucial facet of deliberative democracy. The preference formation models from the previous chapter give a realistic account of how agents form their preferences in terms of their knowledge/belief, their motivational state and the properties that hold of the alternatives. An important aspect of deliberation is that people justify their preferences to one another. Therefore, a proper way to model democratic deliberation is with a model transformer for the preference formation models.

Section 5.1 introduces such a model transformer and discusses its relationship with deliberative democracy. This model transformer is an adaptation of the one used in Chapter 3. In Section 5.2, we formally extend the logic of preference formation with a dynamic modality for deliberation and obtain the logic of democratic deliberation (LDD), of which we prove soundness and completeness.

5.1 Formalising Democratic Deliberation

This section is concerned with formalising democratic deliberation. In Section 5.1.1, we discuss the process of deliberation in a deliberative democracy and introduce a model transformer that is able to model this process in one go. Section 5.1.2 shows that two important claims from the literature on deliberative democracy, namely the claim that deliberation can induce preference change and the claim that deliberation leads to a better understanding among the agents, can be modelled in our framework.

5.1.1 The Model Transformer for Democratic Deliberation

According to all theories of deliberative democracy, deliberation should play a central role in the political realm and is crucial for political decision making. This does not mean, however, that all citizens are required to participate in deliberative processes.[1] As Gutmann and Thompson (2002: 29) point out, "most deliberative democrats ... do not insist that ordinary citizens regularly take part in public deliberations, and most favor some form of representative democracy. On these version of the theory, citizens rely on their representatives to do the deliberating for them". In fact, Gutmann and Thompson (1996: 131) believe that "from a deliberative perspective representation is not only necessary but also desirable", because the number of people who can have a conversation together is limited. Thus, deliberative democracy is consistent with representation and acknowledges that most deliberative processes take place within political institutions.

[1] See, for instance, Cohen (1989) and Gutmann & Thompson (1996, 2002).

The fact that democratic deliberation mostly occurs in political institutions is important for formalising this process. In democratic institutions, there often is a strict protocol for political debates. And in many cases that protocol amounts to speaking in turns. Consider, for instance, parliamentary debates in the Netherlands. Each political party appoints one of their members of parliament as a representative for the party with respect to the issue at stake. Before the debate starts, these representatives ask for a certain amount of speaking time. The debate itself consists of either one or two terms, depending on the complexity, importance and salience of the decision at hand. During each term, each party gets one turn in which its representative can justify the party's viewpoint and ask questions to the government. Although a speaker might be interrupted during his turn by other members of parliament, the essence of the deliberation protocol is that parties – or rather, their representatives – speak in turn. In summary, the deliberation protocols in many political institutions amount to speaking in turns and in his turn an agent justifies his (party's) views and preferences. Often, he does this by sharing his knowledge about and his perspective on the issue at stake. Therefore, we introduce the following model transformer which models the fact that an agent shares both his hard information and his motivational state, i.e. which models what happens models during an agent's turn:

Definition 5.1. Let $\mathcal{N}$ be a finite set of agents and $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i\in\mathcal{N}}$ a model for preference formation. Let $j \in \mathcal{N}$. The *updated model after agent j's turn* is defined as $M^{!j} = (W^{!j}, \sim_i^{!j}, \leq_i^{!j}, S_i^{!j}, \succeq_i^{!j}, r^{!j})_{i\in\mathcal{N}}$, where

- $W^{!j} := W$,
- $\sim_i^{!j} := \sim_i \cap \sim_j$,
- $\leq_i^{!j} := \leq_i \cap \ \sim_i^{!j}$,
- for all $w \in W^{!j}$, $S_i^{!j}(w) := S_i(w) \cup S_j(w)$,
- $\succeq_i^{!j} := \succeq_i$ and
- $r^{!j} = r$.

Firstly, notice that this model transformer models the fact that agent j tells all he knows. The updated epistemic indistinguishability relation of agent i is his old one intersected with the one of agent j. In other words, agent i combines his hard information with j's. Secondly, notice that this model transformer reflects the fact that agent j shares his perspective or motivational state. The new motivational state of agent i consists of the properties that were already motivationally salient to him before and of the properties that are motivationally salient for agent j. In other words, agent i takes the perspective of agent j into account. This is in line with deliberative democracy, because, as Gutmann and Thompson (2002) argue, people have to justify their preferences on the basis of mutually acceptable reasons. A reason is mutually acceptable if it can, in principle, be accepted by other people. That is, even though people might disagree on the relative importance of the reasons or properties that play a role in the decision at hand, they have to acknowledge the intrinsic worth of the perspectives of their co-deliberators. Therefore, it makes sense that an agent's new motivational state combines his old one with the one of the agent who has spoken.

For reasons of mathematical simplicity, Section 3.1 introduced a model transformer that was capable of modelling the entire deliberation process in one go. Here, we do the same:

Definition 5.2. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i\in\mathcal{N}}$ be a model for preference formation. The *updated model after deliberation* is defined as $M^! = (W^!, \sim_i^!, \leq_i^!, S_i^!, \succeq_i^!, r^!)_{i\in\mathcal{N}}$, where

- $W^! := W$,
- $\sim_i^! := \bigcap_{j\in\mathcal{N}} \sim_j$,
- $\leq_i^! := \leq_i \cap \ \sim_i^!$,

- for all $w \in W^!$, $S_i^!(w) := \bigcup_{j \in \mathcal{N}} S_j(w)$,
- $\succeq_i^! := \succeq_i$ and
- $r^! = r$.

Observation 5.3. Let $\mathcal{N} = \{1, ..., n\}$ be a finite set of agents and $f : \mathcal{N} \to \mathcal{N}$ be a permutation among the agents. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a model for preference formation. Then $M^! = (M^{!f(1)})...)^{!f(n)}$.

Notice that the model transformer for deliberation models that all agents share both their hard information and their motivational state. Furthermore, the situation after deliberation is the same one as the one that is obtained when all agents speak in turn. Thus, [!] is an elegant way to model democratic deliberation.

As the model transformer for preference formation models transforms the epistemic indistinguishability and the plausibility relations of the agents in the same way as the model transformer from Chapter 3, all the formal results from Section 3.2 apply here as well. In particular, whenever the common prior assumption holds, [!] realises distributed knowledge and belief and ensures that the epistemic doxastic state of all the agents is the same:

Observation 5.4. Let M be a common prior model for preference formation and $M^!$ the updated model after deliberation. The following hold:

- For all $i \in \mathcal{N} : \sim_i^! = \sim_{\mathcal{N}}$.
- For all $i \in \mathcal{N} : \leq_i^! = \sim_{\mathcal{N}}$.
- If M is standard, then for all $i \in \mathcal{N} : \to_i^! = \to_{\mathcal{N}}$.

Proof. For the proofs, see Propositions 3.5, 3.7 and 3.13 respectively. □

5.1.2 Deliberative Democracy and [!]

The previous section introduced a model transformer for democratic deliberation in line with the theory of deliberative democracy. This section shows that that two important claims for the philosophical literature on deliberative democracy can be formalised in our framework. We start by showing that deliberation can induce preference change. Afterwards, we show that deliberation leads to a better understanding among the agents.

In a deliberation, people might change their preferences because they are confronted with new information and new perspectives.[2] Precisely because, as Dryzek and List (2003: 1) point out, "deliberation involves discussion in which individuals are amenable to scrutinizing and changing their preferences in the light of persuasion", deliberative democracy regards people's preferences as dynamic. In our framework, preferences might change as a result of [!]. Firstly, [!] transforms the epistemic indistinguishability relations, which reflects the fact that agents learn new information. As the beliefs of the agents are defined in terms of their plausibility and their hard information, their doxastic state changes as well. Secondly, [!] transforms the motivational states, which reflects the fact that agents are confronted with new perspectives. Recall that in our framework the preferences of an agent are determined by his epistemic doxastic state, his motivational state or perspective and the properties that hold of the alternatives. As [!] changes both the epistemic indistinguishability relations and the motivational states, our framework is able to model preference change due to new information and new perspectives, as the following example shows:

[2] This claim is made by all deliberative democrats. See, for instance, Bohman (1998), Cohen (1989), Elster (1986), Habermas (1996) and Gutmann & Thompson (1996, 2002).

Example. *Deliberating about the energy policy*

Suppose three politicians are deliberating about which energy policy the government should fund: wind energy (w), nuclear energy (n) or shale gas (s). For simplicity, assume that only the following three properties play a role in the decision at hand:

- P_1: sustainability
- P_2: low energy price
- P_3: short implementation time

Thus, $\mathbb{P} := \{P_1, P_2, P_3\}$. In the actual world w, exactly the following properties are true of the alternatives:

- Wind energy: P_1 and P_3.
- Nuclear energy: P_2.
- Shale gas: P_2 and P_3.

Suppose agent 1 is a representative of the green party. Sustainability is so important for him that this is his only motivationally salient property. Agent 2 is extremely worried about the upcoming energy shortage. Therefore, it is of the utmost importance to him that the alternative the government will fund has a short implementation time. Additionally, a low energy price is motivationally salient for him. The motivational state of agent 3 consists of all the properties that play a role in the decision at hand. Thus, we get the following motivational states in the actual world w:

- $S_1(w) = \{P_1\}$.
- $S_2(w) = \{P_2, P_3\}$.
- $S_3(w) = \mathbb{P}$.

Suppose, furthermore, that the meta-level prefrences of the agents are as follows:

- Let $\succeq_1^w$ be the linear order that is induced lexicographically by $P_1 \gg_1^w P_3 \gg_1^w P_2$.
- Let $\succeq_2^w$ be the linear order that is induced lexicographically by $P_3 \gg_2^w P_1 \gg_2^w P_2$.
- Let $\succeq_3^w$ be the linear order that is induced lexicographically by $P_2 \gg_3^w P_1 \gg_3^w P_3$.

In addition, assume that agents 1 and 2 are well-informed. That is, they possess hard information about which properties hold of which alternatives. Agent 3 is correctly informed with respect to properties P_2 and P_3, but believes that both wind energy and nuclear energy are sustainable, for he has mistakenly equated sustainability with low CO_2-emissions. Under these assumptions, the preferences of the agents before deliberation are as follows, where $\mathbf{P}_i^w$ and $\mathbf{I}_i^w$ denote strict preference and indifference respectively:[3]

- $w\mathbf{P}_1^w s\mathbf{I}_1^w n$.
- $s\mathbf{P}_2^w w\mathbf{P}_2^w n$.
- $n\mathbf{P}_2^w s\mathbf{P}_2^w w$.

Agent 1 prefers wind energy because it is the only sustainable option. As the other options are not sustainable, he is indifferent between them. Agent 2 prefers wind energy and shale gas to nuclear energy, because both have a low implementation time. Moreover, he prefers shale gas to wind energy because the former gives a low energy price. Notice that we have not taken sustainability into account, since this property does not belong to the motivational state of the agent. For agent 3, low energy prices have top priority. Therefore, wind energy is his least preferred option. He prefers nuclear energy over shale gas, because he (falsely) believes that nuclear energy is sustainable whereas shale gas is not.

[3]More formally, let $x, y \in \mathcal{X}$ and let $x\mathbf{R}_i^w y$ denote the fact that agent i weakly prefers alternative x to y at world w. Then $x\mathbf{P}_i^w y := x\mathbf{R}_i^w y \wedge \neg(y\mathbf{R}_i x)$ and $x\mathbf{I}_i^w y := x\mathbf{R}_i^w y \wedge y\mathbf{R}_i^w x$.

Suppose that deliberation takes place and, thus, that the agents share their hard information and their perspectives. After deliberation, the motivational state of all the agents equals $\mathbb{P}$, because the agents have to take each other's perspectives into account. Furthermore, all the agents have correct information about which properties hold of which alternatives, because this was distributed knowledge. More specifically, agent 1 takes the implementation time and the energy price into account. He comes to prefer shale gas to nuclear energy, because of its short implementation time. Options with a short implementation time are interesting for him, because they can can function as a bridge between a non-sustainable present and a sustainable future. Agent 2 takes into account sustainability as a result of deliberation. Because he is worried about the energy shortage, he has become convinced that sustainable options are better than non-sustainable ones, because sustainable energy sources cannot be depleted. Thus, he prefers wind energy to shale gas. Lastly, agent 3 learns in the deliberation that nuclear energy is not sustainable, because sustainability is not just about low CO_2-emissions, but also about the safety of the workers and waste reduction. Now that he has the right information, he prefers shale gas to nuclear energy. In summary, the preferences after deliberation are as follows:

- $w\mathbf{P}_1^w s\mathbf{P}_1^w n$.
- $w\mathbf{P}_2^w s\mathbf{P}_2^w n$.
- $s\mathbf{P}_2^w n\mathbf{P}_2^w w$.

Notice that the preferences of all the agents have changed as a result of deliberation. The preferences of agents 1 and 2 have changed due to the confrontation with new perspectives, whereas the preferences of agent 3 have changed because of new information.[4]

Proponents of deliberative democracy praise deliberation not only because it might induce preference change, but also because it causes participants to better understand one another. As Knight & Johnson (1994: 289) argue, successful deliberation induces "common understandings of what is at stake in a given political conflict". In our formal framework, this is reflected by the fact that [!] leads to a common motivational state. For all $i, j \in \mathcal{N}$ and all $w \in W$, it is the case that after deliberation $S_i^!(w) = S_j^!(w) = \bigcup_{i\in\mathcal{N}} S_i(w)$. In other words, after deliberation an agent considers a property or reason to be motivationally salient iff the other agents do as well. Thus, the perspective of the agents upon the issue at stake is the same. Furthermore, deliberative democrats often claim that each agent's perspective should be taken seriously. The following proposition shows that this is reflected in our formal framework:

[4]For a complete formalisation of this example, let $\mathcal{N} = \{1, 2, 3\}$ be the set of agents, $\mathcal{X} = \{a, b, c\}$ the set of alternatives and $\mathbb{P} = \{P_1, P_2, P_3\}$ the set of properties. Let the preference of all the agents be defined as belief-based weak dominance. The preference formation model $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i\in\mathcal{N}}$ corresponding to our example is defined as follows:

- $W = \{w, w'\}$ and w is the actual world
- $\sim_1 = \sim_2 = \{(w, w), (w', w')\}$ and $\sim_3 = W \times W$.
- $\leq_1 = \leq_2 = \{(w, w), (w', w')\}$ and $\leq_3 = \{(w, w), (w', w'), (w, w')\}$.
- $S_1(w) = S_1(w') = \{P_1\}, S_2(w) = S_2(w') = \{P_2, P_3\}$ and $S_3(w) = S_3(w') = \mathbb{P}$.
- The meta-level preferences of the agent are defined as follows:
  - Agent 1: $\mathbb{P} \succeq_1^w \{P_1, P_3\} \succeq_1^w \{P_1, P_2\} \succeq_1^w \{P_1\} \succeq_1^w \{P_2, P_3\} \succeq_1^w \{P_3\} \succeq_1^w \{P_2\} \succeq_1^w \emptyset$ and $\succeq_1^{w'} = \succeq_1^w$.
  - Agent 2: $\mathbb{P} \succeq_2^w \{P_1, P_3\} \succeq_2^w \{P_2, P_3\} \succeq_2^w \{P_3\} \succeq_2^w \{P_1, P_2\} \succeq_2^w \{P_1\} \succeq_2^w \{P_2\} \succeq_2^w \emptyset$ and $\succeq_2^{w'} = \succeq_2^w$.
  - Agent 3: $\mathbb{P} \succeq_3^w \{P_1, P_2\} \succeq_3^w \{P_2, P_3\} \succeq_3^w \{P_2\} \succeq_3^w \{P_1, P_3\} \succeq_3^w \{P_1\} \succeq_3^w \{P_3\} \succeq_3^w \emptyset$ and $\succeq_3^{w'} = \succeq_3^w$.
- The valuation function is defined as follows:
  - $r(w, a) = r(w', a) = \{P_1, P_3\}$.
  - $r(w, b) = \{P_2\}$ and $r(w', b) = \{P_1, P_2\}$.
  - $r(w, c) = r(w', c) = \{P_2, P_3\}$.

Proposition 5.5. $\models S_iP \to [!] \bigwedge_{j \in \mathcal{N}} S_jP$.

Proof. Left to the reader. □

This proposition says that if a certain property or reason is motivationally salient to one of the agents before deliberation, it is motivationally salient for all of them after deliberation. In other words, each agent's perspective is taken seriously.

In conclusion, our framework reflects two important claims from deliberative democracy about the power of deliberative processes, namely (i) that deliberation might induce preference change and (ii) that it leads to a better understanding among the agents. Furthermore, the model transformer [!] for deliberation has been developed in close connection with the deliberative democratic literature. Therefore, we may conclude that the models for preference formation together with the dynamic operator [!] are a good mathematical presentation of democratic deliberation.

5.2 The Logic of Democratic Deliberation

The language of the logic of democratic deliberation (LDD) is obtained by extending the language of LPF with a dynamic operator [!] for deliberation:

Definition 5.6. Let $\mathcal{N}$ be a finite set of agents, $\mathcal{X}$ a finite set of alternatives and $\mathbb{P}$ a finite set of properties or reasons. The set $\mathcal{L}_{LDD}$ of formulas of φ of LDD is defined recursively:

$$\varphi ::= Px \mid J \succeq_i J' \mid S_iP \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid \Box_i\varphi \mid D_K\varphi \mid D_\Box\varphi \mid [!]\varphi$$

where $x \in X$, $P \in \mathbb{P}$, $i \in \mathcal{N}$ and $J, J' \subseteq \mathbb{P}$. We define $\bot := Px \wedge \neg Px$ and $\top := \neg\bot$. The Boolean connectives $\vee$ and $\to$ are defined in terms of $\neg$ and $\wedge$ in the standard manner. The duals of the modal operators are defined in the following way: $\hat{K}_i := \neg K_i \neg$ and $\Diamond_i := \neg\Box_i\neg$. Belief is defined as $B_i := K_i\Diamond_i\Box_i$ and distributed belief as $D_B := D_K D_\Diamond D_\Box$, where $D_\Diamond := \neg D_\Box \neg$.

As [!] models deliberation, the intended meaning of $[!]\varphi$ is as follows: After deliberation, that is after all the agents have shared their information and their perspectives, φ is the case. The models of LDD are common prior models for preference formation:

Definition 5.7. A tuple $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ is a model for deliberation iff M is a common prior model for preference formation.

As we have seen in Section 3.3, the common prior assumption is needed for proving completeness. The truth conditions for LDD are those of LPF extended with one for formulas of the form $[!]\varphi$:

Definition 5.8. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a model for deliberation, $w \in W$ and $i \in \mathcal{N}$. The *satisfaction* relation $\models$ between pairs (M, w) and formulas $\varphi \in \mathcal{L}_{LDD}$ is defined as in Definition 4.43, extended with the following clause:

- $M, w \models [!]\varphi$ iff $M^!, w \models \varphi$.

A formula φ is *valid* if for all $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ and all $w \in W$, we have $M, w \models \varphi$. We denote validity by $\models \varphi$.

In order to prove that LDD is sound and complete, we extend the proof system of LPF with the following reduction axioms:

Definition 5.9. The proof system of LDD (notation: Λ_{LDD}) is the proof system Λ_{LPF} extended with reduction axioms 2-8 of Definition 3.17 together with the following:

1. $[!]Px \leftrightarrow Px$
2. $[!](J \succeq_i J') \leftrightarrow (J \succeq_i J')$
3. $[!]S_iP \leftrightarrow \bigvee_{j \in \mathcal{N}} S_jP$

The first reduction axiom says that ontic facts about which properties hold of which alternatives do not change due to deliberation. The second reduction axiom expresses that deliberation does not change the meta-level preferences of the agents. The third reduction axiom says that after deliberation property P is motivationally salient for agent i iff P was motivationally salient for at least one of the agents before deliberation.

The following theorem shows that the reduction axioms are sound on models for deliberation:

Theorem 5.10. Λ_{LDD} is sound with respect to the class of all models for deliberation.

Proof. The proof system of LDD extends the one of LPF with reduction axioms. According to Corollary 4.46, Λ_{LPF} is sound with respect to the class of all common prior models for preference formation. As these are precisely the models for deliberation, it follows that the axioms and inference rules of LPF are sound on all models for deliberation. Therefore, it remains to be shown that the reduction axioms are sound.

Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a model for deliberation and let $w \in W$. From the definition of the models for deliberation, it follows that the tuple $F = (W, \sim_i, \leq_i)_{i \in \mathcal{N}}$ is a common prior frame. Theorem 3.18 shows that reduction axioms 2-8 of Definition 3.17 are sound on all common prior frames. Hence, they are sound on all models for deliberation. Only reduction axioms 1-3 of Definition 5.9 remain. The soundness of axiom 1 follows from the fact that [!] does not change ontic facts about the world. Moreover, as [!] does not change the meta-level preferences of the agents, reduction axiom 2 is sound. For reduction axiom 3, suppose that $M, w \models [!]S_iP$. Using the definition of $M^!$ and the truth conditions gives us: $M, w \models [!]S_iP$ iff $M^!, w \models S_iP$ iff $P \in S_i^!(w) = \bigcup_{j \in \mathcal{N}} S_j(w)$. This holds iff there exists a $j \in \mathcal{N}$ such that $P \in S_j(w)$, which holds iff $M, w \models \bigvee_{j \in \mathcal{N}} S_jP$. □

Similarly to in Section 3.3, we can prove that LPF and LDD are co-expressive. In order to do this, we start by defining a map from $\mathcal{L}_{LDD}$ to $\mathcal{L}_{LPF}$:

Definition 5.11. We define a map $! : \mathcal{L}_{LDD} \to \mathcal{L}_{LPF}$ by recursion on the complexity of φ in the same way as in Definition 3.19 where the base case is replaced by the following:

- $!(Px) := Px$
- $!(J \succeq_i J') := (J \succeq_i J')$
- $!(S_iP) := \bigvee_{j \in \mathcal{N}} S_jP$

Observation 5.12. The map $! : \mathcal{L}_{LDD} \to \mathcal{L}_{LPF}$ is well-defined.

Using the reduction axioms, one can prove that:

Proposition 5.13. For every formula $\varphi \in \mathcal{L}_{LPF}$, the following holds: $\vdash_{LPF} [!]\varphi \leftrightarrow !(\varphi)$.

Proof. The proof is by induction on the complexity of φ. □

We now define a translation function which translates formulas in $\mathcal{L}_{LDD}$ to formulas in $\mathcal{L}_{LPF}$:

Definition 5.14. Let $tr : \mathcal{L}_{LDD} \to \mathcal{L}_{LPF}$ be the *translation function* defined by recursion on the complexity of φ in the same way as in Definition 3.22 where the base case is replaced by the following:

- $tr(Px) := Px$
- $tr(J \succeq_i J') := (J \succeq_i J')$
- $tr(S_i P) := \bigvee_{j \in \mathcal{N}} S_j P$

Observation 5.15. The translation function $tr : \mathcal{L}_{LDD} \to \mathcal{L}_{LPF}$ is well-defined.

Observation 5.16. For all $\varphi \in \mathcal{L}_{LPF}$, it holds that $tr(\varphi) = \varphi$.

We can now show that LDD proves that every formula $\varphi \in \mathcal{L}_{LDD}$ is equivalent to a formula $\psi \in \mathcal{L}_{LPF}$:

Proposition 5.17. For all $\varphi \in \mathcal{L}_{LDD}$, it holds that $\vdash_{LDD} \varphi \leftrightarrow tr(\varphi)$.

Proof. The proof is similar to the one of Proposition 3.25. □

Consequently, LPF and LDD are co-expressive:

Theorem 5.18. EDL and EDLD are co-expressive.

Proof. The proof is similar to the one of Theorem 3.26. □

With the co-expressivity in place, we can prove the completeness of Λ_{LDD}:

Theorem 5.19. Λ_{LDD} is weakly complete with respect to the class of all models for deliberation.

Proof. According to Proposition 2.38, it suffices to show that any consistent LDD-formula is satisfiable on some model for deliberation. Let φ be a consistent LDD-formula. According to Proposition 5.17, there exists an LPF-formula ψ such that $\vdash_{EDLD} \varphi \leftrightarrow \psi$. According to Theorem 4.47, there exists a common prior model for preference formation, i.e. a model for deliberation, $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ and a world $w \in W$ such that $M, w \models \psi$. Since $\models_{LDD} \varphi \leftrightarrow \psi$, it follows that $M, w \models \varphi$. Hence Λ_{LDD} is weakly complete with respect to the class of all models for deliberation. □

Chapter 6

Democratic Deliberation and Single-Peakedness

In the previous chapters, we have developed a formal framework for democratic deliberation in line with the philosophical literature on deliberative democracy. This framework consists of preference formation models together with a model transformer for deliberation. The preference formation models allow us to define the preferences of the agents in terms of their epistemic doxastic state, their motivational state and the properties that hold of the alternatives. Deliberation itself is modelled as a process in which all agents share their information and their perspective. Our framework models two important claims from the deliberative democratic literature, namely that deliberation might cause agents to change their preferences due to the confrontation with new perspectives and new information and that it leads to a better understanding of the values or reasons that motivate other people. As our formal framework for democratic deliberation incorporates essential ideas and assumptions from deliberative democracy, it can be used to formally investigate philosophical claims about the power of deliberative processes. This chapter focuses on the claim that deliberation before voting is a solution to Arrow's impossibility theorem, as it ensures single-peaked preferences profiles, which can be aggregated unproblematically.

Section 6.1 defines the notion of single-peakedness and discusses its role in social choice theory and deliberative democracy. Section 6.2 introduces common conceptual space models, which play a crucial role in formalising the claim that deliberation ensures unproblematic aggregation. In Section 6.3, we prove that, in specific circumstances, democratic deliberation provides a solution to Arrow's impossibility theorem and we discuss how these formal results contribute to the philosophical debate on deliberation and single-peakedness.

6.1 Introduction to Single-Peakedness

This section forms an introduction to the social choice theoretic and the deliberative democratic literature on single-peakedness. Section 6.1.1 formally introduces this idea and shows why aggregation is unproblematic in the case of single-peaked preferences profiles. In Section 6.1.2, we dive into the philosophical discussion about the claim that deliberation induces single-peaked preferences.

6.1.1 Single-Peakedness and Social Choice Theory

As mentioned in the introduction, Condorcet's paradox can be generalised to Arrow's theorem, which says that it is impossible to aggregate individual preference orderings into a collective ordering if we want the aggregation method to satisfy certain democratic and seemingly intuitive principles. This section starts with some terminology and a formal statement of Arrow's theorem. Afterwards, it introduces the idea of single-peakedness and shows that the impossibility result is avoided when preferences

are single-peaked.

Let $\mathcal{N} := \{1, ..., n\}$ be a finite set of individuals and $\mathcal{X} := \{x_1, ..., x_m\}$ with $|\mathcal{X}| \geq 3$ be a finite set of alternatives. For all $i \in \mathcal{N}$, let $\mathbf{R}_i \subseteq \mathcal{X} \times \mathcal{X}$ be a linear order denoting agent i's preferences. In other words, $x\mathbf{R}_i y$ represents the fact that agent i considers alternative x to be at least as good as y. Furthermore, we use the following abbreviation: $x\mathbf{P}_i y := x\mathbf{R}_i y \wedge \neg(y\mathbf{R}_i x)$.

Let $L(\mathcal{X})$ denote the set of all linear orders on $\mathcal{X}$. A *preference profile* $(\mathbf{R}_i)_{i\in\mathcal{N}} \in L(\mathcal{X})^n$ is a vector of linear orders, where $\mathbf{R}_i$ is the preference ordering of agent i. A *social welfare function*, or aggregation method, is a function $F : D \to L(\mathcal{X})$ with $D \subseteq L(\mathcal{X})^n$, that assigns to each preference profile a linear order on $\mathcal{X}$, which is to be interpreted as the collective ordering. For a preference profile $(\mathbf{R}_i)_{i\in\mathcal{N}}$, we let $\mathbf{R} := F((\mathbf{R}_i)_{i\in\mathcal{N}})$ denote the collective ordering. In other words, $x\mathbf{R}y$ means that alternative x is at least as good as y from the perspective of the group. Furthermore, we use the following abbreviation: $x\mathbf{P}y := x\mathbf{R}y \wedge \neg(y\mathbf{R}x)$.

Arrow (1963) has proposed some minimal conditions that any aggregation method should satisfy:

Definition 6.1. Let $F : D \to L(\mathcal{X})$ be a social welfare function. *Arrow's conditions* are the following:

- Universal Domain (U): $D = L(\mathcal{X})^n$.
- Weak Pareto Principle (P): For all preference profiles $(\mathbf{R}_i)_{i\in\mathcal{N}} \in D$, $x\mathbf{P}_i y$ for all $i \in \mathcal{N}$ implies $x\mathbf{P}y$.
- Independence of Irrelevant Alternatives (IIA): For any two preference profiles $(\mathbf{R}_i)_{i\in\mathcal{N}}$, $(\mathbf{R}'_i)_{i\in\mathcal{N}} \in D$, $\{i \in \mathcal{N} \mid x\mathbf{R}_i y\} = \{i \in \mathcal{N} \mid x\mathbf{R}'_i y\}$ implies $x\mathbf{R}y$ iff $x\mathbf{R}'y$.
- Non-dictatorship (ND): There does not exists an $i \in \mathcal{N}$ such that for all $(\mathbf{R}_i)_{i\in\mathcal{N}} \in D$, $\mathbf{R} = \mathbf{R}_i$.

The universal domain conditions captures the idea that any logically possible preference profile is a valid input for the social welfare function. In other words, the aggregation method always returns a collective ordering no matter what the preference orderings of the individuals are. The weak Pareto principle says that if all individuals strictly prefer some alternative over another, then so should society. Independence of irrelevant alternatives says that the collective ranking of two alternatives should only depend on the way the individuals rank those two alternatives. Notice that many positional scoring rules, such as Borda count, violate this constraint, because the collective ranking of alternatives x and y is determined by looking at the entire linear order of the individuals.[1] Lastly, non-dictatorship says that it should not be the case that, no matter what the individual preferences are, the social ordering always equals the preference ordering of some fixed individual $i \in \mathcal{N}$. In order to be democratic, an aggregation method should surely satisfy the weak Pareto principle and non-dictatorship. Furthermore, universal domain and independence of irrelevant alternatives also seem intuitive. Arrow (1963) has shown, however, that no social welfare function can satisfy all these conditions simultaneously:

Theorem 6.2. (Arrow's Theorem) There does not exists a social welfare function F that satisfies (U), (P), (IIA) and (ND).[2]

In other words, Arrow's theorem says that any aggregation method will violate universal domain, the weak Pareto condition, independence of irrelevant alternatives or non-dictatorship. As these conditions are normatively appealing, a pessimistic interpretation of Arrow's result is that aggregation of preferences is arbitrary and, therefore, might lead to political instability. A more optimistic interpretation is that weakening any of the Arrovian axioms might provide an escape route from the

[1] Recall that Borda count is a positional scoring rule that works as follows. All individuals submit a preference ordering and each option gets a number of points equal to the number of alternatives which it beats. The collective ordering is determined by counting how many points each alternative has received in total. Notice that this scoring rule does not satisfy independence of irrelevant alternatives, because the relative position of x and y in the collective ordering is determined by considering the complete preference ordering of all the individuals.

[2] The original proof can be found in Arrow (1963). This proof is quite verbose, so more mathematically inclined readers might prefer the proof of Sen (1986).

impossibility result. One extensively studied way is relaxing the universal domain condition. The research field occupied with domain restrictions studies under which restrictions Arrow's impossibility result disappears.[3] The best-known example of such a domain restriction is single-peakedness.

The notion of single-peakedness was introduced by Black (1948). The preference profile of a group of agents satisfies single-peakedness if there exists a natural ordering on the alternatives such that every individual has a favourite option – his peak – on this dimension and the further an option is removed from an agent's peak, the less he likes it.[4] Consider, for example, the left-right ordering in politics and suppose that agent i's peak is a centre-left party. Intuitively, he should then prefer a centre party to a right party to an extreme right party. This is precisely what single-peakedness expresses. Single-peakedness does not say anything about options on different sides of the peak. Consider, for instance, the preferences of a group of agents over the legal drinking age and suppose that agent i believes that the legal drinking age should be 18. Obviously, agent i will choose 17 if he has to choose between 16 and 17. Single-peakedness does not require, however, that he prefers 17 over 20 even though 17 is closer to 18 than 20.

Mathematically, Black (1948) defines single-peakedness as follows: Let $\gg$ be a strict linear order on $\mathcal{X}$ and $\succeq$ a linear order on $\mathcal{X}$. An alternative $p \in \mathcal{X}$ is a peak iff $p \succeq x$ for all $x \in \mathcal{X}$. $\succeq$ is single-peaked with respect to $\gg$ iff there exists an alternative $p \in \mathcal{X}$ such that p is the peak and for all $x, y \in \mathcal{X}$ the following holds:

- If $x \gg y \gg p$, then $y \succeq x$.
- If $p \gg y \gg x$, then $y \succeq x$.

Figure 1 gives an example of three preferences orderings over the set of alternatives $\mathcal{X} = \{a, b, c, d, e\}$ that are single-peaked with respect to $a \gg b \gg c \gg d \gg e$:[5]

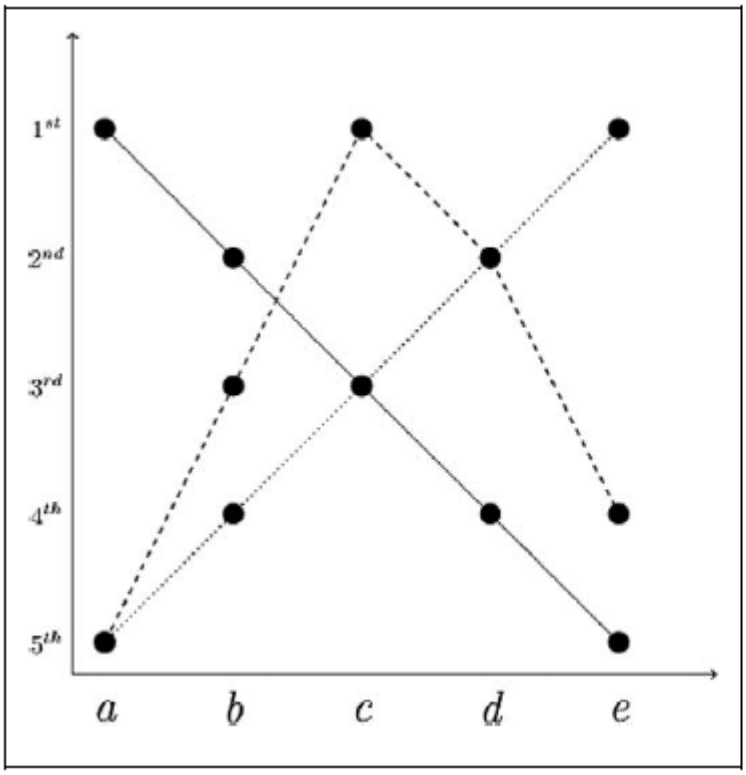

Figure 1: Examples of preference orderings that are single-peaked with respect to $a \gg b \gg c \gg d \gg e$.

[3] For a nice overview of domain restrictions in social choice theory, see Gaertner (2001).
[4] For a discussion of single-peakedness, see Riker (1982): 124-128.
[5] This figure is copied from Ottonelli & Porello (2013): 73.

From this figure, some features of single-peakedness become immediately clear. First, notice that all individuals have a peak and that options are less good the further they are removed from the peak. Moreover, notice that a linear order is single-peaked with respect to $\gg$ iff it is single-peaked with respect to the opposite order $\ll$. Intuitively, this makes sense as, for instance, it does not matter whether we speak of the left-right or the right-left ordering in politics. Lastly, notice that in order for someone's preferences to be single-peaked with respect to a certain linear order $\gg$, it must be the case that his or her least preferred option is either at the top or the bottom of $\gg$.

In this thesis, we work with the following definition of single-peakedness, which is equivalent to Black's:

Definition 6.3. Let $\mathcal{X}$ be a finite set of alternatives and $\gg$ a strict linear order on $\mathcal{X}$. A linear order $\succeq \subseteq \mathcal{X} \times \mathcal{X}$ is *single-peaked* with respect to $\gg$ iff for any three alternatives $x, y, z \in \mathcal{X}$ the following holds: $x \gg y \gg z$ or $x \ll y \ll z$ implies $\neg(z \succeq x \succeq y)$.

Proposition 6.4. Let $\mathcal{X}$ be a finite set of alternatives and $\gg$ a strict linear order on $\mathcal{X}$. Let $\succeq$ be a linear order on $\mathcal{X}$. Consider the following two conditions:

1. There exists an alternative $p \in \mathcal{X}$ such that (i) for all $x \in \mathcal{X} : p \succeq x$ and (ii) for all $x, y \in \mathcal{X}$ the following holds:
   - If $x \gg y \gg p$, then $y \succeq x$.
   - If $x \ll y \ll p$, then $y \succeq x$.
2. For all $x, y, z \in \mathcal{X}$, $x \gg y \gg z$ or $x \ll y \ll z$ implies $\neg(z \succeq x \succeq y)$.

Then $\succeq$ satisfies condition 1 iff $\succeq$ satisfies condition 2.

Proof. For the left-to-right direction, suppose condition 1 holds. Let x, y, z be such that $x \gg y \gg z$. Because $\gg$ is a strict linear order, it holds that $x \neq y \neq z$. The following cases can be distinguished:

- $p \gg y$. In this case, we have $p \gg y \gg z$. Condition 1 gives us that $y \succeq z$. Suppose, for contradiction, that $z \succeq x \succeq y$. By transitivity, $z \succeq y$. As $\succeq$ is a linear order, it follows from $y \succeq z$ and $z \succeq y$ that $y = z$, which is a contradiction.
- $p = y$. Suppose, for contradiction, that $z \succeq x \succeq y$. Because y is the peak, it must be the case that $y \succeq x$. From the fact that $\succeq$ is a linear order, $x \succeq y$ and $y \succeq x$ imply $x = y$, which is a contradiction.
- $y \gg p$. In this case, we have $x \gg y \gg p$. Condition 1 gives us that $y \succeq x$. Suppose, for contradiction, that $z \succeq x \succeq y$. Similarly to the previous item, we can conclude that $x = y$, which is a contradiction.

No matter whether the peak lies left of y, right of y or is equal to y, assuming that $z \succeq x \succeq y$ leads to a contradiction. Thus, $\neg(z \succeq x \succeq y)$. The case where $x \ll y \ll z$ is similar.

For the right-to-left direction, let $p \in \mathcal{X}$ be the peak. Notice that such a peak always exists, since $\succeq$ is a linear order over the finite set $\mathcal{X}$. Let $x, y \in \mathcal{X}$ be such that $x \gg y \gg p$. From condition 2 it follows that $\neg(p \succeq x \succeq y)$. Hence either $\neg(p \succeq x)$ or $\neg(x \succeq y)$. The former cannot be the case, as p is the peak. Thus, it must be the case that $\neg(x \succeq y)$. From the totality of $\succeq$ it now follows that $y \succeq x$. The case where $x, y \in \mathcal{X}$ are such that $x \ll y \ll p$ is similar. □

From the definition of single-peakedness, it immediately follows that single-peakedness is preserved when we go to suborders:

Observation 6.5. Let $\mathcal{X}$ be a finite set of alternatives and $\gg$ a strict linear order on $\mathcal{X}$. Let $\succeq$ be a linear order on $\mathcal{X}$ such that $\succeq$ is single-peaked with respect to $\gg$. Let $X \subseteq \mathcal{X}$. Then $\succeq \cap (X \times X)$ is single-peaked with respect to $\gg \cap (X \times X)$.

Notice that the concept of single-peakedness is vacuous when we consider only one preference ordering on $\mathcal{X}$, because any linear order is single-peaked with respect to itself. Single-peakedness only becomes interesting when we have multiple preference orderings, because we can ask ourselves whether there exists a linear order $\gg$ on $\mathcal{X}$ with respect to which all preference orderings are single-peaked. This motivates the following definition:

Definition 6.6. Let $\mathcal{N} = \{1, ..., n\}$ be a finite set of agents and $\mathcal{X}$ a finite set of alternatives. A preference profile $(\mathbf{R}_i)_{i\in\mathcal{N}} \in L(\mathcal{X})^n$ is called *single-peaked* if there exists a strict linear order $\gg$ on $\mathcal{X}$ such that for all $i \in \mathcal{N}$, $\mathbf{R}_i$ is single-peaked with respect to $\gg$.

Black (1948) has shown that if a preference profile is single-peaked, there exists a Condorcet winner. That is, there exists an alternative that beats every other in pairwise majority voting. This result is generalised in the following theorem, which shows that in the case of single-peaked preference profiles pairwise majority voting always yields acyclic collective preferences:[6]

Theorem 6.7. Let $\mathcal{N} = \{1, ..., n\}$ with n being odd be a finite set of agents and $\mathcal{X} = \{x_1, ..., x_m\}$ a finite set of alternatives. Let $(\mathbf{R}_i)_{i\in\mathcal{N}} \in L(\mathcal{X})^n$ be a preference profile. If $(\mathbf{R}_i)_{i\in\mathcal{N}}$ is single-peaked, then pairwise majority voting yields a linear order on the alternatives.[7]

Proof. Recall that $x\mathbf{R}y$ denotes the fact that the group prefers alternative x to y.[8] The definition of pairwise majority voting is as follows: $x\mathbf{R}y$ iff $|\{i \in \mathcal{N} \mid x\mathbf{R}_i y\}| \geq |\{i \in \mathcal{N} \mid y\mathbf{R}_i x\}|$. Totality follows from the definition of pairwise majority voting. For antisymmetry, let $x, y \in \mathcal{X}$ be such that $x\mathbf{R}y$ and $y\mathbf{R}x$. By definition, it follows that $|\{i \in \mathcal{N} \mid x\mathbf{R}_i y\}| = |\{i \in \mathcal{N} \mid y\mathbf{R}_i x\}|$. Notice that for all $i \in \mathcal{N}$, $\mathbf{R}_i$ is a linear order. Hence if two alternatives are different, then the agent strictly prefers one to the other. Combining this with the fact that the number of individuals is odd, it must be the case that $x = y$. For transitivity, it suffices to show that no cycles occur. Suppose, for contradiction, that majority voting generates a cycle. We start by proving the following claim:

> Claim: If pairwise majority voting yields a cycle of length $k \leq m$, then there is also a cycle of length 3.
> Proof: Let $y_1\mathbf{R}y_2\mathbf{R}...\mathbf{R}y_k\mathbf{R}y_1$, with $y_i \in \mathcal{X}$ for all $i \leq k$ and $y_i \neq y_j$ if $i \neq j$, be a cycle of k-alternatives generated by pairwise majority voting. We have $y_1\mathbf{R}y_2$ and $y_2\mathbf{R}y_3$. By totality of pairwise majority voting, either $y_3\mathbf{R}y_1$ or $y_1\mathbf{R}y_3$. If the former, we have found a cycle of length 3, namely $y_1\mathbf{R}y_2\mathbf{R}y_3\mathbf{R}y_1$. If the latter, we have $y_1\mathbf{R}y_3$ and $y_3\mathbf{R}y_4$. By totality, either $y_4\mathbf{R}y_1$, in which case we have the cycle $y_1\mathbf{R}y_3\mathbf{R}y_4\mathbf{R}y_1$, or $y_1\mathbf{R}y_4$. We continue this procedure until we reach a cycle of length 3 or until we reach $y_1\mathbf{R}y_{k-1}$. If $y_1\mathbf{R}y_{k-1}$, then we get the cycle $y_1\mathbf{R}y_{k-1}\mathbf{R}y_k\mathbf{R}y_1$.

Thus, there exists a cycle of three alternatives. Let these alternatives be denoted by x, y and z. Notice that there are two possible cycle types:

1. $x\mathbf{R}y\mathbf{R}z\mathbf{R}x$.
2. $x\mathbf{R}z\mathbf{R}y\mathbf{R}x$.

As there are three alternatives, there are $3! = 6$ possible preference profiles. Correspondingly, we define the following sets:

[6] Even though single-peakedness is a sufficient condition for avoiding cycles in the case of pairwise majority voting, it is by no means the only one. A more general condition that suffices is Sen's (1966) triplewise value-restriction. This condition says that for any three alternatives $x, y, z \in \mathcal{X}$ there exists an $x^* \in \{x, y, z\}$ and a value $v \in \{$"best", "middle", "worst"$\}$ such that none of the agents assigns value v to x^*. Elsholtz and List (2002) have introduced a condition that is both necessary and sufficient for the avoidance of cycles. Although this condition is even more general than Sen's, Elsholtz and List acknowledge that their condition is intuitively unappealing.

[7] The condition that the number of agents is odd is an innocuous assumption made for technical reasons. As Riker (1982) argues, harmless ties might occur if the number of agents is even. These do not, however, undermine the importance of Black's result.

[8] The proof of this theorem is an adaptation of the proof from Elsholtz and List (2005).

- $A := \{i \in \mathcal{N} \mid x\mathbf{R}_i y\mathbf{R}_i z\}$
- $B := \{i \in \mathcal{N} \mid x\mathbf{R}_i z\mathbf{R}_i y\}$
- $C := \{i \in \mathcal{N} \mid y\mathbf{R}_i x\mathbf{R}_i z\}$
- $D := \{i \in \mathcal{N} \mid y\mathbf{R}_i z\mathbf{R}_i x\}$
- $E := \{i \in \mathcal{N} \mid z\mathbf{R}_i x\mathbf{R}_i y\}$
- $F := \{i \in \mathcal{N} \mid z\mathbf{R}_i y\mathbf{R}_i x\}$

Notice that the following inequalities hold if we have a cycle of type 1:

1. $|A| + |B| + |E| > |C| + |D| + |F|$
2. $|A| + |C| + |D| > |B| + |E| + |F|$
3. $|D| + |E| + |F| > |A| + |B| + |C|$

Notice that this inequalities are strict ones, as the number of individuals is odd. By adding 1 and 2, 1 and 3 and 2 and 3 respectively, we obtain:

- $|A| > |F|$
- $|E| > |C|$
- $|D| > |B|$

Doing the same for a cycle of type 2 yields:

- $|B| > |D|$
- $|C| > |E|$
- $|F| > |A|$

By assumption, $(\mathbf{R}_i)_{i\in\mathcal{N}}$ is single-peaked. Let $\gg \subseteq X \times X$ be the linear order with respect to which all the agent's preference orderings are single-peaked. Notice that if a preference profile is single-peaked with respect to some order $\gg$, it is also single-peaked with respect to the opposite order $\ll$. Thus, there are only three different cases:

1. $x \gg y \gg z$
2. $x \gg z \gg y$
3. $y \gg x \gg z$

We now show that all possibilities lead to contradiction. In case 1, it follows from the definition of single-peakedness that for all $i \in \mathcal{N} : \neg(z\mathbf{R}_i x\mathbf{R}_i y)$ and $\neg(x\mathbf{R}_i z\mathbf{R}_i y)$. Thus, $E = B = \emptyset$. Notice that our cycle is either of type 1 or of type 2. Therefore, it must be the case that $|E| > |C|$ or $|B| > |D|$. However, since $|\emptyset| = 0$, we have a contradiction. In case 2, it follows from the definition of single-peakedness that for all $i \in \mathcal{N} : \neg(y\mathbf{R}_i x\mathbf{R}_i z)$ and $\neg(x\mathbf{R}_i y\mathbf{R}_i z)$. Thus, $C = A = \emptyset$. This also gives rise to a contradiction. In case 3, we get that $F = D = \emptyset$ and again obtain a contradiction. Thus, there are no cycles. □

Since pairwise majority voting satisfies the Pareto condition, independence of irrelevant alternatives and non-dictatorship, we get the following result as a corollary:

Corollary 6.8. Let $\mathcal{N} = \{1, ..., n\}$ with n being odd be a finite set of agents and $\mathcal{X}$ a finite set of alternatives. Let $D \subseteq L(\mathcal{X})^n$ be the set of all single-peaked preference profiles. Then there exists a social welfare function $F : D \to L(\mathcal{X})^n$ that satisfies (P), (IIA) and (ND).

In other words, if we drop the universal domain condition and instead restrict the domain of social welfare functions to all single-peaked preferences profiles, Arrow's impossibility result is avoided.

6.1.2 Single-Peakedness and Deliberative Democracy

This section discusses the philosophical claim that deliberative democracy provides an escape route from Arrow's impossibility theorem. Some proponents of deliberative democracy, especially the earlier ones, argue that this is the case because deliberation tends to promote consensus by asking people's preferences to be justifiable in terms of rational and mutually acceptable reasons.[9] Or, as Elster (1986: 112) puts it:

> The core of the theory [of deliberative democracy] ... is rather than aggregating or filtering preferences, the political system should be set up with a view to changing them by public debate and confrontation. The input to the social choice mechanism would then not be raw, quite possibly selfish or irrational preferences ... but informed and other-regarding preferences. Or rather, there would not be any need for an aggregation mechanism, since a rational discussion would tend to produce unanimous preferences.

The idea of unanimous preferences is referred to be List (2002) as *substantive agreement*. Substantive agreement or unanimity is highly demanding and many other deliberative democrats acknowledge that this is too much to ask.[10]

A more commonly heard argument, proposed by both Miller (1992) and Dryzek and List (2003), is that if deliberation is able to promote meta-agreement, Arrow's impossibility result is circumvented. According to List (2002: 73), "individuals agree at meta-level to the extent that they agree on a common dimension in terms of which an issue is to be conceptualized. They may reach perfect agreement at meta-level while at the same time disagreeing substantively on what the most preferred option on that dimension is." For instance, suppose the national government has to decide between publically funded religious education, privately funded religious education or a complete ban on religious education. It might well be the case that politicians agree that the relevant issue dimension is the availability of religious education, without agreeing on the most preferred option on that dimension. In other words, there can be meta-agreement without substantive agreement. Meta-agreement has at least two advantages over substantive agreement. Firstly, it is less demanding because it expects "processes of political deliberation to produce agreement on what the relevant *questions* are rather than on what the *answers* should be" (List, 2002: 74).[11] Secondly, as meta-agreement does not ask people to agree on the answers, it leaves room for pluralism.

In situations where there is meta-agreement, Arrow's impossibility result does not apply as meta-agreement implies single-peakedness: "If different individuals agree on a common dimension along which each individual's preferences are systematically aligned in the requisite way, then the profile of preferences orderings across these individuals satisfies single-peakedness" (Dryzek & List, 2003: 14). Thus, the philosophical argument that democratic deliberation may provide a solution to Arrow's impossibility theorem can be broken down in two claims:

1. Democratic deliberation induces meta-agreement.
2. Meta-agreement implies single-peaked preferences profiles.

In fact, empirical evidence suggests that deliberation increases the degree of single-peakedness.[12] It should be noted, however, that this evidence does not say anything on whether this single-peakedness is achieved via meta-agreement. Meta-agreement is a sufficient, but not a necessary condition for single-peakedness, which is merely a formal condition on the set of preference orderings.

Thus far, agreement at meta-level is defined as agreement on a common dimension in terms of which an issue is conceptualised. Saying that this kind of meta-agreement leads to single-peaked preferences

[9] See, for instance, Elster (1986), Cohen (1989), Habermas (1996) and Bohman (1998).

[10] See, for instance, Gutmann & Thompson (1996, 2002) and Dryzek (2000).

[11] Italics in the original.

[12] See, for example, List et al. (2000, 2013) and Farrar et al. (2010). The degree of single-peakedness is defined as $|M|/|N|$, where N is the group of agents and $M \subseteq N$ is a largest subgroup whose preference profile is single-peaked.

seems to rely, as Ottonelli and Porello (2013: 75) argue, "on the fundamental intuition that once we have conceptualized the different options according to a single semantic dimension, public deliberation will also lead parties to align the options along one and the same geometrical dimension corresponding to the semantic dimension, and the parties will then choose their preferred options along that geometrical dimension."[13] However, as Aldred (2004) shows, this fundamental intuition is not right in all circumstances.

First of all, there might be semantic or issue dimensions that do not correspond to linear orders. For example, suppose that the electorate collectively bases their preference for a political party on the favourite colour of the party representative. Although a common semantic dimension has been identified, there need not be a geometrical dimension that corresponds to it. Furthermore, even if there is a geometric dimension corresponding to the semantic dimension in terms of which the options are conceptualised, preferences might not be aligned according to that dimension. For instance, suppose the government has to decide how many soldiers to send to a conflict area and they agree that this is the dimension of interest. A politician might prefer 1000 over 0 over 100 soldiers, because in the case of 100 soldiers the military operation is ineffective anyway. In this case, the agent has agreed to the common issue dimension, but his preferences are based on some other, hidden issue dimension, namely effectiveness.

These examples show that meta-agreement is not just about agreeing on a common issue dimension. This dimension should also be translatable into a geometrical dimension and the agents should really base their preferences on the dimension that was agreed upon. In line with these observations, List (2002, 2007) proposed to break down the hypothesis that deliberation induces meta-agreement into three subhypotheses:

1. "Group deliberation leads people to identify a single shared issue-dimension in terms of which the issue at stake is to be conceptualized."

2. "For a given issue-dimension, group deliberation leads people to agree on the position of each option on that dimension."

3. "Once an issue dimension has been identified as relevant, group deliberation leads each individual to determine a most preferred option on that dimension, with decreasing preference as options are increasingly distant from that most preferred position."

Subhypothesis 1 is called *normative meta-agreement*, as it asks the group to agree on the normative question which conceptual dimension is the (most) relevant one for the issue at stake. Normative meta-agreement is highly demanding, because often several issue dimensions are important for the issue at stake. Subhypothesis 2 is *factual meta-agreement*. As Ottonelli and Porello (2013: 81) argue, it requires agents to agree "on the facts about the policy alternatives that are relevant for positioning the options along the shared dimension". Subhypothesis 3 corresponds to *rationality meta-agreement*, because the preferences of the agents should indeed be based upon the commonly shared issue dimension and the common beliefs. Thus, in order for meta-agreement to lead to single-peaked preference profiles, one has to work with a rather thick notion of meta-agreement consisting of normative, factual and rationality meta-agreement. Because all these components have their own problems, Ottonelli and Porello (2013) are of the opinion that meta-agreement is not necessarily less demanding than substantive agreement. In addition, they argue that meta-agreement is hostile to pluralism because it excludes the possibility that agents prefer the same option for different reasons.

In conclusion, we can say that Dryzek and List (2003) and List (2002, 2007) argue that deliberation should seek to produce meta-agreement, because meta-agreement provides an escape from Arrow's impossibility result by guaranteeing single-peaked preference profiles.

[13] In the literature on deliberative democracy and single-peakedness, the term semantic dimension is used for an issue dimension in terms of which the issue at stake can be conceptualised. In other words, semantic dimensions are about the aspects or properties that play a role in the decision at hand. The term geometrical dimension is used to refer to a linear order on the set of alternatives.

6.2 Common Conceptual Space Models

In order to formally study the relationship between democratic deliberation and Arrow's impossibility result, we work with a special kind of models: common conceptual space models. Section 6.2.1 introduces these models and discusses their relation with meta-agreement. To get a better grasp on these models and to get an intuitive idea of what is needed for obtaining single-peaked preference profiles, we work out two examples in Section 6.2.2.

6.2.1 Common Conceptual Space Models and Meta-Agreement

In our framework, the philosophical claim that democratic deliberation induces single-peaked preferences can only be proven in common conceptual space models:

Definition 6.9. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a model for deliberation. M is a *common conceptual space model* iff the following holds:

- For all $w \in W$, there exists a strict linear order $\gg^w$ on $\mathcal{P}(\mathbb{P})$ such that for all $i \in \mathcal{N}$, $\succeq_i^w$ is single-peaked with respect to $\gg^w$.

Thus, a model for deliberation is a common conceptual space model if the meta-level preferences of the agents are single-peaked.

Common conceptual space models formalise the idea that on some subconscious level agents approach the issue at stake in the same way. In other words, there is a common conceptual space regarding the decision at hand. This is reflected by two assumptions that are made in our formal framework.

First, our framework assumes that there is implicit or potential agreement about the relevant reasons or properties that play a role in the decision at hand, as the set $\mathbb{P}$ of reasons or properties is fixed and agent-independent. Since the set of motivationally salient properties might differ from agent to agent, this agreement only exists on a subconscious level. This implicit agreement does, however, imply – as is reflected by the model transformer for deliberation – that if an agent introduces a reason or property that is motivationally salient for him, the other agents acknowledge that this property plays a role in the issue at stake. Thus, the implicit meta-agreement about $\mathbb{P}$ models the deliberative democratic idea that reasons should be mutually acceptable.

Second, the condition on common conceptual space models requires the agents' meta-level preferences to be single-peaked. In other words, it requires the agents to structure the property packages in the same way. Recall that these property packages can be thought of as justifications. Thus, the spirit of this assumption is that each agent has a most preferred justification/property package and that the other property packages/justifications are assessed solely on the basis of how far they are removed from the most preferred one. In principle, this assumption is rather intuitive when interpreted n-dimensionally. The problem, however, is that our formalisation requires one dimensionality. It is not clear to us why there should be a common, virtual dimension with respect to which all agents order their property packages, especially since often many independent properties or reasons play a role in political decisions. This assumption seems only likely when there already is some natural dimension accompanying the problem, for instance the left-right ordering in politics. We come back to this problem in Section 6.3.2. For now we assume that this assumption holds, as the one dimensionality problem is not specific to our framework, but problematic for the single-peakedness discussion in general.

According to List (2002: 74), normative meta-agreement is about having "a single shared issue dimension in terms of which an issue is to be conceptualized". That is exactly what these two assumptions capture, albeit implicitly or virtually. The condition on common conceptual space models requires the existence of a common dimension with respect to which the agents structure the justifications for their preferences and, thus, the issue at stake. This dimension is, however, subconscious or virtual. In reality, agents might not explicitly agree on this dimension, because they may not agree on

the properties that are motivationally salient in the decision at hand. Notice that for such a dimension to exist subconsciously, there should be implicit agreement about the set $\mathbb{P}$ of properties that play a role in the decision at hand. Therefore, the two discussed assumptions together capture the idea of *virtual normative meta-agreement.*

In conclusion, we can say that the assumption of a common conceptual space is much like the common prior assumption, because both capture the idea that agents are of the same kind. The common prior assumption says that agents are of the same kind doxastically speaking. At some (fictional) initial state, agents had the same plausibility. Differences in their current beliefs are solely due to differences in information. The common conceptual space assumption is about virtual normative meta-agreement, which says that agents are of the same kind in the way they conceptualise the issue at stake. Differences in their current conceptualisations are solely due to differences in their motivational state.

6.2.2 Examples

This section discusses two examples in order to give the reader an idea of the concrete situations that our framework can model and to give a feeling of which things deliberation needs to ensure to get single-peaked preference profiles.

Example. *Voting for a political party.*

Elections are on their way and there are three political parties that can be voted for. Party a is a left-wing party, b a centre party and c a right-wing party. For simplicity, we assume that there are only three voters and four important issues that play a role in the upcoming elections:

- P_1: social security
- P_2: public health care
- P_3: healthy treasury
- P_4: low taxes

Thus, $\mathbb{P} = \{P_1, P_2, P_3, P_4\}$. Not all combinations of these properties make sense. Therefore, we construct a set $\mathcal{J} \subseteq \mathcal{P}(\mathbb{P})$ of relevant property packages or justifications. Ideally, both the agents and the political parties would like to realise all these things, but everyone knows that is financially impossible. Suppose that it is common knowledge that each party focuses on precisely two of these properties, i.e. for all $J \in \mathcal{J} : |J| = 2$. Furthermore, the money for a good social security or public health care system has to come from taxes, i.e. for all $J \in \mathcal{J} : P_1 \in J$ or $P_2 \in J$ implies that $P_4 \notin J$. Lastly, if a party finds a strong social security system important it also has to be in favour of a strong public health care system, i.e. $\forall J \in \mathcal{J} : P_1 \in J$ implies $P_2 \in J$. Given these constraints, we get that $\mathcal{J} := \{\{P_1, P_2\}, \{P_2, P_3\}, \{P_3, P_4\}\}$.

Given the fact that a is a left-wing, b a centre and c a right-wing party exactly the following properties are true of each party in the actual world w:

- Party a: P_1 and P_2.
- Party b: P_2 and P_3.
- Party c: P_3 and P_4.

Suppose that the meta-level preferences of the agents are as follows:

- Agent 1: $\{P_1, P_2\} \succeq_1^w \{P_2, P_3\} \succeq_1^w \{P_3, P_4\}$.
- Agent 2: $\{P_2, P_3\} \succeq_2^w \{P_3, P_4\} \succeq_2^w \{P_1, P_2\}$.
- Agent 3: $\{P_3, P_4\} \succeq_3^w \{P_2, P_3\} \succeq_3^w \{P_1, P_2\}$.

Notice that we can model this example with a common conceptual space model, because the meta-level preference orderings of the agents are single-peaked with respect to $\{P_1, P_2\} \gg^w \{P_2, P_3\} \gg^w \{P_3, P_4\}$. This ordering can be thought of as the left-right ordering in politics.

Suppose that the motivational state of all agents is the same and that they consider all four properties to be motivationally salient. Furthermore, suppose that agents 1 and 2 are very interested in politics and, thus, well-informed. They know precisely which party focuses on which issues. Agent 3, however, is not interested in politics at all. He knows that party c is in favour of low taxes and a healthy treasury, but he has confused parties a and b. That is, he believes that both support public health care but that a wants a healthy treasury whereas b favours a strong social security system. Notice that all our agents are opinionated. Lastly, suppose that the agents preferences are belief-based. Recall that in opinionated models all the notions of belief-based preference collapse, hence it is unnecessary to specify which preference type the agents have.

With all of these things in place, it follows that the agents have the following preference orderings over the alternatives before deliberation:

- Agent 1: $a\mathbf{P}_1^w b\mathbf{P}_1^w c$.
- Agent 2: $b\mathbf{P}_2^w c\mathbf{P}_2^w a$.
- Agent 3: $c\mathbf{P}_3^w a\mathbf{P}_3^w b$.

Notice that since each alternative occurs at the bottom place once, there is no ordering with respect to which the preferences of the agents over the alternatives are single-peaked. In order to get single-peaked preferences profiles, it is necessary that agent 3 has the same information as agents 1 and 2. Thus, deliberation leads to single-peaked preferences if the agents share all their knowledge.[14]

[14]For a complete formalisation of this example, let $\mathcal{N} = \{1, 2, 3\}$ be the set of agents, $\mathcal{X} = \{a, b, c\}$ the set of alternatives and $\mathbb{P} = \{P_1, P_2, P_3, P_4\}$ the set of properties. Let the preferences of all the agents be defined as belief-based weak dominance. The common conceptual space model $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i\in\mathcal{N}}$ corresponding to our example is defined as follows:

- $W = \{w, w'\}$ and w is the actual world
- $\sim_1 = \sim_2 = \{(w, w), (w', w')\}$ and $\sim_3 = W \times W$.
- $\leq_1 = \leq_2 = \{(w, w), (w', w')\}$ and $\leq_3 = \{(w, w), (w', w'), (w, w')\}$.
- For all $i \in \mathcal{N} : S_i(w) = S_i(w') = \mathbb{P}$.
- The meta-level preferences of the agent are defined as follows:
  - Agent 1: $\{P_1, P_2\} \succeq_1^w \{P_2, P_3\} \succeq_1^w \{P_3, P_4\}$ and $\succeq_1^{w'} = \succeq_1^w$.
  - Agent 2: $\{P_2, P_3\} \succeq_2^w \{P_3, P_4\} \succeq_2^w \{P_1, P_2\}$ and $\succeq_2^{w'} = \succeq_2^w$.
  - Agent 3: $\{P_3, P_4\} \succeq_3^w \{P_2, P_3\} \succeq_3^w \{P_1, P_2\}$ and $\succeq_3^{w'} = \succeq_3^w$.

  One can easily extend the agents' meta-level preference to linear orders on the full $\mathcal{P}(\mathbb{P})$ by putting all the elements in $\mathcal{P}(\mathbb{P}) \setminus \mathcal{J}$ on the bottom of each agent's ordering. For example, for all $i \in \mathcal{N}$ we could put the following at the bottom: $\mathbb{P} \succeq_i^w \{P_1, P_2, P_3\} \succeq_i^w \{P_1, P_2, P_4\} \succeq_i^w \{P_1, P_3, P_4\} \succeq_i^w \{P_2, P_3, P_4\} \succeq_i^w \{P_1, P_3\} \succeq_i^w \{P_1, P_4\} \succeq_i^w \{P_2, P_4\} \succeq_i^w \{P_1\} \succeq_i^w \{P_2\} \succeq_i^w \{P_3\} \succeq_i^w \{P_4\} \succeq_i^w \emptyset$. Notice that it is important to do this in the same way for all the agents, because in that case their meta-level preferences remain single-peaked. In particular, they are single-peaked with respect to the extension of $\gg^w$ which is defined by putting the elements in $\mathcal{P}(\mathbb{P}) \setminus \mathcal{J}$ on the bottom of $\gg^w$ in the same way as was done for the individuals.
- The valuation function is defined as follows:
  - $r(w, a) = r(w', b) = \{P_1, P_2\}$.
  - $r(w, b) = r(w', a) = \{P_2, P_3\}$.
  - $r(w, c) = r(w', c) = \{P_3, P_4\}$.

Example. *Religious education.*[15]

The government has to make a decision about the availability and funding of religious eduction. Suppose there are three agents and three alternatives: publicly funded religious education (a), privately funded religious education (b) or a complete ban on religious education (c). The following reasons can be used by the agents to justify their preferences:

- P_1: All schools – no matter what their ideological viewpoint is – should get equal state funding.
- P_2: There is a constitutional right to religious education, hence it should be available.
- P_3: The separation of church and state entails that religious education is not a matter of the state.
- P_4: All children should receive the same, scientifically based education.

Thus, $\mathbb{P} = \{P_1, P_2, P_3, P_4\}$. Not all combinations of these reasons make sense. Therefore, we construct a set $\mathcal{J} \subseteq \mathcal{P}(\mathbb{P})$ of relevant property packages or justifications. Firstly, if someone is convinced that all schools, irrespective of their ideological viewpoint, should get equal state funding then he should also be of the opinion that religious eduction should be available: $\forall J \in \mathcal{J} : P_1 \in J \rightarrow P_2 \in J$. Secondly, if one is convinced that all children should get the same scientifically based education, he cannot at the same time argue for religious education: $\forall J \in \mathcal{J} : P_4 \in J \rightarrow P_1, P_2 \notin J$. Lastly, if one believes that religious eduction is not a state affair, he cannot at the same time argue that religious schools should get state funding $\forall J \in \mathcal{J} : P_3 \in J \rightarrow P_1 \notin J$. Given these constraints, we get that $\mathcal{J} = \{\{P_2\}, \{P_3\}, \{P_4\}, \{P_1, P_2\}, \{P_2, P_3\}, \{P_3, P_4\}\}$.

For each alternative, the agents can use the following reasons to justify their preferences in the actual world w:

- Alternative a: P_1 and P_2.
- Alternative b: P_2 and P_3.
- Alternative c: P_3 and P_4.

Suppose that the meta-level preferences of the agents are as follows:

- Agent 1: $\{P_1, P_2\} \succeq_1^w \{P_2, P_3\} \succeq_1^w \{P_2\} \succeq_1^w \{P_3, P_4\} \succeq_1^w \{P_3\} \succeq_1^w \{P_4\}$.
- Agent 2: $\{P_2, P_3\} \succeq_2^w \{P_3, P_4\} \succeq_2^w \{P_3\} \succeq_2^w \{P_1, P_2\} \succeq_2^w \{P_2\} \succeq_2^w \{P_4\}$.
- Agent 3: $\{P_3, P_4\} \succeq_3^w \{P_3\} \succeq_3^w \{P_4\} \succeq_3^w \{P_2, P_3\} \succeq_3^w \{P_1, P_2\} \succeq_3^w \{P_2\}$.

Notice that we can model this example with a common conceptual space model, because the meta-level preference orderings of the agents are single-peaked with respect to $\{P_2\} \gg^w \{P_1, P_2\} \gg^w \{P_2, P_3\} \gg^w \{P_3, P_4\} \gg^w \{P_3\} \gg^w \{P_4\}$. The middle part of this ordering contains property packages of two elements. The more a property package contains reasons in favour of the availability of religious education, the more it is to the left. At the ends of this ordering, we find the singletons. Notice that at the far ends, we find the most extreme reasons.

Suppose that agents 1 and 2 consider all reasons to be motivationally salient and that they know which reasons can be used in favour of which alternatives. The same holds for agent 3, except that the separation between church and state is not motivationally salient for him. Since the agents have full knowledge about which reasons can be used in favour of which alternatives, it does not matter whether we take belief or knowledge-based preferences. Also, we do not need to specify in which way they form their preferences because all the agents are opinionated.

The above information gives us the following preference profiles before deliberation:

- Agent 1: $a\mathbf{P}_1^w b\mathbf{P}_1^w c$.
- Agent 2: $b\mathbf{P}_2^w c\mathbf{P}_2^w a$.

[15] This example is an elaboration of Ottonelli & Porello (2013): 75.

- Agent 3: $c\mathbf{P}_3^w a\mathbf{P}_3^w b$.

Notice that since each alternative occurs at the bottom place once, there is no ordering on the alternatives with respect to which the preferences of the agents are single-peaked. In order to get single-peaked preference profiles, it is necessary that reason P_3 becomes motivationally salient for agent 3. Thus, deliberation leads to single-peakedness if it induces a common motivational state.[16]

Both examples show that having a common conceptual space does not necessarily ensure single-peakedness over the options. In the first example, the preferences were not single-peaked because the agents had different information. In the second example, the preferences were not single-peaked because the agents considered different properties to be motivationally salient. In general, a common conceptual space leads to single-peakedness over the alternatives if there is a common epistemic doxastic and a common motivational state.

6.3 Obtaining Actual Meta-Agreement

Section 6.3.1 contains the formal proofs of the fact that, under some circumstances, deliberation before voting ensures unproblematic aggregation and, thus, that democratic deliberation is a solution to Arrow's impossibility result. In Section 6.3.2, we interpret these formal results and discuss their connection to the existing literature on deliberative democracy and single-peakedness. In particular, we argue that deliberation before voting is useful because it induces actual meta-agreement. Lastly, Section 6.3.3 shows that variants of the theorems proven in Section 6.3.1 can be encoded as validities in the logic of democratic deliberation (LDD).

6.3.1 The Formal Results

This section formalises the following claim: If there is a common conceptual space and if, in addition, there is sufficient information about which properties hold of which alternatives or all the agents are opinionated, then deliberation before voting ensures acyclic collective preferences. This result is proven in two steps. First, we show under which formal conditions majority voting generates acyclic collective preferences. Afterwards, we show that in cases of sufficient information or opinionatedness, deliberation ensures that these formal conditions are met.

We start by proving the following lemma:

[16] For a complete formalisation of this example, let $\mathcal{N} = \{1, 2, 3\}$ be the set of agents, $\mathcal{X} = \{a, b, c\}$ the set of alternatives and $\mathbb{P} = \{P_1, P_2, P_3, P_4\}$ the set of properties. Let the preferences of all the agents be defined as belief-based weak dominance. The common conceptual space model $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i\in\mathcal{N}}$ corresponding to our example is defined as follows:

- $W = \{w\}$ and w is the actual world.
- $\sim_1 = \sim_2 = \sim_3 = W \times W$.
- $\leq_1 = \leq_2 = \leq_3 = W \times W$.
- $S_1(w) = S_2(w) = \mathbb{P}$ and $S_2(w) = \{P_1, P_2, P_4\}$.
- The meta-level preferences of the agent are defined as follows:
  - Agent 1: $\{P_1, P_2\} \succeq_1^w \{P_2, P_3\} \succeq_1^w \{P_2\} \succeq_1^w \{P_3, P_4\} \succeq_1^w \{P_3\} \succeq_1^w \{P_4\}$.
  - Agent 2: $\{P_2, P_3\} \succeq_2^w \{P_3, P_4\} \succeq_2^w \{P_3\} \succeq_2^w \{P_1, P_2\} \succeq_2^w \{P_2\} \succeq_2^w \{P_4\}$.
  - Agent 3: $\{P_3, P_4\} \succeq_3^w \{P_3\} \succeq_3^w \{P_4\} \succeq_3^w \{P_2, P_3\} \succeq_3^w \{P_1, P_2\} \succeq_3^w \{P_2\}$.

  Notice that just as in the previous example, we can extend the agents' meta-level preferences orderings to linear orders on the full $\mathcal{P}(\mathbb{P})$ without losing single-peakedness.
- The valuation function is defined as follows:
  - $r(w, a) = \{P_1, P_2\}$.
  - $r(w, b) = \{P_2, P_3\}$.
  - $r(w, c) = \{P_3, P_4\}$.

Lemma 6.10. Let $\mathcal{N} = \{1, ..., n\}$ with $n \in \mathbb{N}$ being odd be a set of agents, $\mathcal{X} = \{x_1, ..., x_m\}$ a finite set of alternatives and $\mathbb{P}$ a finite set of properties. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a common conceptual space model with $w \in W$ as the actual world, such that the following hold:

1. For all $i, j \in \mathcal{N}$, $S_i(w) = S_j(w) =: \mathcal{S}$.
2. The preference definitions of the agents are such that there exists a $w^* \in W$ such that $w \sim_{\mathcal{N}} w^*$ and for all $i \in \mathcal{N}$ and all $x, y \in \mathcal{X}$, it holds that $x \mathbf{R}_i^w y$ iff $r(w^*, x) \cap \mathcal{S} \succeq_i^{w^*} r(w^*, y) \cap \mathcal{S}$.

Then pairwise majority voting in the actual world w generates a total pre-order on the alternatives in $\mathcal{X}$.

Proof. First, notice that condition 2 says the preferences of the agents are determined by looking at the motivationally salient property packages corresponding to the alternatives in some world w^* and comparing these. We show that if one agent is indifferent between two alternatives in world w^*, all agents are:

> Claim 1: For all $i, j \in \mathcal{N}$, $\approx_i^{w^*} = \approx_j^{w^*} =: \approx_{w^*}$.
> Proof: Let $i, j \in \mathcal{N}$ be arbitrary. For the $\subseteq$-inclusion, let $x, y \in \mathcal{X}$ be such that $x \approx_i^{w^*} y$. By Definition 4.4, this holds iff $r_i(w^*, x) = r_i(w^*, y)$. Notice that $w \sim_{\mathcal{N}} w^*$ implies that $w \sim_i w^*$. As M is a model for deliberation, it follows that $S_i(w^*) = S_i(w) = \mathcal{S}$. Similarly, $S_j(w^*) = S_j(w) = \mathcal{S}$. Thus, $r_i(w^*, x) = r(w^*, x) \cap S_i(w^*) = r(w^*, x) \cap S_j(w^*) = r_j(w^*, x)$. Similarly, $r_i(w^*, y) = r_j(w^*, y)$. As a consequence, $r_j(w^*, x) = r_j(w^*, y)$ which holds iff $x \approx_j^{w^*} y$. Therefore, $\approx_i^{w^*} \subseteq \approx_j^{w^*}$. The proof of the $\supseteq$-inclusion is similar.

Notice that $\approx_{w^*}$ is an equivalence relation. Let $\mathcal{X}^\sim := \{[x_1], ..., [x_m]\}$, with $[x_k] := \{y \in \mathcal{X} \mid x_k \approx_{w^*} y\}$ for all $k \leq m$, be the partition of $\mathcal{X}$ into equivalence classes. Notice that for all $i \in \mathcal{N}$ and all $x \in [x]$, it holds that $r_i(w^*, x) = r(w^*, x) \cap \mathcal{S}$. Thus, the motivationally salient property packages corresponding to the alternatives within one equivalence class are the same and agent-independent. Let $\mathcal{J} := \{J \subseteq \mathbb{P} \mid \exists x \in \mathcal{X}(J = r(w^*, x) \cap \mathcal{S})\}$ be the set of all motivationally salient property packages that correspond to an equivalence class.

For all $i \in \mathcal{N}$, we define the preferences over equivalence classes as follows:

$$[x] \mathcal{R}_i^w [y] \text{ iff } r(w^*, x) \cap \mathcal{S} \succeq_i^{w^*} r(w^*, y) \cap \mathcal{S}$$

The reader may check that this definition is independent of the chosen representants. Notice that by condition 2, it is the case that $[x] \mathcal{R}_i^w [y]$ iff $x \mathbf{R}_i^w y$. Furthermore, notice that $\mathbf{R}_i^w$ is a total pre-order whereas $\mathcal{R}_i^w$ is a linear order, because two alternatives belong to the same equivalence class iff the agent is indifferent between them.

Since M is a common conceptual space model, there exists a linear order $\gg^{w^*}$ on $\mathcal{P}(\mathbb{P})$ such that for all $i \in \mathcal{N}$, $\succeq_i^{w^*}$ is single-peaked with respect to $\gg^{w^*}$. If we restrict ourselves to the motivationally salient property packages that are actually considered by the agents in determining their preferences, single-peakedness is preserved, as noted in Observation 6.5. Formally, for all $i \in \mathcal{N}$, $\succeq_i^{w^*} \cap (\mathcal{J} \times \mathcal{J})$ is single-peaked with respect to $\gg^{w^*} \cap (\mathcal{J} \times \mathcal{J}) =: \gg$.

$\mathcal{J}$ contains precisely the motivationally salient property packages that correspond to an equivalence class. Therefore, there exists a linear order $\gg^\sim$ on $\mathcal{X}^\sim$ with respect to which the agents' preferences over the equivalence classes are single-peaked. Formally, for all $i \in \mathcal{N}$, $\mathcal{R}_i^w$ is single-peaked with respect to $\gg^\sim$ where $\gg^\sim$ is defined as follows:

$$[x] \gg^\sim [y] \text{ iff } r(w^*, x) \cap \mathcal{S} \gg r(w^*, y) \cap \mathcal{S}$$

According to Theorem 6.7, it follows that pairwise majority voting on the equivalence classes results in a linear order $\mathcal{R}^w$ on $\mathcal{X}^\sim$. A total pre-order $\mathbf{R}^w$ over the alternatives, representing the collective preferences, is obtained by defining for all $x, y \in \mathcal{X}$:

$$x\mathbf{R}^w y \text{ iff } [x]\mathcal{R}^w[y]$$

In order to show that majority voting on the alternatives yields a total pre-order on $\mathcal{X}$, it suffices to show the following:

> Claim 2: Let $x, y \in \mathcal{X}$. $x\mathbf{R}^w y$ iff $|\{i \in \mathcal{N} \mid x\mathbf{R}_i^w y\}| \geq |\{i \in \mathcal{N} \mid \neg(x\mathbf{R}_i^w y)\}|$.
> Proof: By definition of $\mathbf{R}^w$, it is the case that $x\mathbf{R}^w y$ iff $[x]\mathcal{R}^w[y]$. The order $\mathcal{R}^w$ is obtained by majority voting over the equivalence classes. In other words, $[x]\mathcal{R}^w[y]$ iff $|\{i \in \mathcal{N} \mid [x]\mathcal{R}_i^w[y]\}| \geq |\{i \in \mathcal{N} \mid \neg([x]\mathcal{R}_i^w[y])\}|$. Thus, we have $x\mathbf{R}^w y$ iff $|\{i \in \mathcal{N} \mid [x]\mathcal{R}_i^w[y]\}| \geq |\{i \in \mathcal{N} \mid \neg([x]\mathcal{R}_i^w[y])\}|$. In order to finish the proof, we need to show that $\{i \in \mathcal{N} \mid [x]\mathcal{R}_i^w[y]\} = \{i \in \mathcal{N} \mid x\mathbf{R}_i^w y\}$ and $\{i \in \mathcal{N} \mid \neg([x]\mathcal{R}_i^w[y])\} = \{i \in \mathcal{N} \mid \neg(x\mathbf{R}_i^w y)\}$. This follows from the fact that for all $i \in \mathcal{N}$, it holds that $[x]\mathcal{R}_i^w[y]$ iff $x\mathbf{R}_i^w y$.

□

The first condition of Lemma 6.10 expresses that in the actual world all agents have the same motivational state. Thus, if one agent considers a certain property or reason relevant for the decision at hand, all agents do. The second condition is harder to interpret and only has a clear philosophical meaning in specific instances. There are, however, some things that we can already mention at this point. For starters, notice that the preference definitions are only agent-dependent in $\succeq_i^w$. That is, differences in preferences over the alternatives are solely due to differences in the meta-level preferences of the agents. In addition, the preferences of all the agents are determined by comparing the motivationally salient property packages corresponding to the alternatives in a specific possible world w^*. In a moment, we consider specific situations in which this happens, but the general idea of this is that the epistemic doxastic states of the agents are similar enough.

Lemma 6.10 shows that if both these conditions hold, Arrow's impossibility result is circumvented. Crucial for the proof is that these conditions lead to single-peakedness of the preferences over the equivalence classes. In other words, there is single-peakedness over the options provided that we collapse the options between which the agents are indifferent. Such a collapse is allowed, because all the agents are indifferent between the same alternatives. Therefore, we will from now on speak about single-peaked preferences over options rather than over equivalence classes.

The next theorem shows that if for any alternative and any property it is distributed knowledge whether or not that property holds of that alternative, deliberation ensures that majority voting yields a non-cyclic collective preferences:

Theorem 6.11. Let $\mathcal{N} = \{1, ..., n\}$ with $n \in \mathbb{N}$ being odd be a set of agents, $\mathcal{X}$ a finite set of alternatives and $\mathbb{P}$ a finite set of properties. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a sufficient information common conceptual space model and let $w \in W$ be the actual world. Suppose that for all $i \in \mathcal{N}$, $\mathbf{R}_i^w$ is defined as knowledge or belief-based weak dominance, maximin, leximin, maximax or maximax. Let deliberation be modelled by [!]. Then after deliberation, pairwise majority voting in w generates a total pre-order on the alternatives in $\mathcal{X}$.

Proof. As we want to prove something about the situation after deliberation, we consider the model $M^!$. In order to prove that pairwise majority voting generates a total pre-order on the alternatives in $\mathcal{X}$, we will show that condition 1 and 2 of Lemma 6.10 hold. For condition 1, notice that Definition 5.2 of [!] gives us that $S_i^!(w) = \bigcup_{j \in \mathcal{N}} S_j(w)$ for all $i \in \mathcal{N}$. Let $\mathcal{S} := \bigcup_{j \in \mathcal{N}} S_j(w)$. In order to show that condition 2 holds, we first show that all preference definitions amount to $x\mathbf{R}_i^w y$ iff $r(w, x) \cap \mathcal{S} \succeq_i^w r(w, y) \cap \mathcal{S}$. This is done in a series of claims:

> Claim 1: For all $i \in \mathcal{N}$, $x \in \mathcal{X}$ and $w' \in W$, the following holds: $w \sim_i^! w'$ implies $r_i^!(w', x) = r(w, x) \cap \mathcal{S}$.

PROOF: Let $i \in \mathcal{N}$ and $x \in X$ be arbitrary. Let $w' \in W$ be such that $w \sim_i^! w'$. By definition $r_i^!(w', x) = r^!(w', x) \cap S_i^!(w')$. Using the definition of [!], this amounts to $r_i^!(w', x) = r(w', x) \cap \bigcup_{j \in \mathcal{N}} S_j(w')$. Since $\sim_i^! = \sim_{\mathcal{N}}$, it follows that $S_j(w') = S_j(w)$ for all $j \in \mathcal{N}$. Thus, $\bigcup_{j \in \mathcal{N}} S_j(w') = \mathcal{S}$. Consequently, $r_i^!(w', x) = r(w', x) \cap \mathcal{S}$. According to Definition 4.36 of sufficient information models, it follows from $w \sim_{\mathcal{N}} w'$ that $r(w', x) \cap \mathcal{S} = r(w, x) \cap \mathcal{S}$. Thus, $r_i^!(w', x) = r(w, x) \cap \mathcal{S}$.

CLAIM 2: For all $i \in \mathcal{N}$ and $x \in \mathcal{X}$: $X_i^B(w) = X_i^K(w) = \{r(w, x) \cap \mathcal{S}\}$.
PROOF: Let $i \in \mathcal{N}$ and $x \in X$ be arbitrary. Firstly, recall that $X_i^K(w) = \{J \subseteq \mathbb{P} \mid \exists w' \in W(w \sim_i^! w' \wedge J = r_i^!(w', x))\}$. By claim 1, it follows that $X_i^K(w) = \{r(w, x) \cap \mathcal{S}\}$. Secondly, suppose that $J \in X_i^B(w)$. By Definition 4.10, it follows that there exists a $w' \in W$ such that $w \sim_i^! w'$ and for all $w'' \in W$ such that $w \leq_i^! w''$ there exists a $w''' \in W$ such that $w'' \leq_i^! w'''$ and $r_i^!(w''', x) = J$. Due to the reflexivity of $\leq_i^!$, it must be the case that there exists a $v \in W$ such that $w' \leq_i^! v$ and $r_i^!(v, x) = J$. Since $w \sim_i^! w' \leq_i^! v$ and $\leq_i^! \subseteq \sim_i^!$, it follows that $w \sim_i^! v$. This means $J \in X_i^K(w)$. Thus, $X_i^B(w) \subseteq X_i^K(w)$. Thirdly, according to Proposition 4.12, $X_i^B(w) \neq \emptyset$. Combining all of this gives us that $\emptyset \neq X_i^B(w) \subseteq X_i^K(w) = \{r(w, x) \cap \mathcal{S}\}$. Hence $X_i^B(w) = X_i^K(w) = \{r(w, x) \cap \mathcal{S}\}$.

CLAIM 3: For all $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$ the following holds: If $x\mathbf{R}_i^w y$ is defined as knowledge or belief-based weak dominance, maximin, leximin, maximax or leximax, then $x\mathbf{R}_i^w y$ iff $r(w, x) \cap \mathcal{S} \succeq_i^w r(w, y) \cap \mathcal{S}$.
PROOF: Let $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$ be arbitrary. For all possible preference definitions, we will show that $x\mathbf{R}_i^w y$ iff $r(w, x) \cap \mathcal{S} \succeq_i^w r(w, y) \cap \mathcal{S}$:

- Suppose that $x\mathbf{R}_i^w y$ is defined as knowledge-based weak dominance. That is, $x\mathbf{R}_i y$ iff for all $w' \in W$ such that $w \sim_i^! w'$, it holds that $r_i^!(w', x) \succeq_i^{w'} r_i^!(w', y)$. The left-to-right direction follows immediately from claim 1 and the fact that $w \sim_i^! w'$. For the right-to-left direction, suppose that $r(w, x) \cap \mathcal{S} \succeq_i^w r(w, y) \cap \mathcal{S}$. Let $w' \in W$ be such that $w \sim_i^! w'$. Since $\sim_i^! = \bigcap_{j \in \mathcal{N}} \sim_j$, it holds that $w \sim_i w'$. Given the fact that M is a model for deliberation, it holds that $\succeq_i^{w'} = \succeq_i^w$. Claim 1 gives us that $r_i^!(w', x) = r(w, x) \cap \mathcal{S}$ and that $r_i^!(w', y) = r(w, y) \cap \mathcal{S}$. Thus, from $r(w, x) \cap \mathcal{S} \succeq_i^w r(w, y) \cap \mathcal{S}$, we can conclude that $r_i^!(w', x) \succeq_i^{w'} r_i^!(w', y)$.

- Suppose that $x\mathbf{R}_i^w y$ is defined as belief-based weak dominance. That is, $x\mathbf{R}_i y$ iff for all $w' \in W$ such that $w \sim_i^! w'$ there exists a $w'' \in W$ such that $w' \leq_i^! w''$ and for all $w''' \in W$ such that $w'' \leq_i^! w'''$, it holds that $r_i^!(w''', x) \succeq_i^{w'''} r_i(w''', y)$.
  For the left-to-right direction, notice that $w \sim_i^! w$. Thus, there exists a $v \in W$ such that $w \leq_i^! v$ and for all $v' \in W$ such that $v \leq_i^! v'$, it holds that $r_i^!(v', x) \succeq_i^{v'} r_i^!(v', y)$. Due to the reflexivity of $\leq_i^!$, it follows that $r_i^!(v, x) \succeq_i^v r_i^!(v, y)$. Since $\leq_i^! \subseteq \sim_i^!$, it follows that $w \sim_i^! v$. By claim 1, it follows that $r_i^!(v, x) = r(w, x) \cap \mathcal{S}$ and $r_i^!(v, y) = r(w, y) \cap \mathcal{S}$. Moreover, since $\sim_i^! \subseteq \sim_i$, it follows that $\succeq_i^v = \succeq_i^w$. Therefore, we can conclude that $x\mathbf{R}_i y$ implies $r(w, x) \cap \mathcal{S} \succeq_i^w r(w, y) \cap \mathcal{S}$.
  For the right-to-left direction, suppose that $r(w, x) \cap \mathcal{S} \succeq_i^w r(w, y) \cap \mathcal{S}$. Let $v \in W$ be such that $w \sim_i^! v$. Obviously, $v \leq_i^! v$. Let $v' \in W$ be such that $v \leq_i^! v'$. We need to show that $r_i^!(v', x) \succeq_i^{v'} r_i^!(v', y)$. Notice that from $w \sim_i^! v \leq_i^! v'$ and the fact that $\leq_i^! \subseteq \sim_i^!$, it follows that $w \sim_i^! v'$. Thus, according to claim 1, $r_i^!(v', x) = r(w, x) \cap \mathcal{S}$ and that $r_i^!(v', y) = r(w, y) \cap \mathcal{S}$. Moreover, $w \sim_i^! v$ implies $w \sim_i v$. As M is a model for deliberation, this implies that $\succeq_i^{w'} = \succeq_i^w$. Thus, from $r(w, x) \cap \mathcal{S} \succeq_i^w r(w, y) \cap \mathcal{S}$, we can conclude that $r_i^!(v', x) \succeq_i^{v'} r_i^!(v', y)$.

- Suppose that $x\mathbf{R}_i^w y$ is defined as knowledge or belief-based maximin, leximin, maximax or leximax. From claim 2, it follows that $X_i^K(w) = X_i^B(w)$ contains precisely one element, namely $r(w, x) \cap \mathcal{S}$. The same holds for y. Therefore, $x\mathbf{R}_i^w y$ iff $r(w, x) \cap \mathcal{S} \succeq_i^w r(w, y) \cap \mathcal{S}$.

Finally, in order to show that condition 2 of Lemma 6.10 holds, notice that $w \sim_{\mathcal{N}} w$. Let $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$ be arbitrary. Let $x\mathbf{R}_i^w y$ be defined as knowledge or belief-based weak dominance, maximin, leximin, maximax or leximax. According to claim 3, it is the case that $x\mathbf{R}_i^w y$ iff $r(w,x) \cap \mathcal{S} \succeq_i^w r(w,y) \cap \mathcal{S}$. Thus, condition 2 is satisfied.

In conclusion, since conditions 1 and 2 of Lemma 6.10 are satisfied, it is the case that, after deliberation, pairwise majority voting in w generates a total pre-order on the alternatives in $\mathcal{X}$. □

Condition 1 of Lemma 6.10 is satisfied, because deliberation always induces a common motivational state. Condition 2 is met because of several reasons. Firstly, the common motivational state in the actual world w ensures that in w, and in all worlds that none of the agents can distinguish from w, the motivationally salient property packages corresponding to the alternatives are agent-independent. The fact that the preference definitions of all agents boil down to a comparison of the motivationally salient property packages in the actual world follows from the sufficient information condition. This condition ensures that for all $i \in \mathcal{N}$, all $x \in \mathcal{X}$ and all $w' \in W$, $w \sim_i^! w'$ implies $r_i^!(w', x) = r(w, x) \cap \mathcal{S}$. Thus, crucial for the proof of condition 2 is that in cases of sufficient information, deliberation ensures that all agents have hard information about the motivationally salient property packages corresponding to the alternatives.

The next theorem shows that deliberation ensures that majority voting yields non-cyclic collective preferences if all the agents are opinionated:

Theorem 6.12. Let $\mathcal{N} = \{1, ..., n\}$ with $n \in \mathbb{N}$ being odd be a set of agents, $\mathcal{X}$ a finite set of alternatives and $\mathbb{P}$ a finite set of properties. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be an opinionated common conceptual space model and let $w \in W$ be the actual world. For all $i \in \mathcal{N}$, let $\mathbf{R}_i^w$ be defined as belief-based weak dominance, maximin, leximin, maximax or leximax. Let deliberation be modelled by [!]. Then after deliberation, pairwise majority voting in w generates a total pre-order on the alternatives in $\mathcal{X}$.

Proof. As we want to prove something about the situation after deliberation, we consider the model $M^!$. We start by showing that $M^!$ is opinionated:

> Claim 1: If M is opinionated, then $M^!$ is as well.
> Proof: Firstly, according to Definition 5.2, $\leq_i^! := \leq_i \cap \sim_i^!$ for all $i \in \mathcal{N}$. Thus, if $\leq_i$ is converse well-founded, then so is $\leq_i^!$. Hence $M^!$ is a standard. Secondly, let $i \in \mathcal{N}$, $x \in \mathcal{X}$ and $P \in \mathbb{P}$ be arbitrary. Let $v, v' \in W$ be such that $v \sim_i^! v'$, $P \in S_i^!(v)$, $P \in r(v, x)$ and $P \notin r(v', x)$. In order to show that $M^!$ is opinionated, we need to show that $v <_i^! v'$ or $v' <_i^! v$. Suppose, for contradiction, that $v \leq_i^! v'$ and $v' \leq_i^! v$. Notice that since $S_i^!(v) = \bigcup_{j \in \mathcal{N}} S_j(v)$, there exists an agent $j \in \mathcal{N}$ such that $P \in S_j(v)$. As M is opinionated, it follows that $v <_j v'$ or $v' <_j v$. Moreover, by Proposition 3.7, it follows that $\leq_i^! = \bigcap_{i \in \mathcal{N}} \leq_i$. Thus, $v \leq_j v'$ and $v' \leq_j v$, which is a contradiction. Thus, $v <_i^! v'$ or $v' <_i^! v$.

In order to prove that pairwise majority voting generates a total pre-order on the alternatives in $\mathcal{X}$, we will show that condition 1 and 2 of Lemma 6.10 hold. For condition 1, notice that Definition 5.2 of [!] gives us that $S_i^!(w) = \bigcup_{j \in \mathcal{N}} S_j(w)$ for all $i \in \mathcal{N}$. Let $\mathcal{S} := \bigcup_{i \in \mathcal{N}} S_i(w)$. For condition 2, we consider the preference definitions in the updated model. In order to see what the definitions of belief-based preference amount to, let $w^* \in W$ be such that $w \to_{\mathcal{N}} w^*$. By Definition 3.10, $w \to_{\mathcal{N}} w^*$ iff $w \sim_{\mathcal{N}} w^*$ and $w^* \in \mathrm{Max}(\leq_{\mathcal{N}})$. M is a model for deliberation, thus $F = (W, \sim_i, \leq_i)_{i \in \mathcal{N}}$ is a standard common prior frame. According to Proposition 3.7, it then holds that $\leq_{\mathcal{N}} = \bigcap_{i \in \mathcal{N}} \leq_i$. Since the plausibility relations of all the agents are converse well-founded, $\leq_{\mathcal{N}}$ is as well and such a w^* exists. We now prove the following claim:

CLAIM 2: For all $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$, $x\mathbf{R}_i^w y$ iff $r(w^*, x) \cap \mathcal{S} \succeq_i^{w^*} r(w^*, x) \cap \mathcal{S}$.
PROOF: Let $i \in \mathcal{N}$ and $x, y \in \mathcal{X}$. Claim 1 gives us that $M^!$ is an opinionated model. According to the proof of Theorem 4.40, no matter which belief-based preference type agent i has, the following holds: $x\mathbf{R}_i^w y$ iff $r_i^!(w^{**}, x) \succeq_i^w r_i^!(w^{**}, y)$ for some $w^{**} \in W$ with $w \rightarrow_i^! w^{**}$. In order to prove the claim, we show the following equalities:

- $\succeq_i^w = \succeq_i^{w^*}$. From $w \sim_{\mathcal{N}} w^*$ and $\sim_{\mathcal{N}} \subseteq \sim_i$, it follows that $w \sim_i w^*$. Since M is a model for preference formation, it is the case that $\succeq_i^w = \succeq_i^{w^*}$.
- $r_i^!(w^{**}, x) = r(w^*, x) \cap \mathcal{S}$. As F is a standard common prior frame, it follows from Proposition 3.13 that $\rightarrow_i^! = \rightarrow_{\mathcal{N}}$ and, hence, that $w \rightarrow_i^! w^*$. $M^!$ is opinionated and both $w \rightarrow_i^! w^*$ and $w \rightarrow_i^! w^{**}$ hold. Thus, according to Lemma 4.39, it is the case that $r_i^!(w^{**}, x) = r_i^!(w^*, x) = r(w^*, x) \cap \bigcup_{j \in \mathcal{N}} S_j(w^*)$. Since $w \sim_{\mathcal{N}} w^*$ and $\sim_{\mathcal{N}} = \bigcap_{j \in \mathcal{N}} \sim_j$, it follows that $S_j(w) = S_j(w^*)$ for all $j \in \mathcal{N}$. Thus, $\bigcup_{j \in \mathcal{N}} S_j(w^*) = \bigcup_{j \in \mathcal{N}} S_j(w) = \mathcal{S}$. Therefore, $r_i^!(w^{**}, x) = r_i^!(w^*, x) = r(w^*, x) \cap \mathcal{S}$.
- $r_i^!(w^{**}, y) = r(w^*, y) \cap \mathcal{S}$. The proof is similar to the previous item.

From this claim and the fact that $\sim_{\mathcal{N}} = \sim_{\mathcal{N}}^!$, it follows that there exists a $w^* \in W$ such that $w \sim_{\mathcal{N}}^! w^*$ and for all $i \in \mathcal{N}$ and all $x, y \in \mathcal{X}$, it holds that $x\mathbf{R}_i^w y$ iff $r(w^*, x) \cap \mathcal{S} \succeq_i^{w^*} r(w^*, x) \cap \mathcal{S}$.

As conditions 1 and 2 of Lemma 6.10 are satisfied, it follows that pairwise majority voting in w generates a total pre-order on the alternatives in $\mathcal{X}$. □

As before, condition 1 of Lemma 6.10 is satisfied, because deliberation induces a common motivational state. In turn, the common motivational state ensures that the motivationally salient property packages corresponding to the alternatives are agent-independent. This together with the fact that after deliberation both the epistemic indistinguishability and the plausibility relations are the same for all agents ensures that condition 2 is met. Thus, crucial for the proof of condition 2 is that in cases where the agents are opinionated, deliberation induces a common epistemic doxastic state.

6.3.2 Discussion

The previous section showed that if there is a common conceptual space and if, additionally, there is sufficient information about the alternatives or the agents are opinionated, then deliberation ensures that pairwise majority voting yields acyclic collective preferences. It does so by inducing single-peakedness on the preferences over options, provided we collapse the options between which the agents are indifferent. Since all agents are indifferent between the same options, such a collapse is allowed and we can simply speak of single-peakedness over the options. This section discusses the formal results from the previous section. It starts by showing that deliberation leads to meta-agreement and then focuses on the question of how realistic it is that deliberation induces single-peaked preference profiles.

Actual Meta-Agreement

In Section 6.2.1, it was argued that common conceptual space models formalise the idea of virtual normative meta-agreement. The previous section showed that if there is a common conceptual space, deliberation can, in specific circumstances, ensure single-peaked preference profiles. We will now argue that this happens, because deliberation ensures both normative and factual meta-agreement.

First, let us consider normative meta-agreement. Recall that a common conceptual space encloses virtual normative meta-agreement. That is, it requires the existence of some subconscious common dimension with respect to which the agents conceptualise the issue at stake. In reality, however, the agents might conceptualise the decision at hand differently. These different conceptualisations are solely due to differences in the motivational state of the agents. Thus, if the motivational state of

the agents is the same, there is no difference between virtual and actual normative meta-agreement. Deliberation enforces a common motivational state and, by doing so, it turns virtual into actual normative meta-agreement.

Second, let us consider factual meta-agreement. Or better yet, factual agreement. In cases of sufficient information or opinionatedness, deliberation leads to factual agreement. For starters, notice that in sufficient information models, deliberation ensures that each agent has the right information about which motivationally salient properties hold of which alternative:

Proposition 6.13. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a sufficient information common conceptual space model and let $w \in W$ be the actual world. Then for all $P \in \mathbb{P}$ and all $x \in X$, the following hold:

- $M, w \models Px \wedge \bigvee_{i \in \mathcal{N}} S_i P$ implies $M^!, w \models \bigwedge_{i \in \mathcal{N}} K_i Px$.
- $M, w \models \neg Px \wedge \bigvee_{i \in \mathcal{N}} S_i P$ implies $M^!, w \models \bigwedge_{i \in \mathcal{N}} K_i \neg Px$.

Proof. Left to the reader. □

Moreover, in opinionated models deliberation ensures that agents have the same beliefs about which motivationally salient reasons or properties hold of which alternatives:

Proposition 6.14. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be an opinionated common conceptual space model and let $w \in W$ be the actual world. Then for all $P \in \mathbb{P}$ and all $x \in \mathcal{X}$ the following holds:

- $M, w \models \bigvee_{i \in \mathcal{N}} S_i P$ implies $M^!, w \models \bigwedge_{i \in \mathcal{N}} B_i Px \vee \bigwedge_{i \in \mathcal{N}} B_i \neg Px$.

Proof. Left to the reader. □

The proof of this proposition makes essential use of the common prior assumption. Because the informational differences between the agents disappear as a result of deliberation, their beliefs will be the same. Thus, these propositions show that in cases of sufficient information and opinionatedness, deliberation leads to agreement with respect to the agents' knowledge and belief about which properties hold of which alternatives.

In conclusion, whenever there is a common conceptual space, deliberation ensures that there is normative and factual meta-agreement. In particular, deliberation accomplishes this by turning virtual into actual normative meta-agreement and by enforcing factual agreement. By doing so, the formal conditions of Lemma 6.10, which lead to single-peakedness, are enforced. Thus, our results show that deliberation leads to single-peaked preference profiles via meta-agreement.

Can Deliberation Ensure Single-Peakedness?

In our formal framework, deliberation ensures that pairwise majority voting yields acyclic social preferences, because it induces single-peaked preference profiles. In order to determine whether it is likely that deliberation produces single-peakedness through meta-agreement in reality, the assumptions made in our formal framework need to be discussed. In particular, we discuss the sufficient information assumption, opinionatedness, the assumption on the preference types of the agents and the concept of a common conceptual space.

The sufficient information assumption says that for any $P \in \mathbb{P}$ that is motivationally salient for at least one of the agents and for any $x \in \mathcal{X}$, it is distributed knowledge whether or not P holds of x. Recall that the elements of $\mathbb{P}$ can both be interpreted as properties that hold of the alternatives or as reasons that can be used in justifying one's preference for a certain alternative. In order to discuss whether the sufficient information assumption is likely, we need to consider both options separately.

For starters, the question whether a certain property holds of a certain alternative is clearly a factual question. Assuming distributed knowledge about the relevant ontic facts of the world, i.e. about which motivationally salient properties hold of which alternatives, might seem like a rather strong assumption at first sight, but it is actually less demanding than it seems. Political debates are often institutionalised. They take place in town councils, the national parliament, the European parliament and so forth. The participants of these debates are usually specialists on a specific topic and they enter the debate well-informed. In circumstances like these, distributed knowledge about which property holds of which alternative is quite likely.

When we think of the elements of $\mathbb{P}$ as reasons that agents can appeal to in justifying their preferences, it is harder to argue in favour of the sufficient information assumption. The main reason for this is that people usually use political watchwords as reasons and, as Ottonelli and Porello (2013: 85) argue, "political debates are fought not only over policies ... but also and very importantly over the meaning of fundamental words". Consider, for instance, the example about the availability of religious education in which the government has to decide between a complete ban on religious education (a), privately funded religious education (b) or publicly funded religious eduction (c). Agent 1 might argue for option a on the grounds of fairness, whereas agent 2 might argue for option c also on the grounds of fairness. The problem is that both agents have a different concept of fairness in mind. Agent 1 may consider option a to be fair because this option ensures that the government does not favour any religion over another, whereas agent 2 may consider option c to be fair because it enables children to get the education that their parents see fit for them. The problem here is that agents 1 and 2 disagree about the meaning of fairness in the debate about the availability of religious education. This problem is less likely to arise in a deliberative democracy, however, because people cannot just use political watchwords to justify their preferences. They also have to justify why this watchword applies. In the example above, the agents should not justify their preference by appealing to fairness but to the underlying reasons. As it is the underlying reasons and not the political watchwords that count, a situation in which it is distributed knowledge which reasons can be used for justifying one's preference for which alternatives is achievable. Nevertheless, when we interpret the elements of $\mathbb{P}$ as reasons the sufficient information assumption is much more demanding than when we interpret them as properties.

The assumption on opinionated models says that for any $P \in \mathbb{P}$ that is motivationally salient for at least one of the agents and any $x \in \mathcal{X}$, each agent either believes that P holds of x or he believes that P does not hold of x.[17] That is, all the agents have an opinion on whether a certain property holds of a certain alternative or whether a certain reason can be appealed to in justifying one's preference for a certain alternative. This assumption is relatively undemanding. In reality, it might well be the case that for a given proposition φ someone acknowledges that he does not know whether φ holds or not, but it is unlikely that he does not have beliefs about this. Not having any beliefs at all with regard to a certain proposition seems strange. Therefore, the assumption of opinionated agents is unproblematic.

The theorems from the previous section also make assumptions about the preference types of the agents. In sufficient information models, the preferences of the agents should be defined as knowledge or belief-based weak dominance, maximin, leximin, maximax or leximax. In opinionated models, the preferences should be belief-based. That is, although agents might form their preferences differently, there is meta-agreement about the rational ways to form preferences under uncertainty. Therefore, this condition is related to the concept of *rationality meta-agreement*.

List (2002: 74) argues that there is rationality meta-agreement if the agents determine their preferences by determining a peak on the chosen issue dimension "with decreasing preference as options are increasingly distant from that most preferred option", i.e. if the agents preferences are single-peaked with respect to the identified relevant issue dimension. However, elsewhere Dietrich and List (2013a, 2013b) argue that preferences over options are determined by their underlying properties and fundamental values. In other words, preferences over options are determined by meta-level preferences

[17] Recall that in addition to this assumption, opinionated models satisfy standardness. Since the assumption of standardness has solely been made for reasons of mathematical simplicity, we will not discuss it any further.

over more stable factors. It is this approach that we have worked with in this thesis, because it is more in line with the philosophical literature on deliberative democracy which stresses that preferences are secondary and should be justifiable.

In this approach, preference formation is rational if the preferences over the options are determined in a rational way on the basis of the agent's knowledge/belief, his motivational state and the properties that hold of the alternatives. Therefore, we can interpret this assumption as rationality meta-agreement. The assumption that agents base their preferences on the above-mentioned things is very likely. The assumption that this happens rationally is ensured by deliberation. In a democratic deliberation, people have to justify their preferences. If someone forms his preferences in an irrational matter, they will not be justifiable and this will be pointed out in a deliberation. The agent in question is then forced to form his preference in a justifiable and hence reasonable manner. Therefore, we may conclude that the assumption about the preference types of the agents is unproblematic.

Lastly, the assumption of a common conceptual space requires some discussion. The condition on common conceptual space models formalises virtual normative meta-agreement. The spirit of virtual normative meta-agreement is that each agent has a favourite justification/property package and that the further a property package is away from the most preferred one, the less appealing it is. This idea is rather intuitive when interpreted n-dimensionally. In making complex decision, people often know what properties or characteristics the ideal option should have. Other options are then weighed on how far they are removed from the optimal option. The problem of our framework, however, is that a common conceptual space requires the existence of one dimension in terms of which all the agents subconsciously conceptualise the issue at stake. This is rather unrealistic, because in many decisions there are several independent aspects that play a role. In considering which energy policy to fund, for instance, one might take into account sustainability, efficiency, the energy price, implementation time and so on. It is unlikely that there exists a linear order on all possible property packages with respect to which all the agents order their meta-level preferences. Most issues are too complex for a one dimensional representation. That seems only reasonable in cases where there really is only one natural issue dimension accompanying the problem. This is the case, for instance, in elections in countries with a powerful left-right ordering, in decisions about the availability of religious education, in determining the legal drinking age, et cetera. In those scenarios, the virtual normative meta-agreement assumption is likely. It is, however, not at all clear why this should hold in general.

List's (2002: 74) original idea of normative meta-agreement is less demanding than ours, because it only requires that "group deliberation leads people to identify a single shared issue-dimension in terms of which the issue at stake is to be conceptualized". There might be several important issue dimensions, but the only thing that matters is that deliberation somehow identifies a single issue dimension as the most relevant. There is, however, also a big problem with this requirement.

To begin with, it is not at all clear why deliberation would lead people to focus on one of the possibly many issue dimensions that play a role in the decision at hand. On the contrary, deliberation confronts people with new perspectives and new ideas. Consequently, deliberation might show just how complex an issue really is.[18] Hoping that deliberation will lead to the identification of a single relevant issue dimension is unlikely and, more importantly, not in line with the spirit of deliberative democracy, which asks the deliberators to take each other's perspective seriously. Even worse, it is hostile to pluralism precisely because everyone conceptualises the possible options in exactly the same way. As Ottonelli and Porrello (2013: 85) argue, "recognizing pluralism in the realm of politics means respecting ... diversity and the importance of disagreement". Thus, List's (2002) hypothesis that deliberation leads to normative meta-agreement by singling out issue dimension is hostile to pluralism and not in line with deliberative democracy.

Furthermore, even if we assume that deliberation leads to agreement on what the most relevant issue dimension is, our problems do not disappear. If deliberation has such great persuasive power, it is not clear why deliberation would be able to lead to meta-agreement but not to substantive agreement. Especially because, as Ottonelli and Porello (2013) argue, substantive agreement need not be more demanding than meta-agreement because it allows people to prefer the same option for different

[18]For this argument, see Miller (1992): 64-66 and Knight & Johnsson (1994): 286.

reasons.

Concludingly, we can say that the hypothesis that group deliberation miraculously creates normative meta-agreement out of the blue is problematic. Hence there is no reason to prefer this hypothesis to our problematic assumption of virtual normative meta-agreement. We believe we should accept that the philosophical applicability of single-peakedness is restricted to scenarios in which there already is a single issue dimension accompanying the decision problem. In other cases, one needs other arguments and another formal framework in order to show that deliberative democracy is able to overcome the impossibility results of social choice theory.

Conclusion

In conclusion, the formal results show that if there is a common conceptual space and (i) sufficient information about which properties hold of which alternatives or (ii) opinionatedness, then deliberation ensures single-peaked preference profiles. Deliberation does this by turning virtual normative meta-agreement into actual normative meta-agreement and by enforcing factual meta-agreement. As the sufficient information assumption is demanding but realisable and opinionatedness is plausible, our theorems show – as the literature on deliberative democracy suggests – that deliberation might help circumvent Arrow's impossibility result in cases where there is a natural issue dimension accompanying the decision problem. It remains unclear, however, why and how a deliberative process in line with the principles of deliberative democracy would out of the blue lead to agreement on which issue dimension is the most relevant.

This conclusion is best summarised in the words of Mueller (1989: 89-90), who holds that "given that we have a single-dimensional issue, single-peakedness does not seem to be that strong an assumption, what is implausible is the assumption that the issue space is one dimensional." Our formal results affirm the first claim, because they show that the conditions needed for proving that deliberation leads to single-peakedness in the case of a one dimensional issue space, are rather mild.

6.3.3 Expressing the Formal Results in LDD

Section 6.3.1 showed that if there is a common conceptual space and if, additionally, there is sufficient information about the alternatives or the agents are opinionated, then deliberation ensures that pairwise majority voting yields acyclic collective preferences. This section states variants of these result as validities in the logic of democratic deliberation (LDD).

For starters, we fix the following parameters for the rest of this section. Let $\mathcal{N}$ be a finite set of agents, $\mathcal{X}$ a finite set of alternatives and $\mathbb{P}$ a finite set of properties. Let L denote the set of all linear order on $\mathcal{P}(\mathbb{P})$. Moreover, let $\alpha : \mathcal{N} \to \{K, B\}$ be an attitude function, where $\alpha(i)$ denotes whether i's preferences are knowledge or belief-based. Additionally, let $\pi : \mathcal{N} \to \{wd, mmin, lmin, mmax, lmax\}$ be a protocol function, where $\pi(i)$ denotes the protocol used by i to form his preferences.

It is possible to state weakenings of the conditions needed for proving our formal results about deliberation in $\mathcal{L}_{LDD}$. First, consider common conceptual space models. Recall that a model $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i\in\mathcal{N}}$ for deliberation is a common conceptual space model iff for all $w \in W$, there exists a strict linear order $\gg^w$ on $\mathcal{P}(\mathbb{P})$ such that for all $i \in \mathcal{N}$, $\succeq_i^w$ is single-peaked with respect to $\gg^w$. We cannot state the full condition in $\mathcal{L}_{LDD}$, since this language does not have a universal modality. However, in the proof of Theorems 6.11 and 6.12, we only used that there is a common conceptual space in the actual world. This is expressible in $\mathcal{L}_{LDD}$:

Definition 6.15. Let *common* be the following abbrevation denoting that, locally speaking, there is a common conceptual space:

$$common := \bigvee_{\gg \in L} \bigwedge_{i \in \mathcal{N}} \Big(\bigwedge_{\substack{J,J',J'' \subseteq \mathbb{P} \\ J \gg J' \gg J''}} \neg(J'' \succeq_i J \succeq_i J') \wedge \bigwedge_{\substack{J,J',J'' \subseteq \mathbb{P} \\ J'' \gg J' \gg J}} \neg(J'' \succeq_i J \succeq_i J') \Big)$$[19]

Proposition 6.16. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a common conceptual space model. Then for all $w \in W$, $M, w \models common$.

Proof. Left to the reader. □

Second, consider the sufficient information condition. Recall that a model $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ for deliberation is a sufficient information model iff for all $x \in X$, all $P \in \mathbb{P}$ and all $w, w' \in W$, the following holds: if $w \sim_{\mathcal{N}} w'$ and $P \in \bigcup_{i \in \mathcal{N}} S_i P(w)$, then $P \in r(w, x)$ iff $P \in r(w', y)$. Again, this cannot be expressed in $\mathcal{L}_{LDD}$ due to the lack of a universal modality. However, we again only need the fact that there is sufficient information about which motivationally salient properties hold of which alternatives in the actual world for proving Theorems 6.11 and 6.12. This local variant can be expressed as follows:

Definition 6.17. Let $sufinfo$ be the following abbrevation denoting that, locally speaking, there is sufficient information:
$$sufinfo := \bigwedge_{\substack{P \in \mathbb{P} \\ x \in \mathcal{X}}} \Big(\bigvee_{i \in \mathcal{N}} S_i P \rightarrow (D_K Px \vee D_K \neg Px) \Big)$$

Proposition 6.18. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be a sufficient information model. Then for all $w \in W$, $M, w \models sufinfo$.

Proof. Left to the reader. □

Third, consider opinionatedness. Recall that a model $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ for deliberation is opinionated iff it is standard and for all $i \in \mathcal{N}$, all $x \in X$, all $P \in \mathbb{P}$ and all $w, w' \in W$ the following holds: if $w \sim_i w'$, $P \in S_i(w)$, $P \in r(w, x)$ and $P \notin r(w', x)$, then $w <_i w'$ or $w' <_i w$. Neither this nor the local variant of this condition can be expressed in $\mathcal{L}_{LDD}$. However, we can introduce an abbrevation which is valid over all opinionated models and which says that in the actual world the group is opinionated with respect to properties that are motivationally salient for at least one of its members:

Definition 6.19. Let *opinionatedgroup* be the following abbreviation denoting that, locally speaking, the group is opinionated:
$$opinionatedgroup := \bigwedge_{\substack{P \in \mathbb{P} \\ x \in \mathcal{X}}} \Big(\bigvee_{i \in \mathcal{N}} S_i P \rightarrow (D_B Px \vee D_B \neg Px) \Big).$$

Proposition 6.20. Let $M = (W, \sim_i, \leq_i, S_i, \succeq_i, r)_{i \in \mathcal{N}}$ be an opinionated model for deliberation. Then for all $w \in W$, $M, w \models opinionatedgroup$.

Proof. Suppose, for contradiction, that there exists a $w \in W$ such that $M, w \not\models opinionatedgroup$. Then there exists a property $P \in \mathbb{P}$, an alternative $x \in \mathcal{X}$ and an agent $i \in \mathcal{N}$ such that $M, w \models S_i P$ and $M, w \models \neg(D_B Px \vee D_B \neg Px)$. That is, $M, w \models \neg D_B Px \wedge \neg D_B \neg Px$. By Proposition 3.11, there exists a $w' \in W$ such that $w \rightarrow_{\mathcal{N}} w'$ and $M, w' \models \neg Px$ and there exists a $w'' \in W$ such that $w \rightarrow_{\mathcal{N}} w''$ and $M, w'' \models Px$. Recall that Definition 3.10 of $\rightarrow_{\mathcal{N}}$: for all $v, v' \in W$, $v \rightarrow_{\mathcal{N}} v'$ iff $v \sim_{\mathcal{N}} v'$ and $v' \in \mathrm{Max}(\leq_{\mathcal{N}})$. Thus, $w' \in \mathrm{Max}(\leq_{\mathcal{N}})$ and $w'' \in \mathrm{Max}(\leq_{\mathcal{N}})$. As both are maximal worlds, we get that $w' \leq_{\mathcal{N}} w''$ and $w'' \leq_{\mathcal{N}} w'$. Since $\leq_{\mathcal{N}} = \bigcap_{i \in \mathcal{N}} \leq_i$, it follows that $w' \leq_i w''$ and $w'' \leq_i w'$. Furthermore, notice that since $\sim_{\mathcal{N}}$ is an equivalence relation and $\sim_{\mathcal{N}} = \bigcap_{i \in \mathcal{N}} \sim_i$, it follows that $w' \sim_i w''$. Thus, there exist two worlds $w', w'' \in W$ with $w' \sim_i w''$ and a property $P \in S_i(w) = S_i(w') = S_i(w'')$ such that $P \notin r(w', x)$ and $P \in r(w'', x)$. Moreover, $w' \leq_i w''$ and $w'' \leq_i w'$. This contradicts opinionatedness. Therefore, $M, w \models opinionatedgroup$. □

[19] Needless to say, for all $i \in \mathcal{N}$ and all $J, J', J'' \subseteq \mathbb{P}$: $J \succeq_i J' \succeq_i J'' := (J \succeq_i J') \wedge (J' \succeq_i J'')$.

Thus far, we have stated weakings of the assumptions of a common conceptual space, sufficient information and opinionatedness in $\mathcal{L}_{LDD}$. These weaker conditions are sufficient for showing that deliberation ensures that pairwise majority voting yields acyclic preferences. Now, we turn to stating the idea of acyclic collective preferences in $\mathcal{L}_{LDD}$. First, we define collective or group preferences:

Definition 6.21. For all $x, y \in X$, let $Pref^{\pi}_{\alpha}(x, y)$ be the following abbreviation denoting that under majority voting the group weakly prefers x to y:
$Pref^{\pi}_{\alpha}(x, y) := \bigvee_{\substack{N,N' \subseteq \mathcal{N} \\ N \cup N' = \mathcal{N} \\ N \cap N' = \emptyset \\ |N| \geq |N'|}} (\bigwedge_{i \in N} Pref^{\pi(i)}_{i,\alpha(i)}(x, y) \wedge \bigwedge_{i \in N'} \neg(Pref^{\pi(i)}_{i,\alpha(i)}(x, y)))$

Notice that if the collective preferences are a total pre-order, they are acyclic. Before expressing the fact that the group preferences form a total pre-order, we first introduce abbreviations expressing that they are reflexive, transitive and total:

Definition 6.22. Let $reflexivity$, $transitivity$ and $totality$ be the following abbreviations denoting that the collective preferences are reflexive, transitive and total respectively:
$reflexivity := \bigwedge_{x \in \mathcal{X}} Pref^{\pi}_{\alpha}(x, x)$
$transitivity := \bigwedge_{x,y,z \in \mathcal{X}} ((Pref^{\pi}_{\alpha}(x, y) \wedge Pref^{\pi}_{\alpha}(y, z)) \rightarrow Pref^{\pi}_{\alpha}(x, z))$
$totality := \bigwedge_{x,y \in \mathcal{X}} (Pref^{\pi}_{\alpha}(x, y) \vee Pref^{\pi}_{\alpha}(y, x))$

Definition 6.23. Let $tpo(Pref^{\pi}_{\alpha})$ be the following abbreviation denoting that the collective preferences are a total pre-order:
$tpo(Pref^{\pi}_{\alpha}) := reflexivity \wedge transitivity \wedge totality$

With these abbreviations in place, it is possible to state variants of Theorems 6.11 and 6.12 as validities in LDD:

Theorem 6.24. $\models (common \wedge sufinfo) \rightarrow [!]tpo(Pref^{\pi}_{\alpha})$

Proof. The proof of this theorem is along the lines of the proof of Theorem 6.11. □

Theorem 6.25. If $\alpha : \mathcal{N} \rightarrow \{K, B\}$ is such that $\alpha(i) = B$ for all $i \in \mathcal{N}$, then the following holds: $\models (common \wedge opinionatedgroup) \rightarrow [!]tpo(Pref^{\pi}_{\alpha})$

Proof. The proof of this theorem is along the lines of the proof of Theorem 6.12. □

As the proof system of LDD is complete, it is the case that the above results are provable in the proof system Λ_{LDD}:

Corollary 6.26. $\vdash (common \wedge sufinfo) \rightarrow [!]tpo(Pref^{\pi}_{\alpha})$.
Corollary 6.27. Suppose $\alpha : \mathcal{N} \rightarrow \{K, B\}$ is such that for all $i \in \mathcal{N}$, $\alpha(i) = B$. Then the following holds: $\vdash (common \wedge opinionatedgroup) \rightarrow [!]tpo(Pref^{\pi}_{\alpha})$

In conclusion, Λ_{LDD} proves that if in the actual world there is a common conceptual space and either sufficient information or opinionatedness of the group, then deliberation leads to acyclic collective preferences. In other words, variants of Theorems 6.11 and 6.12 are provable in the proof system of LDD. One might object that we cheated a little bit by using local variants of the assumptions of a common conceptual space and sufficient information and by weaking the assumption of opinionatedness. However, as the same results are proven under weaker conditions, Theorems 6.24 and 6.25 are in fact stronger than their counterparts from Section 6.3.1.

Chapter 7

Conclusion

The goal of this thesis was twofold. First and foremost, this thesis aimed at developing a formal framework for democratic deliberation, i.e. for a political deliberation that is ideal from the perspective of deliberative democracy. Furthermore, this thesis aspired to use this formal framework to study the philosophical claim that deliberation can play a role in circumventing social choice theory's impossibility results. This final chapter of the thesis contains some concluding remarks. We start by summarising the main achievements of this thesis and, afterwards, we discuss several ideas for future work.

The primary goal of this thesis was the development of a formal framework for democratic deliberation. According to deliberative democrats, an essential feature of democratic deliberation is that the participants justify their preferences to one another. Deliberative democracy does not see preferences as fixed and inexplicable, but considers them to be derived from something more fundamental. Therefore, this thesis started by developing a formal framework for preference formation. Our models for preference formation combine epistemic doxastic plausibility models with the framework of Dietrich and List (2013a, 20313b), who define preferences in terms of the motivational state of the agent and his meta-level preferences over the properties that hold of the alternatives. As our models combine the two, we were able to define an agent's preferences in terms of (i) his knowledge or belief, (ii) his motivational state and (iii) his meta-level preferences over the properties that hold of the alternatives or, alternatively, the reasons that can be used to justify one's preference for a certain alternative. By doing so, we have contributed to the preference formation literature, because to the best of our knowledge other formal accounts of preference formation only take into account two of the three above mentioned factors.

In order to obtain a formal framework for democratic deliberation, these preference formation models were combined with the model transformer [!] for deliberation. Thus, given a model of some initial, pre-deliberative situation, the model for the situation after deliberation is obtained by applying [!] to the initial model. This model transformer captures the spirit of democratic deliberation, because it models the fact that (i) all agents share their information and their perspectives with one another and (ii) all the agents accept the information and perspectives provided by the others. Adherents of deliberative democracy acknowledge that deliberation without representation is impossible. In most institutional political debates, the party representatives speak in turn and they share their knowledge and their perspectives. Therefore, the first item is in line with deliberative democracy. The second item also captures the deliberative democratic spirit, because, in a democratic deliberation, people justify their preferences on the basis of mutually acceptable and generally accessible reasons. Because of this mutual acceptability, it makes sense that the agents take the reasons and perspectives provided by others into account.

Our formal framework for democratic deliberation affirms two claims that are made in almost all articles and books on deliberative democracy. Firstly, it shows that preferences might change as a result of deliberation, because the participants are confronted with new information and new per-

spectives. Secondly, it shows that deliberation ensures a better understanding among the agents by inducing a common motivational state.

The second goal of this thesis was to use our formal framework to investigate the philosophical claim that deliberation can play a role in circumventing social choice theory's impossibility results. In particular, we studied the claim that deliberation provides a solution to Arrow's theorem by ensuring single-peaked preferences. Our formal results show that if we have (i) a common conceptual space and (ii.a) sufficient information about which properties hold of which alternatives or (ii.b) opinionatedness, then deliberation ensures single-peaked preferences provided that we collapse the options between which the agents are indifferent. Deliberation achieves this by turning virtual normative meta-agreement into actual normative meta-agreement and by enforcing factual meta-agreement.

The condition of opinionatedness is relatively undemanding, because it solely asks each participant to have an opinion on whether a certain property holds of an alternative or whether a certain reason can be used to justify one's preferences for that alternative. The sufficient information condition requires distributed knowledge of which motivationally salient properties hold of which alternatives. This is more demanding than opinionatedness, but not unlikely in the case of democratic deliberation, because we may expect our politicians to be well-informed. The common conceptual space requirement says that all agents conceptualise the issue at stake in the same way. More precisely, it says that all agents have a favourite property package and that the further the motivationally salient property package corresponding to a certain alternative is removed from the optimal one, the less good that option is. Although debatable, this is not at all untenable in the n-dimensional case. However, in order to prove single-peakedness, we need to assume one-dimensionality. And it is not at all clear why there would be one dimension with respect to which all agents order their property packages. Issue complexity often forbids this. Thus, the formal requirement of a common conceptual space is only likely in cases where there is one natural dimension accompanying the issue at stake, for example the left-right ordering in politics. Therefore, the right conclusion that follows from our formal results is the following: Given the fact that we are dealing with a single dimensional issue, it is likely that deliberation induces single-peaked preferences and thus helps to avoid Arrow's impossibility result, because the conditions needed to show that deliberation ensures single-peaked preference profiles are relatively undemanding.

For future work, there are two main directions that can be taken. The first one involves expanding the framework in such a way that other important aspects from deliberative democracy are taken into account as well, in order to get an even more precise model of the ideal deliberative situation. The second one is about moving away from this ideal situation by modifying the framework in such a way that it can model events that happen in real life communicative situations. We briefly discuss both directions.

From the perspective of deliberative democracy, a deliberation is a process in which the participants justify their views and preferences to one another and in which they learn new information, come into contact with new perspectives and are critically questioned on their own views. Our framework models that, in a deliberation, people justify their preferences to one another and acquire new information and new perspectives. What our framework is not able to model, however, is (i) that agents justify their views to one another and (ii) that their views and beliefs are critically examined. A promising way to do this is to use an evidence logic similar to the ones proposed by Baltag et al. (2012, 2014), which are based on Artemov's (2008) justification logic. More specifically, one could extend the language of LPF with a set of evidence terms and a relation saying which propositions are supported by which pieces of evidence. This allows us to define explicit knowledge and explicit belief as follows: an agent explicitly knows/believes a proposition φ iff he knows/believes φ and has evidence for φ. This setting enables us to model the fact that agents justify their views to one another, by asking them to not only share their explicit knowledge but also the supporting evidence.[1]

[1]Suppose an agent explicitly knows a proposition φ and the evidence he has for this is t. The model transformer $[!t]$, proposed by Baltag et al. (2012), is able to model this as it represents an update with a piece of hard evidence. In order to model deliberation as a process in which agents share all their explicit knowledge together with the supporting evidence and all their perspectives, we have to do two things. First, we have to generalise $[!t]$ in the same way as $[!]$ generalises $[!\varphi]$. In other words, we have to define an operation which models the fact that all agents share all their hard

Furthermore, it enables us to model the fact that agents are critically questioned on their own views, as this setting works with agents that are not logically omniscient. Critically questioning someone with respect to his views could mean a lot of things, but deliberative democrats agree that it at least encompasses the idea that agents are made aware of the (logical) implications of their judgements. In a framework with logically omniscient agents, this can never be modelled because the problem is not reflected in the framework. However, as the proposed framework deals with non-logical omniscience, possibilities for modelling this open up.[2]

Up till now, we have only considered deliberative situations that were ideal from the perspective of deliberative democracy. An interesting route for future work would be to move away from this ideal situation and to enable our formal framework to model things that happen in real life deliberative situations. Two things that often happen in real life, but are forbidden in a democratic deliberation, are (i) framing effects and (ii) the effect that only one or a few agents speak up, for instance due to power differences. The first is forbidden by democratic deliberation, because agents have to justify their preferences on the basis of mutually acceptable reasons. If others are confronted with these reasons, mutual acceptability forces them to take these reasons into account. This precludes framing effects. The second is forbidden because in a true democratic deliberation all participants have an equal say. However, if we want to model these two things, the logic of preference formation is still a good starting point. In a true democratic deliberation, the set of motivationally salient properties can only grow because every perspective is taken into account. Framing has the opposite effect. It makes one perspective or one way of looking at reality dominant. In order to model that a certain property becomes motivationally salient for agent i or, on the contrary, that it loses salience, we can use the model transformers $[+_i P]$ and $[-_i P]$ which respectively add or delete property P from i's motivational state.[3] For changes in information, we can use $[!\varphi]$ to model the fact that φ is publicly announced or $[!j]$, as defined in Chapter 3, to model the fact that agent j speaks up. A framework with model transformers like $[+_i P]$, $[-_i P]$ and $[!j]$ can *reflect* real life situations, like the fact that only one or a few agents speak up, that a property becomes motivationally salient for some agent or that it loses salience. However, if we want prove things about deliberation in our formal framework, we need to define specific deliberation protocols which formally state who speaks up, when a property is added and when it is deleted from the agent's motivational state and so forth. Defining specific deliberation protocols and developing a logic for them is another great challenge for the future.

In conclusion, our formal framework forms a nice starting point for modelling political deliberation as aspired by deliberative democracy, since it reflects the fact that, in a deliberation, agents justify their preferences to one another and are confronted with new information and new perspectives, which might lead to preference change. For the future, it might be interesting to refine this model so that it can also account for the justification of views and for fact that deliberation makes agents aware of the implications of their judgements. Furthermore, it might be interesting to adjust our models in such a way that we can model real-life situations. Both of these ways give us more insight into which effects of deliberation can be proven under which assumptions. One might argue that complex situations like deliberation can never be adequately captured in a formal model. And although we believe this is true, it is important that we keep trying. Because of its structured approach and the need to make all possibly hidden assumptions explicit, mathematical reasoning can help shine a new light on philosophical and sociological issues.

---

evidence. Afterwards, we have to modify this operation in such a way that it also takes into account that all agents share all the properties that they consider to be motivationally salient. There is, however, an important problem in a framework that works with explicit knowledge and that is that deliberation does not ensure that distributed knowledge is realised, i.e. that agents end up with the same epistemic indistinguishability relation. The reason for this is that agents share their explicit knowledge, whereas the epistemic indistinguishability relations reflect their implicit knowledge.

[2] In fact, Baltag et al. (2012) introduced the model transformer $t \otimes s$ reflecting the event that an agent combines evidence term t and s and applies modus ponens.

[3] These model transformers are similar to, and indeed inspired by, the ones discussed in Van Benthem and Velázquez-Quesada (2010).

Bibliography

Aldred, J. (2004). Social choice theory and deliberative democracy: A comment. *British Journal of Political Science*, *34*, 747–752.

Arrow, K. (1963). *Social Choice and Individual Values*. John Wiley and Sons.

Artemov, S. (2008). The logic of justification. *Review of Symbolic Logic*, *1*, 477–513.

Aumann, R. (1976). Agreeing to disagree. *Annals of Statistics*, *4*, 1236–1239.

Baltag, A., Moss, L., & Solecki, S. (1998). The logic of public announcements, common knowledge and private suspicions. *Proceedings of the 7th Conference on Theoretical Aspects of Knowledge and Rationality*, (pp. 43–56).

Baltag, A., Renne, B., & Smets, S. (2012). The logic of justified belief change, soft evidence and defeasible knowledge. In L. Ong, & R. de Queiroz (Eds.) *Logic, Language, Information and Computation*, vol. 7456 of *Lecture Notes in Computer Science*, (pp. 168–190). Springer-Verlag.

Baltag, A., Renne, B., & Smets, S. (2014). The logic of justified belief, explicit knowledge and conclusive evidence. *Annals of Pure and Applied Logic*, *165*, 49–81.

Baltag, A., & Smets, S. (2006a). Dynamic belief revision over multi-agent plausibility models. *Proceedings of the 7th European Conference on Logic and the Foundations of Game and Decision Theory*.

Baltag, A., & Smets, S. (2006b). The logic of conditional doxastic actions: a theory of dynamic multi-agent belief revision. *Proceedings of the Workshop on Rationality and Knowledge (ESSLLI)*.

Baltag, A., & Smets, S. (2008). A qualitative theory of dynamic interactive belief revision. In G. Bonanno, W. van der Hoek, & M. Wooldridge (Eds.) *Logic and the Foundations of Game and Decision Theory*, vol. 3 of *Texts in Logic and Games*, (pp. 9–58). Amsterdam University Press.

van Benthem, J. (2002). One is a lonely number: on the logic of communication. In P. Koepke, Z. Chatzidakis, & W. Pohlers (Eds.) *ASL Lecture Notes in Logic*, vol. 27, (pp. 96–129). AMS Publications.

van Benthem, J., & Velázquez-Quesada, F. (2010). The dynamics of awareness. *Synthese*, *177*, 5–27.

Berlin, I. (1969). *Four Essays on Liberty*. Oxford University Press.

Black, D. (1948). On the rationale of group decision making. *Journal of Political Economy*, *56*, 23–34.

Blackburn, P., de Rijke, M., & Venema, Y. (2001). *Modal Logic*. Cambridge University Press.

Bohman, J. (1998). The coming age of deliberative democracy. *Journal of Political Philosophy*, *6*, 400–425.

Bohman, J., & Rehg, W. (1998). *Deliberative Democracy*. The MIT Press.

Cohen, J. (1989). Deliberation and democratic legitimacy. In A. Hamlin, & P. Pettit (Eds.) *The Good Polity*, (pp. 17–34). Basil Blackwell.

de Condorcet, N. (1785). *Essai sur l'Application de l'Analyse à la Probalité des Décisions Rendue à la Pluralité des Voix*. Imprimerie Royale.

Dégremont, C., & Roy, O. (2012). Agreement theorems in dynamic-epistemic logic. *Journal of Philosophical Logic*, *41*, 735–764.

Dietrich, F., & List, C. (2013a). A reason-based theory of rational choice. *Nous*, *47*, 104–134.

Dietrich, F., & List, C. (2013b). Where do preferences come from? *International Journal of Game Theory*, *42*, 613–637.

van Ditmarsch, H. (2000). *Knowledge Games*. ILLC Dissertation Series 2000-06. Grafimedia Groningen University.

Dryzek, J. (2000). *Deliberative Democracy and beyond: Liberals, Critics, Contestations*. Oxford University Press.

Dryzek, J., & List, C. (2003). Social choice theory and deliberative democracy: A reconciliation. *British Journal of Political Science*, *33*, 1–28.

Elsholtz, C., & List, C. (2005). A simple proof of Sen's possibility theorem on majority decisions. *Elemente der Mathematik*, *60*, 45–56.

Elster, J. (1986). The market and the forum. In J. Elster, & A. Hylland (Eds.) *Foundations of Social Choice Theory*, (pp. 103–132). Cambridge University Press.

Elster, J. (1998). *Deliberative Democracy*. Cambridge University Press.

Fagin, R., Halpern, J., Moses, Y., & Vardi, M. (2003). *Reasoning about Knowledge*. The MIT Press.

Fagin, R., Halpern, J., & Vardi, M. (1992). What can machines know? *Journal of the Association for Computing Machinery*, *39*, 328–376.

Farrar, C., Fishkin, J., Green, D., List, C., Lushkin, R., & Paluck, E. (2010). Disaggregating deliberation's effects: an experiment within a deliberative poll. *British Journal of Political Science*, *40*, 333–347.

Gaertner, W. (2001). *Domain Restrictions in Social Choice Theory*. Cambridge University Press.

Gerbrandy, J. (1999). *Bisimulations on Planet Kripke*. ILLC Dissertation Series 1999-01. University of Amsterdam.

Gerbrandy, J., & Groeneveld, W. (1997). Reasoning about information change. *Journal of Logic, Language and Information*, *6*, 147–169.

Gibbard, A. (1973). Manipulation of voting schemes: a general result. *Econometrica*, *41*, 587–601.

Gutmann, A., & Thompson, D. (1996). *Democracy and Disagreement*. Princeton University Press.

Gutmann, A., & Thompson, D. (2002). *Why Deliberative Democracy?*. Princeton University Press.

Habermas, J. (1994). Three models of democracy. *Constellations*, *1*, 1–10.

Habermas, J. (1996). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press.

Hendricks, V. (2006). *Mainstraim and Formal Epistemology*. Cambridge University Press.

Hintikka, J. (1962). *Knowledge and Belief: an Introduction to the Logic of the Two Notions*. Cornell University Press.

Hintikka, J. (1969). *Models for Modalities: Selected Essays*. D. Reidel.

Klein, P. (1971). A proposed definition of propositional knowledge. *Journal of Philosophy*, *68*, 471–482.

Lehrer, K. (1990). *Theory of Knowledge*. Routledge.

Lehrer, K., & Paxson, T. (1969). Knowledge: Undefeated justified true belief. *Journal of Philosophy*, *66*, 225–237.

Lenzen, W. (1979). Epistemologische betrachtungen zu [S4,S5]. *Erkenntnis*, *14*, 33–56.

Lenzen, W. (1980). *Glauben, Wissen und Wahrscheinlichkeit: Systeme der epistemischen Logik*. Springer Verlag.

Lewis, D. (1969). *Convention: a philosophical study*. Blackburn.

List, C. (2002). Two concepts of agreement. *The Good Society*, *11*, 72–79.

List, C. (2007). Deliberation and agreement. In S. Rosenberg (Ed.) *Deliberation, Participation and Democracy: Can the People Govern?*, (pp. 64–81). Palgrave.

List, C., Lushkin, R., Fishkin, J., & McLean, I. (2000). Can deliberation induce greater preference structuration: Evidence from deliberative opinion polls. *Proceedings of the American Political Science Association*.

List, C., Lushkin, R., Fishkin, J., & McLean, I. (2013). Deliberation, single-peakedness and the possibility of meaningful democracy: Evidence from deliberative polls. *The Journal of Politcs*, *75*, 80–95.

Liu, F. (2011). *Reasoning about Preference Dynamics*. Springer.

Miller, D. (1992). Deliberative democracy and social choice. *Political Studies*, *40*, 54–67.

van Mill, D. (1996). The possibility of rational outcomes from democratic discourse and procedures. *The Journal of Politics*, *58*, 734–752.

Mueller, D. (1989). *Public Choice II*. Cambridge University Press.

Ottonelli, V., & Porello, D. (2013). On the elusive notion of meta-agreement. *Politics, Philosophy and Economics*, *12*, 68–92.

Peterson, M. (2009). *An Introduction to Decision Theory*. Cambridge University Press.

Plaza, J. (1989). Logic of public communications. *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, (pp. 201–216).

Riker, W. (1982). *Liberalism against Populism: a Confrontation between the Theory of Democracy and the Theory of Social Choice*. W.H Freemann.

Rott, H. (2004). Stability, strength and sensitivity: Converting belief into knowledge. *Erkenntnis*, *61*, 469–493.

Satterthwaite, M. (1975). Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, *10*, 187–217.

Sen, A. (1966). A possibility theorem on majority decisions. *Econometrica*, *34*, 491–499.

Sen, A. (1986). Social choice theory. In K. Arrow, & M. Intriligator (Eds.) *Handbook of Mathematical Econonomics*, vol. 3, (pp. 1073–1181). North-Holland.

Stalnaker, R. (1991). The problem of logical omniscience 1. *Synthese*, *89*, 425–440.

Stalnaker, R. (1999). The problem of logical omniscience 2. In R. Stalnaker (Ed.) *Context and Content: Essays on Intentionality in Speech and Thought*, (pp. 255–273). Oxford University Press.

Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies*, *128*, 169–199.

Steup, M., & Sosa, E. (2005). *Contemporary Debates in Epistemology*. Blackwell.

von Wright, G. (1963). *The Logic of Preference*. Edinburgh University Press.