Manipulating the Manipulators: Richer Models of Strategic Behavior in Judgment Aggregation

MSc Thesis *(Afstudeerscriptie)*

written by

Zoi Terzopoulou
(born August 6, 1993 in Athens, Greece)

under the supervision of **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:**
June 15, 2017	Prof Jan van Eijck
	Dr Ulle Endriss
	Prof Sonja Smets
	Prof Ronald de Wolf

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Contents

Abstract

Judgment Aggregation is a formal framework of collective decision making. When agents that belong to a group express their individual opinions on a set of logically interconnected issues, a good rule is required in order to combine these opinions and induce a representative collective judgment for the group. However, it is often the case that some agent may find a way to achieve a preferable outcome for herself, by reporting a dishonest opinion. This kind of strategic behavior, namely manipulation, constitutes the heart of this thesis. Until now, researchers have been making two very strong assumptions in the context of Judgment Aggregation: first, that all the agents fully know the truthful opinions of all the other members of their group; and second, that every agent thinks that everyone else is always truthful. In this thesis we start with enriching the existing model of Judgment Aggregation in a twofold manner: we account for partially informed agents who behave strategically under various types of uncertainty, as well as for higher-level strategic reasoners, who reflect on the reasoning of their peers. We employ analytical methods, and we explore how the aforementioned assumptions affect the agents' choices and insincere acts in Judgment Aggregation. Moreover, after investigating in depth single-round aggregation processes, a model of Iterative Judgment Aggregation is developed. In our third model, the agents have the possibility to change their initially submitted judgments in a sequential random order, while they are (maybe partially) observing the acts of their peers. We study (a) whether common aggregation rules in iteration reach equilibria states, (b) how fast they do, and (c) what the potential benefits of strategic behavior are, for a group of agents *en masse*.

Acknowledgments

Having read the acknowledgments of other theses before, I thought that mentioning one's supervisor first is just a matter of good manners. But now I know; if other students have had an experience even remotely similar to mine, then their supervisor's principal place in the acknowledgments is more than well-deserved. This was meant to say: "Thank you, Ulle". For believing in me before I even learned how to spell "Gibbard-Satterthwaite", for pushing me forward with your comments in every single meeting, for carefully checking my work and teaching me the importance of aesthetics in writing, and for reminding me to give some examples, in between of my definitions.

Studying in Amsterdam as an independent adult was made possible thanks to the Amsterdam Science Talent Scholarship program, and the Onassis Foundation, that generously financed me. I am very grateful to the people who made that happen. I would also like to thank Jan, Ronald, and Sonja, for being in my committee and taking the time to read this thesis. Moreover, thank you, Sonja, for being there for me, available for discussion and advice every time I needed it during the Master's.

I owe a big thanks to the members of the Computational Social Choice group: Ronald, Sirin, Weiwei. Thank you for the interest you showed in my work since the very first moment, and thank you for the support and motivation that you were constantly offering, even without realizing. Next, Laura, I cannot express how different the last two years would have been without you. You have been an italian sister, mother, and best friend for me. I am honored to have met you, and I am very happy to be the first one from our year to acknowledge how amazing your cakes are. Bonan, thanks for your positive energy, that was filling our flat and was often giving me a new perspective on things. Levin and Lisa, thank you for letting me talk to you about anything that worried me, from science to life, and for sharing your beautiful ideas. Natalia, thank you for always being willing to travel. I wish I could mention every other person that did the Master of Logic with me, but this thesis is already long. You know who you are: your enthusiasm and kindness made this experience unique!

Finally, to my friends and family from Greece: thank you for making sure that I always remember how much trust you have in my potentials, either by giving me the warmest welcome when I come back, or by being there, loving faces behind a screen. To my mother, my father and my brother: you are able to make the dark dutch sky look lighter. Thank you, I will need it. Μαμά, Μπαμπά, Παναγιώτη: Ευχαριστώ!

Chapter 1

Introduction

Judgment Aggregation is a formal framework that has its roots in the area of Social Choice Theory and models a particular class of collective decision making problems emerging from Political Science, Philosophy, Law, but also Economics and Artificial Intelligence. The situations that Judgment Aggregation is concerned with involve individuals (agents) expressing binary (*Yes* / *No*) opinions on possibly interconnected issues, upon which a collective decision has to be made. Since the first appearance of democratic societies, citizens have been participating in various procedures with the aim of aggregating individual judgments and making collective choices. The individuals in a Judgment Aggregation setting may also be people engaging in everyday-life, small-scale collective decision making. For example, groups of friends or families, boards of companies, or juries of courts. On the other hand, computerized intelligent agents may also make use of aggregation methods to make collective decisions. In Section 1.1 we review the framework of Judgment Aggregation and we motivate its applications. In Section 1.2 we give an overview of this thesis, which investigates several aspects of strategic behavior in Judgment Aggregation, and we discuss our broader contributions to related fields.

1.1 About Judgment Aggregation

Consider a committee consisting of three professors that has to evaluate the Master's thesis of a student.[1] The committee is expected to collectively pass or fail the student, depending on the individual judgments of its members. Specifically, according to the regulations of the Master's program, the student can obtain a final passing grade (p) if and only if her thesis is considered satisfactory in both the following two criteria: "difficulty" (d) and "writing" (w). Suppose now that the professors express the judgments depicted on Table 1. A Yes-judgment on a criterion means that the thesis is evaluated

[1]The thesis-evaluation example is inspired by the *doctrinal paradox*, initially observed by Kornhauser and Sager (1986) in the context of Legal Theory, and later discussed by Pettit (2001) (see also Mongin, 2012).

	d	w	p
Professor 1:	Yes	Yes	Yes
Professor 2:	Yes	No	No
Professor 3:	No	Yes	No

Table 1: The thesis-evaluation example.

as good enough regarding that particular criterion. Professor 1 liked the thesis a lot; he thinks that it satisfies both criteria and hence, that the student should pass. However, Professors 2 and 3 disagree in different ways. In Professor 2's opinion, even if the content of the thesis is advanced, its writing-style does not correspond to the Master's level, and this constitutes a sufficient reason for the student to fail. To the contrary, Professor 3 thinks that the thesis is well-written, but she suggests a failing final grade on the basis of the student's poor performance with regard to the difficulty criterion. In total, two out of the three members of the committee evaluate the thesis as satisfying on each one of the stated criteria. So, if a majority rule operates, focusing only on the evaluations of the professors concerning the thesis' difficulty and writing-style, then the student should pass. But if the committee's decision is based only on the final grades that the professors would assign individually, then a majority would prescribe failing the student. Finally, if the decision of the committee is induced by the opinion of the majority on each issue d, w, and p separately, then it will be inconsistent with respect to the regulations, suggesting a "Yes" considering the two criteria, but a "No" regarding the student's passing final grade.

The crucial role of the chosen aggregation method applied in an aggregation problem is now brought into light. Different procedures may prescribe totally opposite, even inconsistent, actions for a group. Abstracting from specific situations, the framework of Judgment Aggregation is able to study the common structure between different aggregation scenaria, and provide answers to questions like: Under which conditions does a given aggregation process result in "appropriate" collective outcomes? What kind of aggregation rules are guaranteed to always induce "good" group decisions? As a formal framework in its own right, Judgment Aggregation was introduced by List and Pettit (2002, 2004) (see Endriss (2016) and List (2012) for two recent reviews of the field). Creating a general basis for numerous types of collective decision making problems, it incorporates a large framework that has been developed independently: Preference Aggregation (that is most prominently exemplified in Voting Theory; see for instance the review by Zwicker, 2016), and is closely connected to Belief Merging (mainly studied in Computer Science, e.g., by Everaere et al., 2015).

Strategizing

The focal point of our work was already remarked upon by Arrow (1951b), in the introductory chapter of his book "Social Choice and Individual Values":

> Once a machinery for making social choices from individual tastes is established, individuals will find it profitable, from a rational point of view, to misrepresent their tastes by their actions [...]

Arrow's claim will become clearer considering again the thesis-evaluation situation. The committee has now chosen to use the first aggregation procedure that we mentioned in the example. That is, the student can pass if and only if, with respect to each criterion, more than half of the professors believe that her thesis is sufficiently good. Then, if the committee reports the judgments of Table 1, the student should indeed pass. However, imagine that during a meeting where the members of the committee discuss about their evaluations of the thesis, Professor 2 gets to know the exact opinions of her colleagues. Hence, she realizes that, if she truthfully declares her judgment, then the committee will pass the student, following the regulations. But this does not please Professor 2, who judges that the student should not obtain a Master's degree because her writing skills are disappointing. Smart as she is, Professor 2 then decides to lie; she pretends that, similarly to Professor 3, she found the difficulty-level of the thesis not satisfying. In that case, more than half of the professors claim that the student's performance on one criterion is insufficient, and as a result, the student will fail. At the end of the day, one individual had the chance to *manipulate* the decision making process, in order to obtain a more desirable outcome for her.

It is shown by Dietrich and List (2007c) that situations like the previous one are often unavoidable in Judgment Aggregation. More precisely, aggregation procedures that meet some reasonable properties are always *manipulable*, when the agents are fully aware of the judgments of their peers and they expect everyone else to be truthful.[2] Subsequently, scholars have been hunting through Dietrich and List's assumptions, for those conditions whose relaxation could reveal some more positive news (see Baumeister et al. (2017) for an overview). First, a successful approach (already pointed out in the same article of Dietrich and List, 2007c) relates to limiting the possible opinions that a group of agents can submit, viz., imposing *domain restrictions*.[3] Second, another renowned method, principally employed by computer scientists (e.g., Endriss et al., 2012), shows that deciding the manipulation problem of certain aggregation procedures is computationally hard, and hence practically impossible to happen (see also Bartholdi et al. (1989) for an older work in Voting).[4]

In this thesis we give an answer to the following natural *research question*: Do the negative implications of Dietrich and List's theorem hold for more refined assumptions concerning the information and the reasoning abilities of the agents?

[2]This result is analogous to the famous impossibility theorem of Gibbard (1973) and Satterthwaite (1975) in Voting Theory.

[3]A domain restriction for instance is *unidimensional alignment* (List, 2003), according to which the members of a group can be aligned in a left-to-right order, such that, for every issue upon which judgments are expressed, all the agents that have the same opinion are either on the left or on the right side of those that disagree with them.

[4]In Voting, alternative ways to escape the inevitability of manipulation have been attempted too, as for example, calculating that its probability of occurrence is small enough (e.g., Aleskerov et al., 2011).

1.2 About this Thesis

Our goal is to extend the basic model of Judgment Aggregation, by accounting for richer aspects of intelligent agents' (either human or artificial) interactions, and examining the correlations with their strategic attitudes. Our study is systematic, in the sense that it consistently investigates the manipulability of common aggregation procedures under various assumptions, yet flexible, since it allows for additional modifications. This thesis is purely formal, but motivated by psychological and behavioral facts. Chapter 2 presents the technical background, that is followed by the main body in Chapters 3, 4 and 5 (further explained below). We conclude in Chapter 6.

Chapter 3. In aggregation procedures that take place in large groups, or that involve confidential issues, etc., it is obvious that the assumption that the agents are fully informed about all the opinions of others is rather stringent. But how would agents reason when they are ignorant of a part of their peers' opinions? Can such uncertainty influence their incentives for manipulation? We design a model of Judgment Aggregation that embodies *partial information*, and we answer those questions.

Chapter 4. Once intelligent agents start reasoning about a decision-making situation, where other agents participate too, it is reasonable to assume that they will be tempted to reason about the reasoning of their peers. So, what happens if an agent realizes that a member of her group has a reason to report an insincere judgment? Are the first agent's possible incentives to manipulate affected? We investigate this issue by developing a model of *higher-level reasoning* in Judgment Aggregation.

Chapter 5. In practice, collective decision making procedures may take place in multiple rounds, instead of a single one. We study the manipulative acts of the agents in a model of *Iterative Judgment Aggregation*. The main questions that we address are: Do iterative aggregation processes reach a terminal state where no-one can profit by a unilateral deviation (i.e., an equilibrium)? Do they reach this state fast? And how beneficial is strategic behavior in the end for a society or a group as a whole?

Our work contributes to the wider collective decision making literature, by making explicit several assumptions that were implicit up till now, and consequently revealing the importance of accounting for multidimensional features of the agents' reasoning. Specifically, we broaden significantly the spectrum of situations that Judgment Aggregation can account for, by designing a full-fledged framework, inspired by well-tried models from the areas of (theoretical and experimental) Voting and Game Theory. Furthermore, this thesis relates to Political Science, because it rigorously analyzes commonly used methods in the field, from richer perspectives. Finally, our work contributes to the Artificial Intelligence literature, opening the way to many applications in the rapidly growing area of Multiagent Systems.

Chapter 2

Background

This chapter is meant to provide a baseline for the formal language of this thesis. We present the model of *formula-based Judgment Aggregation* (Section 2.1), discussing various *agendas* (Section 2.2) on which *aggregation rules* are applied (Section 2.3), and we examine the rules' normative appeal grounded on sets of *axioms* (Section 2.4). We also introduce an alternative but equivalent framework of Judgment Aggregation: *Binary Aggregation with Integrity Constraints* (Section 2.5). Next, individual *preferences* with regard to collective decision making are analyzed (Section 2.6), facilitating the transition of our focus to the study of *strategic behavior* (Section 2.7).

2.1 Formula-based Judgment Aggregation

We now present the basic model of *formula-based Judgment Aggregation*, which is going to be used throughout this thesis. We follow Baumeister et al. (2016); Endriss (2016) and List and Pettit (2002). Consider a finite set of individuals (agents) $\mathcal{N} = \{1, 2, \ldots, n\}$, with $n \geqslant 2$, that constitute a group whose judgments are to be aggregated into one collective decision. The issues that the agents express opinions upon are called *propositions* and are represented as formulas in classical Propositional Logic $\mathcal{L}$. In the sequel we will assume acquaintance of the reader with the basic concepts of Propositional Logic.[5] We define $\sim\phi$ the complement of formula $\phi \in \mathcal{L}$ as follows: $\sim\phi := \phi'$ if $\phi = \neg\phi'$ for some formula ϕ', and $\sim\phi := \neg\phi$ otherwise. The domain of the decision making is an *agenda* Φ, where $\varnothing \neq \Phi \subseteq \mathcal{L}$, and Φ is closed under complementation (i.e., for every formula $\phi \in \mathcal{L}$, if $\phi \in \Phi$, then $\sim\phi \in \Phi$).

Each individual i's *judgment set* (or *opinion*) $J_i \subseteq \Phi$ is the set of propositions that she accepts in Φ. We assume that all individual judgment sets are *consistent*, i.e. consistent sets of propositions in the standard sense of Logic, and *complete*, i.e. they

[5] An extended model of Judgment Aggregation which captures propositions expressed in richer logical languages, such as Predicate Logic, Modal Logic, and Multi-valued or Fuzzy Logic is developed by Dietrich (2007), but goes beyond the scope of our work.

contain at least one member of each proposition-complement pair in Φ.[6] The set of all the consistent and complete subsets of the agenda Φ, which are the possible judgment sets of the agents, is denoted as $\mathcal{J}(\Phi)$. In addition, we say that two judgments J and J' *agree* on proposition ϕ if and only if ϕ belongs to both or neither of them, and they *disagree* otherwise. A *profile* $\boldsymbol{J} = (J_1, \ldots, J_n) \in \mathcal{J}(\Phi)^n$ is a vector of all the agents' judgment sets, and $\boldsymbol{J}_{-i}$ stands for the partial profile of judgments of the whole group except for agent i. We denote with $N_\phi^{\boldsymbol{J}}$ the set of agents who accept ϕ in $\boldsymbol{J}$, and with $\overline{N_\phi^{\boldsymbol{J}}}$ its set-theoretic complement, i.e., the set of agents who reject ϕ and accept $\sim\phi$.

Finally, an aggregation rule F is a function that maps every profile $\boldsymbol{J}$ of individual judgment sets to a set $F(\boldsymbol{J})$ of collective decisions, which are (not necessarily complete or consistent) subsets of the agenda. When $F(\boldsymbol{J})$ is a singleton, i.e., when $F : \mathcal{J}(\Phi)^n \to 2^\Phi$, the rule F is called *resolute*. In general, the practical aim of an aggregation rule is to provide us with an answer about what the collective decision of the agents is, or should be. To that end, resolute rules are essential. Hence, for what follows we will always work with them (even when it is not mentioned, all the definitions and results stated hereafter will refer to resolute rules); when resoluteness is not the case, we will additionally consider a *lexicographic tie-breaking rule* to resolve the ties between the suggested collective opinions. To be precise, a lexicographic tie-breaking rule ranks all the possible subsets of the agenda (the sets in 2^Φ) using a linear order, and if the result of an aggregation rule consists of at least two judgment sets, then it dictates the choice of the one ranked higher. Intuitively, the prescribed linear order can be thought to represent the preference of a leading agent, whose desires are to be satisfied when the aggregation rule does not suggest a unique solution. For example, the president of a company may be asked to resolve the ties when the judgments of the board clash.[7]

2.2 Agendas

The set of propositions on which a collective decision is to be made, i.e., the agenda of an aggregation problem, is a prime constituent of the Judgment Aggregation framework. Several restrictions can be imposed on the structure of an agenda, in order to better capture the essence of specific aggregation situations, or to establish certain desirable properties of the aggregation, such as logical consistency of the collective

[6] These conditions have been relaxed by some authors. For example, Dietrich and List (2008) consider a framework in Judgment Aggregation with incomplete individual opinions.

[7] An alternative technique of breaking ties could exploit *random tie-breaking*. In this thesis however, we insist on the use of lexicographic tie-breaking orders. One of our main reasons is that breaking ties with a linear order satisfies the *independence of irrelevant alternatives principle* (Ray, 1973). The independence of irrelevant alternatives principle, also known as Sen's property α (Sen, 1969, 1970), states that if an alternative J is chosen from a set S, and J is also an element of a subset S' of S, then J must be chosen from S. That is, eliminating some of the unchosen alternatives should not affect the selection of J. We find this condition normatively desirable as far as a tie-breaking rule (and in extension an aggregation rule) is concerned.

outcome. The former case is further discussed below; the latter is part of a research focus that goes by the name *safety of the agenda*, and will not be relevant for this thesis (but for the interested reader we refer to Endriss et al., 2010, 2012).

Two examples of special agendas are the *conjunctive* and the *disjunctive* agendas.

Definition 2.1. A *conjunctive* agenda consists of a set of *premises*, which are propositional variables $a_1, \ldots, a_k$, and a *conclusion* c, where $c \leftrightarrow (a_1 \wedge \ldots \wedge a_k)$, as well as their negations.

The thesis-evaluation example in the Introduction uses an instance of a conjunctive agenda with two premises. Analogously, the conclusion of a disjunctive agenda is equivalent to the disjunction of all the premises. Naturally, conjunctive and disjunctive agendas appear in situations in which a final decision has to be made on a conclusion, but the reasons that lead to that choice, described by the premises, are also important.

A more general class of agendas, which includes the ones mentioned above, is that of the *path-connected* agendas, defined by Dietrich (2007) and related to the concept of *total-blockedness* (Nehring and Puppe, 2007). In the words of Dietrich and List, an agenda of propositions under consideration is path-connected if any two propositions in the agenda are logically connected with each other, either directly or indirectly, via a sequence of (conditional) logical entailments. Formally, proposition ϕ *conditionally entails* proposition ψ if $\{\phi, \neg\psi\} \cup \Psi$ is consistent for some $\Psi \subseteq \Phi$ consistent with ϕ and with $\neg\psi$. The agenda Φ is path connected if for all propositions $\phi, \psi \in \Phi$ that are not tautologies nor contradictions, there is a sequence of propositions $\phi_1, \phi_2, \ldots, \phi_k \in \Phi$ with $\phi = \phi_1$ and $\psi = \phi_k$ such that ϕ_{i-1} conditionally entails ϕ_i, for every $i \in \{2, \ldots k\}$. Almost every standard and interesting agenda is path-connected.

Whenever we assume that the examined agendas are of a specific kind, it will be made clear in the text.

2.3 Rules

The core of the theory of Judgment Aggregation consists of course of the aggregation rules. The first encounter of the reader with aggregation rules in this thesis has been in the Introduction, where we mentioned different possible methods to aggregate the judgments of the committee in the thesis-evaluation example. One of them was the *majority rule*. Re-stated slightly more formally here, the majority rule accepts each issue in the agenda if and only if at least half of the agents accept it (for the strict case, an acceptance by more than half of the group is required). Generalizing the majority rule, the *quota rules* may be defined. According to them, a proposition is part of the collective decision if and only if at least some proportion of the agents (exceeding the relevant *quota*) agrees with it.

Definition 2.2. Consider a fixed quota $q_\phi \in [0, n+1]$ for every proposition ϕ in the agenda Φ. Then, the *quota rule* F^q is such that, for any profile of judgments $\boldsymbol{J}$, $\phi \in F^q(\boldsymbol{J})$ if and only if $|N_\phi^{\boldsymbol{J}}| \geqslant q_\phi$.

A peculiar category of aggregation rules, inspired by the quota rules but being less appealing to our intuitions, consists of the *parity rules*. The odd(even)-parity rule accepts a proposition ϕ if and only if an odd (even) number of agents accepts it.

However, we already saw in the Introduction that when the majority rule is implemented, the collective outcome may end up being logically inconsistent (recall the thesis-evaluation example: the majority of the professors accepts that the student passes each evaluation criterion, but the majority also wants the student to fail). Such unfortunate cases hold for the other quota rules too. To resolve this problem, an effective method is the use of the *premise-based procedure*. We consider this rule with regard to conjunctive agendas (the definition for disjunctive agendas is similar).

Definition 2.3. Consider a conjunctive agenda Φ with $\Phi^p := \{a_1, \ldots, a_k\}$ the set of its premises and c its conclusion. Having a profile $\boldsymbol{J}$ of judgments of n agents, we define the outcome of the *premise-based procedure* F^{pr} as follows: First, a collective decision is made on the premises with respect to the majority rule, that is, for all $a_i \in \Phi^p$, $a_i \in F^{pr}(\boldsymbol{J})$ if $|N^{\boldsymbol{J}}_{a_i}| > \frac{n}{2}$, and $\neg a_i \in F^{pr}(\boldsymbol{J})$ otherwise. Then, $c \in F^{pr}(\boldsymbol{J})$ if $a_i \in F^{pr}(\boldsymbol{J})$ for all $a_i \in \Phi^p$, and $\neg c \in F^{pr}(\boldsymbol{J})$ otherwise.

Since $c \leftrightarrow (a_1 \wedge \ldots \wedge a_k)$, a consistent outcome is then guaranteed. Alternatively, one could also use the *conclusion-based procedure*. In that case, the opinion of the group is aggregated again using the majority rule, only regarding the conclusion, and individual judgments on the premises are ignored. It is clear from the Introduction that the premise-based procedure and the conclusion-based procedure may produce diametrically opposite results. As demonstrated in our thesis-evaluation example, the former rule would advise giving to the student a passing grade, because the student is good enough in each evaluation criterion separately, while the latter would fail her, following the opinion of the majority of the professors on the conclusion. Hence, it is easy to understand why choosing between the premise-based and the conclusion-based procedure has generated numerous philosophical debates among political scientists and economists (e.g., Chapman, 2002; Dietrich and List, 2007c; Elster, 1998; Pettit, 2001). In several parts of this thesis we will contribute to this discussion.

The most trivial and objectionable aggregation rule is the *dictatorship*. Living up to its name, the dictatorship is connected to a single agent, the dictator, whose individual judgment is taken to be the collective judgment independently of the input profile. The dictatorship vacuously guarantees that the opinion of the group, that is exactly the opinion of the dictator, will satisfy all the nice properties of individual opinions, like completeness and consistency.

Fortunately, there are several other aggregation rules that present the above advantage. For example, the *distance-based* or *Kemeny rule* F^K (Endriss et al., 2012; Pigozzi, 2006; Miller and Osherson, 2009) takes into account a notion of *distance* between judgment sets and specifies the winner(s) to be the complete and consistent judgment set(s) that is (are) closer to the profile of the agents' opinions. Specifically, the *Hamming-distance* $H(J, J')$ of two judgment sets $J, J' \in 2^{\Phi}$ is defined as the number of formulas in Φ on which they disagree. Formally,

$$H(J, J') := |\Phi| - |J \cap J'| - |\overline{J} \cap \overline{J'}|$$

Then, the *Hamming-distance* $H(\boldsymbol{J}, J)$ of the profile $\boldsymbol{J} = (J_1, \ldots, J_n)$ and the judgment set J sums the Hamming distances of all judgment sets $J_1, \ldots, J_n$ and J.

$$H(\boldsymbol{J}, J) := \sum_{i \in \{1,\ldots,n\}} H(J_i, J)$$

Given the profile $\boldsymbol{J}$, we have that $F^K(\boldsymbol{J}) = \underset{J \in \mathcal{J}(\Phi)}{\arg\min}\, H(\boldsymbol{J}, J)$. This is a direct translation of the Kemeny rule that appeared in Voting Theory many years earlier in the work of Kemeny (1959). Endriss and Grandi (2014) later introduced the class of *representative-voter rules*, which choose a voter that better represents the input profile according to specific criteria and make her judgment be the collective decision. A natural choice for a representative voter is the one realized by the *average-voter rule* F^{av}. According to it, the collective outcome is taken to be the individual judgment(s) in the profile $\boldsymbol{J}$ for which the Hamming distance to $\boldsymbol{J}$ is minimized (note that contrary to the Kemeny rule, the decision of the group is now selected among the submitted judgment sets only). Formally, for the profile $\boldsymbol{J} = (J_1, \ldots, J_n)$, we have that $F^{av}(\boldsymbol{J}) = \underset{J_i : i \in \{1,\ldots,n\}}{\arg\min}\, H(\boldsymbol{J}, J_i)$.

Lastly, another aggregation rule directly inspired by Voting is the *plurality rule*.

Definition 2.4. The *plurality rule* considers the aggregated outcome to be the judgment set(s) submitted by the largest number of individuals.

In the framework of Voting, the agents are asked to vote for their favorite candidate, and the candidate with the most supporters wins. Obviously, the plurality rule presents severe theoretical limitations. For instance, only each voter's top alternative is represented by the voting procedure, and in settings with few voters but many alternatives, it is very probable that several candidates receive the same amount of support, only by one voter, and hence the tie-breaking process has to play an important role to decide a single winner. Nonetheless, the plurality rule is widely used in practice, employed by many electoral systems (for example in the general elections of the United Kingdom and the national elections of Canada). Furthermore, in a different context motivated by applications to crowdsourcing, Caragiannis et al. (2014) show that an aggregation rule that they call *modal ranking*, and that is equivalent to the plurality rule (for the alternatives being different rankings), is the unique one satisfying specific desirable properties. Having the above facts in mind, in this thesis we wish to explore in depth the status of the plurality rule in Judgment Aggregation, focusing on the agents' strategic behavior on it.

2.4 Axioms

The axiomatic approach serves as a skeleton for both the descriptive and the normative analysis of systems in many branches of formalized Social Science, such as Logic, Decision and Social Choice Theory. In Judgment Aggregation, basic features of aggregation rules have been explored and defined via axioms for example by List and Pettit (2002). Descriptively, axiomatic characterizations provide a structured way to look into aggregation rules, helping us to compare them and better understand them. On the other hand, axioms are constructive from a normative perspective because they directly reflect properties that we wish our designed aggregation rule to adopt. Here, we refer to axiomatic characteristics of resolute rules only.

- We call the aggregation rule F *responsive* if it gives the chance to every proposition to be accepted by the group. Formally, if for every proposition ϕ that is not a tautology nor a contradiction, there exists a profile $\boldsymbol{J}$ such that $\phi \in F(\boldsymbol{J})$, and another profile $\boldsymbol{J}'$ such that $\phi \notin F(\boldsymbol{J}')$.

- The rule F is *anonymous* if it treats all individuals symmetrically, i.e., for any permutation $\pi : \mathcal{N} \to \mathcal{N}$, it holds that $F(J_1, \ldots, J_n) = F(J_{\pi(1)}, \ldots, J_{\pi(n)})$. Intuitively, anonymity is a stronger version of *non-dictatorship*.

- *Neutrality* for F suggests that all propositions ϕ, ψ in Φ are treated symmetrically, that is, for any profile $\boldsymbol{J}$, if $N_\phi^{\boldsymbol{J}} = N_\psi^{\boldsymbol{J}}$, then $\phi \in F(\boldsymbol{J}) \Leftrightarrow \psi \in F(\boldsymbol{J})$.

- *Monotonicity* prescribes that extra support for a proposition $\phi \in \Phi$ can never be damaging. Formally, $\phi \in J_i' \setminus J_i$ entails that $\phi \in F(J_i, \boldsymbol{J}_{-i}) \Rightarrow \phi \in F(J_i', \boldsymbol{J}_{-i})$, for all $(J_i, \boldsymbol{J}_{-i}) \in \mathcal{J}(\Phi)^n$ and $J_i' \in \mathcal{J}(\Phi)$.

- A more controversial property is *independence*, according to which each proposition ϕ in Φ is treated separately by the aggregation rule F. Formally, for all profiles $\boldsymbol{J}, \boldsymbol{J}'$, if $N_\phi^{\boldsymbol{J}} = N_\phi^{\boldsymbol{J}'}$, then $\phi \in F(\boldsymbol{J}) \Leftrightarrow \phi \in F(\boldsymbol{J}')$. It is easy to see that the Kemeny, the average-voter, and the plurality rule are not independent.

- Rule F is said to be *complete* (similarly *consistent*) if $F(\boldsymbol{J})$ is complete (consistent) for every $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$.

A straightforward characterization can be provided for the quota rules, based on a number of the above axioms.

Theorem 2.1 (Dietrich and List, 2007b)**.** *The quota rules are exactly those aggregation rules that satisfy anonymity, neutrality, independence and monotonicity simultaneously.*

However, even without asking for monotonicity, if we require the above properties together with completeness and consistency, the news is negative. Following the

established tradition in Social Choice Theory after Arrow's famous impossibility theorem (Arrow, 1951b) in Preference Aggregation, List and Pettit (2002, 2004) prove that if an agenda Φ contains at least two literals and their conjunction, then there is no complete and consistent judgment aggregation rule on Φ that is also anonymous, neutral, and independent. In the literature of Judgment Aggregation this result was extended by several authors, such as Dietrich (2007); Dietrich and List (2007a); Pauly and Van Hees (2006); and Van Hees (2007). Some ways out of the impossibility have also been attempted, with the most prominent being *domain restrictions* (e.g., Dietrich and List, 2010; List, 2003), that is, limiting our attention to profiles of judgments with specific (convenient) structure.

2.5 Binary Aggregation with Integrity Constraints

An alternative setting of Judgment Aggregation has been explored by Grandi and Endriss (2011) and Grandi (2012). This framework will be useful in certain parts of this thesis, which will be specified in the text (but if not mentioned otherwise, we will stick to Formula-based Judgment Aggregation). In the model of *Binary Aggregation with Integrity Constraints*, instead of an agenda the agents express their opinions on a finite set of *issues* $\mathcal{I} = \{1, \ldots, m\}$. The *ballot* B_i of an agent i, analogously to a judgment set, indicates the issues that agent i accepts. Formally, B_i is a vector in $\{0,1\}^m$, with 0 and 1 denoting acceptance and rejection of the corresponding issue, respectively. We call $\mathcal{D} = \{0,1\}^m$ the *domain*. Collecting all the individual ballots of the group $\mathcal{N}$, we have a profile $\boldsymbol{B} = (B_1, \ldots, B_n)$, which is the input of the aggregation procedure. As expected, the collective decision will be a ballot in $\mathcal{D}$.

Furthermore, depending on the application we have in mind, some elements of the domain may not be suitable to form opinions. Hence, in the terminology of Grandi and Endriss (2011), it is necessary to specify which of them are *rational*. Given the set of m issues $\mathcal{I}$, a set of propositional variables, one for each issue, is created: $PS = \{p_1, \ldots, p_m\}$. We do so using the language of Propositional Logic. Let $\mathcal{L}_{PS}$ be the propositional language constructed by closing PS under boolean connectives. For any formula $\phi \in \mathcal{L}_{PS}$, $Mod(\phi)$ designates the set of propositional assignments that satisfy ϕ. Then, we call an *integrity constraint* some formula $IC \in \mathcal{L}_{PS}$. Thanks to integrity constraints, we can capture logical interconnections between issues, and distinguish between rational and irrational ballots, which now correspond to assignments to the variables $p_1, \ldots, p_m$. So, fixing an integrity constraint IC, a rational ballot B is one that satisfies IC, i.e., an element of $Mod(IC)$. Similarly, a rational profile is an element of $Mod(IC)^n$.

It can be shown that there is a translation from Formula-based Judgment Aggregation to Binary Aggregation with Integrity Constraints and back. One direction is easy. Given a Judgment Aggregation framework defined by an agenda $\Phi = \{\phi_1, \ldots, \phi_m\}$, we can define a set of issues $\mathcal{I}_\Phi = \{i_{\phi_1}, \ldots, i_{\phi_m}\}$ that interprets it, creating an issue i_ϕ for every formula $\phi \in \Phi$. Then, the domain of aggregation is $D := \{0,1\}^{|\Phi|}$. More-

over, having a binary ballot $B \in D$, let us denote by B_j its j^{th} coordinate. In these terms, a ballot B will correspond to a judgment set J as follows: for every formula $\phi_r \in \Phi$, $\phi_r \in J$ if and only if $B_r = 1$ (call this correspondence $(\star)$). Also, an integrity constraint can be constructed, to dictate that only the ballots that correspond to complete and consistent judgment sets are rational. For the other direction, we mention a result due to Dokow and Holzman (2009, 2010) (see also Endriss et al., 2016a). It is shown that for every nonempty subset X of $\{0,1\}^m$, there exists an agenda $\Phi = \{\phi_1, ..., \phi_m\} \subseteq \mathcal{L}_{\{p_1,...,p_m\}}$ such that that $\mathcal{B}_{\mathcal{J}(\Phi)^n} = X$, where $\mathcal{B}_{\mathcal{J}(\Phi)^n}$ is derived by translating every profile of judgment sets in $\mathcal{J}(\Phi)^n$ into a profile of binary ballots, using correspondence $(\star)$.

2.6 Preference Relations

The notion of *preference* is principal in disciplines that study individual and collective decision making, including modern (micro)Economics, Rational Choice and Social Choice Theory. Von Neumann and Morgenstern (1944) initiated the treatment of preferences as formal mathematical relations whose properties can be stated axiomatically, and radically influenced other prominent economists who kept working in that direction afterwards, like Arrow (1951b).[8]

The central idea of this thesis is to interpret an aggregation problem as a strategic situation, where the agents, besides holding individual judgments, also prefer specific collective decisions more than others. In other words, we can think of an aggregation situation in terms of a game: every individual chooses an action, which is the (truthful or untruthful) judgment set she submits, and the outcome is computed by the submitted profile of the group's judgments, in accordance with the aggregation rule that is used. An agent's action, as in every standard game, is directly connected to the agent's preferences over the outcomes. To that end, we assume that every member i of a group $\mathcal{N}$ in an aggregation problem holds some preference relation $\succsim_i$ over all the possible collective judgment sets $J \in 2^{\Phi}$. By writing $J \succsim_i J'$, we mean that agent i wants the collective decision to be the judgment set J at least as much as she wants it to be judgment J'. Considering all judgment sets $J, J', J'' \in 2^{\Phi}$, we take the relation $\succsim_i$ to be *reflexive* ($J \succsim_i J$), *transitive* ($J \succsim_i J'$ and $J' \succsim_i J''$ implies $J \succsim_i J''$), and *complete* (either $J \succsim_i J'$ or $J' \succsim_i J$). Thus, we assume that individuals rank all pairs of possible outcomes relative to each other; no collective results are going to be incomparable.[9] Finally, we write $J \sim_i J'$ if $J \succsim_i J'$ and $J' \succsim_i J$, and we denote by

[8] For a recent work on individual preferences see Dietrich and List (2013).

[9] The requirement of completeness of preferences has triggered lot of discussion among philosophers and economists (e.g., Jeffrey, 1983), and one of the main arguments against it is directly reflected in the Judgment Aggregation framework. The possible collective outcomes will usually be exponentially as many as the issues in the agenda, and the agents have to be able to compare all of them. Nonetheless, one justification of the completeness constraint is based on our interpretation of the agents' preferences over the collective decisions. For example, we may think of preferences expressing "conceivable" acts

$J \succ_i J'$ the strict component of $J \succsim_i J'$, i.e., the case where $J \succsim_i J'$, but not $J \sim_i J'$.

The type of preferences that the agents hold will play a crucial role in our analysis about individual strategic behavior in aggregation problems. So, we will now reflect on some further assumptions that we can make about them. For example, in many aggregation contexts it is natural to suppose that the preferences of an agent depend on the truthful judgment set that this agent holds. Recall for instance the thesis-evaluation example in the Introduction. There, the members of the examination committee express their sincere opinions on whether the student should pass or fail the Master's program, and it would be reasonable to assume (supposing they are all well-intentioned) that they would like the final collective decision to match their own judgment. Hence, we will restrict our study to cases where individual judgments and preferences over collective outcomes are expected to be related. A full identification of scenaria that satisfy our assumptions is an empirical problem, which certainly deserves further investigation.

Consider an agenda Φ. Along the lines of Dietrich and List (2007c), we present three conditions that capture stronger and weaker assumptions with respect to the connection between an agent's judgment and her preferences over the collective decision. In all the following cases, the agents want the judgment of the group to "agree" to some extent with their own individual judgment. Each of the three conditions uniquely defines a class of preferences.

Top-respecting Preferences. For each agent i with truthful judgment set J_i, we define $T(J_i)$ as the set of all preference relations $\succsim_i \subseteq 2^\Phi \times 2^\Phi$ according to which J_i is ranked on top of all the other judgment sets. Formally, $T(J_i) := \{\succsim_i : J_i \succsim_i J, \text{ for all } J \in 2^\Phi\}$. We denote with T the family of all such preference relations with regard to individual judgment sets, i.e., $T := \{T(J_i) : J_i \in \mathcal{J}(\Phi)\}$.

Closeness-respecting Preferences. We say that a judgment set J is *at least as close to* J_i as another judgment set J' if for all formulas $\phi \in \Phi$, if J' agrees with J_i on ϕ, then J also agrees with J_i on ϕ. A preference relation $\succsim$ *respects closeness* to J_i if for any two judgment sets J and J', if J is at least as close to J_i as J', then $J \succsim J'$.

We can now define the class of *closeness-respecting preferences*. For each individual judgment set J_i, let $C(J_i)$ be the set of all preference relations $\succsim_i \subseteq 2^\Phi \times 2^\Phi$ that respect closeness to J_i. We denote with C the family of all such preference relations with respect to individual judgment sets, i.e., $C := \{C(J_i) : J_i \in \mathcal{J}(\Phi)\}$.

Then, $C \subset T$.

---

and not "actual" ones, in the sense that they represent the choice dispositions of the agents (Gilboa, 2009; Sen, 1973). From this perspective, completeness does not imply that the agents should be able to rank a large number of options prior to making a decision about them; instead, it may mean that they possess an intrapersonal method to rank the different judgment sets when these judgments are presented in pairs, which induces a complete ordering. An instance of a plausible such method is defined later in this section, and is constructed via the Hamming-distance.

Hamming-distance Preferences. One particular example of commonly used preferences in the literature that are closeness-respecting (for instance, Botan et al. (2016) base their analysis on those) are the *Hamming-distance preferences*. For every agent i, the Hamming-distance naturally induces a (reflexive, transitive and complete) preference relation $\succsim_i$ on judgment sets. According to it, agent i orders higher the judgments which agree on a greater number of propositions with her truthful opinion J_i:

$$J \succsim_i J' \Leftrightarrow H(J, J_i) \leqslant H(J', J_i)$$

We denote with $H(J_i)$ the unique preference relation $\succsim_i \subseteq 2^{\Phi} \times 2^{\Phi}$ that is defined as above, with respect to a fixed judgment set J_i. The family H contains all the Hamming distance preferences $H(J_i)$, induced by any $J_i \in \mathcal{J}(\Phi)$, i.e., $H := \{H(J_i) : J_i \in \mathcal{J}(\Phi)\}$. Obviously, $H \subset C \subset T$.

2.7 Strategic Manipulation

This section summarizes the main definitions and results in the literature of strategic behavior in Judgment Aggregation, which will constitute a cornerstone for the rest of this thesis. We focus on agents who may reason strategically and resort to untruthful behavior individually, in order to achieve a better outcome for themselves. Moreover, in accordance with the approaches of all the authors in the field to date, in this section we make two implicit assumptions: First, we assume that in every aggregation problem all the agents hold *full information* about the sincere opinions of their peers. Second, all the agents are taken to be *first-level reasoners*, i.e., they think strategically themselves, but they do not consider the possibility that the rest of the group may think strategically too. This research direction was initiated by Dietrich and List (2007c).

To start off, consider a function PR that assigns to each agent i and judgment set $J_i \in \mathcal{J}(\Phi)$ a non-empty set $PR(J_i)$ of reflexive, transitive and complete preference relations $\succsim_i$, which are considered "compatible" with J_i. Then, abusing the notation, we will also denote with PR the class of preferences constructed by that function, i.e., $PR := \{PR(J_i) : J_i \in \mathcal{J}(\Phi)\}$ (examples of such a class are the top-respecting, the closeness-respecting, and the Hamming-distance preferences defined above).[10] So, when does an agent with a truthful judgment J_i and a preference relation $\succsim_i$ have an *incentive* to submit a dishonest judgment in an aggregation problem? Definition 2.5 provides a formal answer. Note that we will use the words "*incentives*" and "*reasons*" interchangeably.

Definition 2.5. Consider a truthful profile of judgments $\boldsymbol{J} = (J_1, \ldots, J_n) \in \mathcal{J}(\Phi)^n$, an aggregation rule F and a class of preferences PR. Then, agent $i \in \mathcal{N}$ who holds

[10] The analysis of Dietrich and List (2007c) does not require completeness. However, to simplify the demonstration of the definitions and the proofs in the thesis we assume completeness, and we underline the fact that all the relevant results of Dietrich and List (2007c) hold for this restriction.

preferences $\succsim_i \in PR(J_i)$ has an *incentive to manipulate* on the profile $\boldsymbol{J}$ if there is at least one individual judgment set $J_i^* \in \mathcal{J}(\Phi)$, such that $F(J_i^*, \boldsymbol{J}_{-i}) \succ_i F(J_i, \boldsymbol{J}_{-i})$.

In words, the potential of obtaining a strictly more desirable result can provide a sufficient reason to an individual to lie. If some agent has an incentive to manipulate an aggregation rule in some aggregation situation, then we say that the aggregation rule is *manipulable*.

Definition 2.6. The aggregation rule F is *manipulable* for the class of preferences PR if there are a profile $\boldsymbol{J} \in \mathcal{J}(\Phi)^n$ and an individual $i \in \mathcal{N}$ holding preferences $\succsim_i \in PR(J_i)$, such that i has an incentive to manipulate on $\boldsymbol{J}$.

The following definition introduces formally the notion of *strategy-proofness* of an aggregation rule, i.e., the absence of incentives for manipulation on it, for all the agents, in all aggregation situations.

Definition 2.7. The aggregation rule F is *strategy-proof* for the class of preferences PR if for all individuals $i \in \mathcal{N}$, all profiles $\boldsymbol{J} = (J_1, \ldots, J_n) \in \mathcal{J}(\Phi)^n$, all preference relations $\succsim_i \in PR(J_i)$ and all individual judgment sets $J_i^* \in \mathcal{J}(\Phi)$, $F(J_i, \boldsymbol{J}_{-i}) \succsim_i F(J_i^*, \boldsymbol{J}_{-i})$.

Definition 2.7 implies that the agents have a noticeable *truth-bias*. That is, in case where they equally like all the possibly induced collective outcomes, they will choose to be honest. Obraztsova et al. (2013) justify this assumption by remarking that strategizing is costly for the agents, for example in time and cognitive effort, so they have a slight preference for truthfulness when they cannot unilaterally affect the outcome. Nevertheless, the gain obtained by being sincere can be assumed to be small enough, so that the agents will still try to manipulate if they can obtain a preferable result. The aggregation rule F is strategy-proof for the class of preferences PR if and only if F is not manipulable for PR.[11] We will often refer to strategy-proofness as *immunity to manipulation* and to manipulability as *susceptibility to manipulation*.

2.7.1 Characterization Results

Dietrich and List (2007c) were able to axiomatize the strategy-proof aggregation rules, considering groups of agents with reflexive and transitive, closeness-respecting preferences. Theorem 2.2 states their result, which establishes that the strategy-proof rules are exactly those that are both independent and monotonic. The idea for one direction can be grasped intuitively: if a rule is independent, it means that each proposition in the agenda is treated separately; together with monotonicity, an agent could never obtain a collective outcome that is closer to her truthful opinion if she tries to lie on some of the formulas. In order to prove that an aggregation rule is strategy-proof only

[11] Note that in order to obtain Definition 2.6 as a contrapositive of Definition 2.7, we make use of the fact that preference relations are taken to be complete.

if it is independent and monotonic, however, Dietrich and List follow a more indirect path, which is explained in the Appendix (see Theorem 2.3).

Theorem 2.2 (Dietrich and List, 2007c)**.** *An aggregation rule F is strategy-proof for all reflexive and transitive closeness-respecting preferences if and only if F is independent and monotonic.*

We further show in Theorem 2.3 that the characterization result of Dietrich and List remains valid for agents whose preferences are also complete. In a trivial manner, we have that any independent and monotonic aggregation rule is strategy-proof for every subset of the class of all reflexive and transitive, closeness-respecting preferences (and the class of all reflexive, transitive and complete, closeness-respecting preferences is obviously such a subset). Moreover, we prove in the Appendix that the non-straightforward direction also holds.

Theorem 2.3. *An aggregation rule F is strategy-proof for all reflexive, transitive and complete closeness-respecting preferences if and only if F is independent and monotonic.*

2.7.2 The Main Impossibility

The attempts to escape manipulability in Social Choice Theory and particularly in Voting Theory (or Preference Aggregation) have been haunted for decades by the famous impossibility theorem of Gibbard (1973) and Satterthwaite (1975). These scholars showed independently that any voting procedure with at least three candidates, where each one of them could be the winner in some voting scenario, is immune to strategic manipulation if and only if it is a dictatorship of some individual.

Dietrich and List (2007c) present a Gibbard-Sattethrwaite-style impossibility result conveyed in the framework of Judgment Aggregation. According to Dietrich and List's theorem, any aggregation rule functioning on an agenda obtained from the large class of path-connected agendas, satisfies a number of reasonable axiomatic properties together with strategy-proofness if and only if it is dictatorial.

Theorem 2.4 (Dietrich and List, 2007c)**.** *For a path-connected agenda Φ, an aggregation rule F is complete, consistent, responsive and strategy-proof for the class of all closeness-respecting preferences if and only if F is a dictatorship of some individual.*

To conclude, we have now introduced all the necessary terminology and we will wend our way through the main body of this thesis, in Chapters 3, 4 and 5.

Chapter 3

Strategic Manipulation under Partial Information

The central goal of this chapter is to provide a richer model of Judgment Aggregation that accounts for partial information, and explore its implications with regard to known aggregation rules and results of the literature. In this framework each agent knows the formation of the group that she belongs to (i.e., the members of the group) and the aggregation rule that is applied. Each agent is also aware of her own truthful judgment set, but she may hold incomplete information about the judgment sets of the rest of the group. For example, a member of a committee may know the opinions of all the other members on one specific issue on the agenda, in case this issue has been already discussed, but be ignorant about their judgments on the rest of the issues. In a different situation, an agent may be fully informed only about the judgments of a subset of the group members, of those people that she happens to be friends with, for instance.

We assume that each agent receives some information about her peers, which may fully or partially describe their truthful judgments. Each agent knows that the information she holds is accurate, in the sense that it only provides valid data about the truthful judgments of the group, and believes that the other agents are going to report their truthful judgments when they are asked to do so (we call agents with that belief *level-1 reasoners*). The latter assumption will be dropped in Chapter 4. Moreover, it is common knowledge that the aggregation procedure takes place in a single round (a refinement will be studied in Chapter 5). Our question is: Under which conditions will an agent be better off by reporting an untruthful judgment, and what kind of (partial) information may protect an aggregation rule from manipulation?

The remainder of this chapter is structured as follows. Section 3.1 introduces *judgment information functions*, which model the information of the agents about the truthful judgments of the rest of the group. Each judgment information function induces an amount of *uncertainty*, which we are able to measure formally. Section 3.2 provides the definitions needed in order to determine when an agent has an incentive to lie under partial information, and Section 3.3 introduces an agent's *best strategies*.

Subsequently, we study the logical connections between full and partial information (Section 3.4), by addressing the following topics. First, is an aggregation rule that is immune to manipulation under full information guaranteed to be immune to manipulation under partial information? Our reply here is positive. On the other direction, if an agent has a reason to manipulate an aggregation procedure when she is fully informed about the truthful judgments of the group, does she still have a reason to manipulate when there is information that she is missing? The answer to this question is slightly more complex; both positive and negative, depending on the aggregation rule that is used in combination with the nature of the information that the agent possesses. We continue in Section 3.5 by introducing the notion of *relevant information* in an aggregation problem. Furthermore, we take an axiomatic perspective and we show for example that if an aggregation rule is independent and monotonic, then it is immune to manipulation under any kind of partial information. Afterwards, we focus on the premise-based procedure (Section 3.6). Inspired by a result of Dietrich and List (2007c) according to which the premise-based procedure is manipulable under full information when the agents only care about the result on the conclusion (and not the premises) of a conjunctive or a disjunctive agenda, we prove that manipulability is also the case under any kind of uncertainty. However, we can obtain results that support immunity to manipulation when we consider agents with preferences that take into account, up to some degree, both the premises and the conclusion. Finally, in Section 3.7 we show how accounting for the partial information that the agents may hold can be crucial in order to circumvent the impossibility result of Dietrich and List (2007c), explained in the Background. This is the main contribution of this chapter. We conclude in Section 3.8.

3.1 Information Functions

The (partial) information that the agents hold in an aggregation problem may be of different types. For example, it may suffice to identify a number of profiles of judgments that are possible to be held by the group, by clarifying how many agents have a specific opinion, but not who holds which judgment, etc. The analysis of this section is largely inspired by previous work in Voting Theory by Reijngoud and Endriss (2012).[12] We call $\mathcal{I}$ the set of all different data about the judgments of the rest of the group that an agent can be informed about before the final reporting of judgments. We formally define a *judgment information function* (JIF) $\pi : \mathcal{N} \times \mathcal{J}(\Phi)^n \to \mathcal{I}$, which maps agents and profiles of judgments to elements of $\mathcal{I}$. Intuitively, a JIF represents the available information for every agent, given the truthful profile of judgments of the group. To ease the notation, we will write $\pi_i(\boldsymbol{J})$ for the information of the agent i on the profile $\boldsymbol{J} := (J_1, \ldots, J_n)$. The following are some possible choices for elements

[12] Recently, a different approach on partial information in Judgment Aggregation has been undertaken by Griffioen (2017), where the agents are associated with cardinal utility functions.

of $\mathcal{I}$, together with their corresponding JIF π.

- *Profile*: The profile-JIF returns the full input profile for every agent:
  $\pi_i(\boldsymbol{J}) = \boldsymbol{J}$, for all $i \in \mathcal{N}$.

- *Anonymous profile* (a-profile): The a-profile-JIF returns the number of agents that accept each formula in the input profile:
  $\pi_i(\boldsymbol{J}) = (|N_\phi^{\boldsymbol{J}}|)_{\phi \in \Phi}$, for all $i \in \mathcal{N}$.

- *Judgment sets' number* (js-number): The js-number-JIF returns the number of agents that submit each judgment set in the input profile:
  $\pi_i(\boldsymbol{J}) = (|\{i \in \mathcal{N} : J_i = J\}|)_{J \in \mathcal{J}(\Phi)}$, for all $i \in \mathcal{N}$.

- *Winner*: The winner-JIF returns the judgment set that is submitted by the largest number of agents in the input profile (if there are more than one such judgment sets, ties are broken by a fixed lexicographic-tie breaking):[13]
  $\pi_i(\boldsymbol{J}) = J \in \mathcal{J}(\Phi) : J \in \underset{J' \in \mathcal{J}(\Phi)}{\arg\max} |\{i \in \mathcal{N} : J_i = J'\}|$, for all $i \in \mathcal{N}$.

- *All_but_ϕ_i*: The all_but_ϕ_i-JIF returns for each agent i the judgments of the rest of the group on each formula apart from the formula $\phi_i \in \Phi$:
  $\pi_i(\boldsymbol{J}) = (N_\phi^{\boldsymbol{J}})_{\phi \in \Phi \setminus \{\phi_i, \neg\phi_i\}}$, for all $i \in \mathcal{N}$.

- *Zero*: The zero-JIF does not return any information; it just gives us a constant value:
  $\pi_i(\boldsymbol{J}) = 0$, for all $i \in \mathcal{N}$.

Our framework allows for the above JIFs to be combined, in the sense that different agents may have access to different types of information. Now, having the information expressed by a JIF π and a truthful profile of judgments $\boldsymbol{J}$, we define the set of (partial) profiles that a level-1 reasoner i considers possible:

$$\mathcal{W}_i^{1,\pi,\boldsymbol{J}} := \{\boldsymbol{J}'_{-i} : \pi_i(J_i, \boldsymbol{J}'_{-i}) = \pi_i(\boldsymbol{J})\}$$

That is, $\mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ contains all the judgments of the rest of the group that are compatible with agent i's information and level-1 reasoning. Note that in the present work we only deal with qualitative beliefs. We assume that the agents cannot or do not want to assign any numerical value (probability) to their beliefs about the possibility of the occurrence of each scenario concerning the judgments of the group.[14] Hence, in order to account for the most general case that only distinguishes between "possible" and "not

[13] The winner-JIF actually conveys the plurality winner.

[14] Due to the extensive discussion in Decision Theory about imprecise beliefs and ambiguity (e.g., Arrow, 1951a; Ellsberg, 1961; Knight, 1921), we hereby wish to avoid making any restricting assumptions concerning the uncertainty domain of the agents.

possible" scenaria, a suitable tool is the notion of *qualitative uncertainty* (Halpern, 2005). As Reijngoud and Endriss (2012) (see also Chopra et al., 2004) observe, $\mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ satisfies the three axioms of reflexivity (REF), symmetry (SYM) and transitivity (TRANS). For all judgment sets J_i and for all (partial) profiles $\boldsymbol{J}_{-i}, \boldsymbol{J}^*_{-i}, \boldsymbol{J}^{**}_{-i}$:

$$\text{(REF)} \quad \boldsymbol{J}_{-i} \in \mathcal{W}_i^{1.\pi,(J_i,\boldsymbol{J}_{-i})}$$

$$\text{(SYM)} \quad \text{if } \boldsymbol{J}_{-i} \in \mathcal{W}_i^{1,\pi,(J_i,\boldsymbol{J}^*_{-i})}, \text{ then } \boldsymbol{J}^*_{-i} \in \mathcal{W}_i^{1,\pi,(J_i,\boldsymbol{J}_{-i})}$$

$$\text{(TRANS)} \quad \text{if } \boldsymbol{J}_{-i} \in \mathcal{W}_i^{1,\pi,(J_i,\boldsymbol{J}^*_{-i})} \text{ and } \boldsymbol{J}^*_{-i} \in \mathcal{W}_i^{1,\pi,(J_i,\boldsymbol{J}^{**}_{-i})}, \text{ then } \boldsymbol{J}_{-i} \in \mathcal{W}_i^{1,\pi,(J_i,\boldsymbol{J}^{**}_{-i})}$$

Axiom (REF) expresses that every agent always considers possible the truthful profile of judgments of the rest of the group. Axioms (SYM) and (TRANS) together state that whenever an agent considers some profile possible, then that profile would also induce the same information set as her current one.

A JIF π represents the amount of information that the agents possess, hence it is useful to define a formal measure of the *uncertainty* that this information induces.

Definition 3.1. Consider a JIF π.

- $U_i^{\boldsymbol{J}}(\pi) := \dfrac{|\mathcal{W}_i^{1,\pi,\boldsymbol{J}}| - 1}{|\mathcal{J}(\Phi)^{n-1}| - 1}$
  is the uncertainty of the agent i on the truthful profile $\boldsymbol{J}$ that the JIF π induces.[15]
- $U_i(\pi) := \max\limits_{\boldsymbol{J} \in \mathcal{J}(\Phi)^n} U_i^{\boldsymbol{J}}(\pi)$
  is the uncertainty of the agent i that the JIF π induces.
- $U(\pi) := \max\limits_{i \in \mathcal{N}} U_i(\pi)$
  is the uncertainty that the JIF π induces.

The uncertainty that the JIF π induces for an agent on a profile is a real number from 0 to 1, i.e., $0 \leqslant U_i^{\boldsymbol{J}}(\pi) \leqslant 1$, where 0 denotes full certainty and 1 total uncertainty. The more partial profiles are possible for the agent, the more her uncertainty increases. For example, according to the profile-JIF the agent only considers possible the truthful partial profile, thus the uncertainty of the profile-JIF is 0. At the other extreme, the uncertainty of the zero-JIF is 1, because according to it the agent considers possible all the partial profiles. Overall, the uncertainty of a JIF π is the maximum of the uncertainty it induces for all the agents and for all the potential truthful profiles.

[15] $U_i^{\boldsymbol{J}}(\pi)$ is defined for all agendas Φ such that $|\mathcal{J}(\Phi)| > 1$; otherwise it is set to 0.

3.2 Strategy-proofness under Partial Information

Definitions 2.6 and 2.7, given in the Background, constitute the standard approach with regard to strategy-proofness in the literature of Judgment Aggregation to date. However, these definitions implicitly assume that every agent is fully informed about the truthful judgments of the group. We now refine them, accounting for agents with incomplete information.

Definition 3.2. Consider a truthful profile $\boldsymbol{J} = (J_1, \ldots, J_n) \in \mathcal{J}(\Phi)^n$. For an aggregation rule F and a class of preferences PR, agent $i \in \mathcal{N}$ holding preferences $\succsim_i \in PR(J_i)$ has an *incentive to π-manipulate* on the profile $\boldsymbol{J}$ if there is at least one individual judgment set $J_i^* \in \mathcal{J}(\Phi)$, such that

1. $F(J_i^*, \boldsymbol{J'}_{-i}) \succ_i F(J_i, \boldsymbol{J'}_{-i})$, for some $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ and
2. $F(J_i^*, \boldsymbol{J''}_{-i}) \succsim_i F(J_i, \boldsymbol{J''}_{-i})$, for all other $\boldsymbol{J''}_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$

This means that an agent has an incentive to manipulate under the (partial) information provided by the JIF π by reporting an untruthful judgment if there is a scenario consistent with her information that will result in a more desirable collective decision for her and there is no scenario where she will be worse off than when reporting a truthful judgment. That is, we adopt the pessimistic perspective (from the agents' standpoint) according to which they are willing to lie only if they are totally safe to do so. Said differently, the agents are taken to be *risk-averse*: if there is at least one possible scenario where lying induces a less desirable result, then they remain truthful. In that sense, agents are *cautious* when it comes to manipulation.

If there is a profile $\boldsymbol{J}$ where at least one agent has an incentive to π-manipulate, then we say that the aggregation rule is π-manipulable.

Definition 3.3. The rule F is *π-manipulable* for the class of preferences PR if there is some profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i}) \in \mathcal{J}(\Phi)$ and at least one individual $i \in \mathcal{N}$ holding preferences $\succsim_i \in PR(J_i)$ such that i has an incentive to π-manipulate on $\boldsymbol{J}$.

The aggregation rule F is π-strategy-proof for the class of preferences PR if and only if F is not π-manipulable for PR.[16]

Definition 3.4. The aggregation rule F is *π-strategy-proof* for the class of preferences PR if for all individuals $i \in \mathcal{N}$, all profiles $\boldsymbol{J} = (J_1, \ldots, J_n) \in \mathcal{J}(\Phi)^n$, all preference relations $\succsim_i \in PR(J_i)$ and all individual judgment sets $J_i^* \in \mathcal{J}(\Phi)$,

1. $F(J_i, \boldsymbol{J'}_{-i}) \succsim_i F(J_i^*, \boldsymbol{J'}_{-i})$, for all $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ or
2. $F(J_i, \boldsymbol{J''}_{-i}) \succ_i F(J_i^*, \boldsymbol{J''}_{-i})$, for some $\boldsymbol{J''}_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$

[16] In order to obtain Definition 3.4 as a contrapositive of Definition 3.3, we make use of the fact that according to our assumptions preference relations are complete.

Definition 3.4 brings out a further assumption concerning the *truth-bias* of the agents. Justifying it as in the case of full information, whenever an agent cannot unilaterally change the outcome, she chooses to be honest, and moreover the same holds if being sincere induces a preferable collective decision in some scenario (no matter whether lying does the same in a different scenario). For what follows we will concentrate our analysis on closeness-respecting preferences. If it is clear from the context, we may write that an aggregation rule F is simply (π-)strategy-proof, omitting the class of preferences we refer to.

Obviously, when π is the profile-JIF, π-strategy-proofness (manipulation) is equivalent to strategy-proofness (manipulation) under full information. We can further understand the importance of partial information on strategy-proofness as follows. Consider an agent i. If this agent possesses full information about the judgments of the rest of the group, then she will manipulate with no second thought in case she finds an untruthful judgment that makes her better off. However, finding such an insincere judgment is not sufficient to make agent i manipulate under partial information. Then, an extra condition needs to be satisfied: for all possible scenaria, the untruthful judgment should induce a result at least as good as the one induced by the agent's truthful judgment. Loosely speaking, this second condition provides an additional layer of safety against manipulation for an aggregation rule.

3.3 Best Strategies

An alternative way to specify the incentives of an agent to manipulate an aggregation process is by looking at her *best strategies* under the information she holds.

Consider an aggregation rule F, a truthful profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$, and an agent $i \in \mathcal{N}$ with preferences $\succsim_i$, who considers possible the set of (partial) profiles $\mathcal{W} \subseteq \mathcal{J}(\Phi)^{n-1}$ to be truthfully held by the group. First, we say that a judgment set $J \in \mathcal{J}(\Phi)$ is *undominated* in the standard game-theoretical sense, if there is no other judgment set J' such that (1) $F(J', \boldsymbol{J'}_{-i}) \succ_i F(J, \boldsymbol{J'}_{-i})$, for some $\boldsymbol{J'}_{-i} \in \mathcal{W}$ and (2) $F(J', \boldsymbol{J''}_{-i}) \succsim_i F(J, \boldsymbol{J''}_{-i})$, for all other $\boldsymbol{J'}_{-i} \neq \boldsymbol{J''}_{-i} \in \mathcal{W}$. Then, recalling that according to our assumptions the agents are truth-biased, we will say that if agent i's truthful judgment J_i is undominated, then it will be her unique best strategy. Otherwise, all the undominated judgment sets can be used by agent i as her best strategies.[17]

Definition 3.5. We define the set $S := S_i^F(\mathcal{W}, \succsim_i, J_i)$ of agent i's *best strategies*.

- If agent i's truthful opinion J_i is undominated, then $S := \{J_i\}$.
- Otherwise, $S := \{J \in \mathcal{J}(\Phi) : J \text{ is undominated}\}$.

[17] In such situation, the agent could practically break her tie in various ways that are not of our concern. For example, she could try to assign different weights (probabilistic or not) to the scenaria she considers possible, or she could use some randomness-providing tool, etc.

Lemma 3.1 is immediate.

Lemma 3.1. *It always holds that*

(a) $S_i^F(\mathcal{W}, \succsim_i, J_i) \neq \varnothing$*;*

(b) $J_i \in S_i^F(\mathcal{W}, \succsim_i, J_i)$ *if and only if* $S_i^F(\mathcal{W}, \succsim_i, J_i) = \{J_i\}$.

Hence, an agent has an incentive to manipulate if some insincere opinion is a best strategy of hers, while there is no incentive for manipulation if the only best strategy of the agent is telling the truth.

Definition 3.6. Consider an aggregation rule F, an agent i holding preferences $\succsim_i$, a truthful profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$, and a JIF π. Agent i has an *incentive to π-manipulate* on $\boldsymbol{J}$ if and only if an untruthful judgment set of hers J_i^* belongs to her best strategies when she considers possible the set of partial profiles $\mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ induced by π. That is, if and only if $J_i^* \in S_i^F(\mathcal{W}_i^{1,\pi,\boldsymbol{J}}, \succsim_i, J_i)$, if and only if $J_i \notin S_i^F(\mathcal{W}_i^{1,\pi,\boldsymbol{J}}, \succsim_i, J_i)$.

3.4 The Interplay between Full and Partial Information

A natural next question is the following: Is there a logical relation between full and partial information with regard to strategy-proofness? We show that as one could expect, any aggregation rule that is strategy-proof under full information is strategy-proof under partial information too, or equivalently, any rule that is manipulable under partial information is also manipulable under full information (Theorem 3.4).

Consider an agenda Φ. Theorem 3.2 builds a basis for a number of results that follow. It generalizes the proof of Dietrich and List (2007c) (recall Theorem 2.2 in the Background), connecting strategy-proofness under partial information with two well-known axioms of judgment aggregation rules: *independence* and *monotonicity*.

Theorem 3.2. *For any JIF π, if an aggregation rule F is independent and monotonic, then F is π-strategy-proof for the class C of all closeness-respecting preferences.*

Proof. Assume, aiming for a contradiction, that the statement of the theorem is not true. Then, there is an independent and monotonic aggregation rule F and a JIF π such that F is not π-strategy-proof for the class C of all closeness-respecting preferences. That is, there are an agent i, a profile $(J_i, \boldsymbol{J}_{-i})$, a closeness-respecting preference $\succsim_i \in C(J_i)$ and a judgment set J_i^* such that $F(J_i^*, \boldsymbol{J'}_{-i}) \succ_i F(J_i, \boldsymbol{J'}_{-i})$, for some $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ (and $F(J_i^*, \boldsymbol{J''}_{-i}) \succsim_i F(J_i, \boldsymbol{J''}_{-i})$, for all other $\boldsymbol{J''}_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$). By the definition of closeness-respecting preferences, we have that there is some $\phi \in J_i$ such that $\phi \in F(J_i^*, \boldsymbol{J'}_{-i})$ and $\phi \notin F(J_i, \boldsymbol{J'}_{-i})$. By independence and monotonicity, this is possible only if $\phi \in J_i^*$ too. But then, by independence, it should be $\phi \in F(J_i^*, \boldsymbol{J'}_{-i})$ if and only if $\phi \in F(J_i, \boldsymbol{J'}_{-i})$, which is a contradiction. □

Interestingly, the JIF π did not have a salient role in the proof of Theorem 3.9. This observation will let us generalize even further the above result in later sections, showing that independent and monotonic rules are resistant to manipulation when we consider richer models of strategic behavior too. Now, since quota rules are independent and monotonic (recall Theorem 2.1 in the Background), Corollary 3.3 holds.

Corollary 3.3. *Quota rules are immune to π-manipulation, for every JIF π.*

Our main claim of this section, namely that strategy-proofness under full information guarantees strategy-proofness under partial information, is proven in Theorem 3.4. (Note that, when it is clear from the context, we simply write "strategy-proof", meaning "strategy-proof under full information", and similarly for "manipulable".)

Theorem 3.4. *For the class C of all closeness-respecting preferences, all aggregation rules that are strategy-proof are also π-strategy-proof, for every JIF π.*

Proof. Dietrich and List (2007c) show that the strategy-proof rules for the class of preferences C are independent and monotonic (see the Background). Also, all independent and monotonic rules are π-strategy-proof, for every JIF π (Theorem 3.2). □

Corollary 3.5. *For the class C of all closeness-respecting preferences and for every JIF π, all aggregation rules F that are π-manipulable are also manipulable.*

Theorem 3.4 can be understood as a special case of a more general fact. Let us call a JIF π *at least as informative as* another JIF σ if for all profiles $\boldsymbol{J}$ and all agents i, $\mathcal{W}_i^{1,\pi,\boldsymbol{J}} \subseteq \mathcal{W}_i^{1,\sigma,\boldsymbol{J}}$. As anticipated, if a JIF π is at least as informative as another JIF σ, then the uncertainty induced by σ is at least as high as the uncertainty induced by π. Lemma 3.6 translates in our framework an analogous result from Voting, worked out by Reijngoud and Endriss (2012), whose proof can be found in the Appendix.

Lemma 3.6. *If a JIF π is at least as informative as another JIF σ, then all aggregation rules that are σ-manipulable for a class of preferences PR are also π-manipulable for PR.*

Corollary 3.7. *If a JIF π is at least as informative as another JIF σ, then all aggregation rules that are π-strategy-proof for the class of preferences PR are also σ-strategy-proof for PR.*

Obviously, the profile-JIF is at least as informative as every other JIF π. Hence, Theorem 3.4 follows by Corollary 3.7.

As Theorem 3.8 shows next, the notions of strategy-proofness and π-strategy-proofness are not equivalent. This means that there are cases where an aggregation rule is not strategy-proof, but still, it is π-strategy-proof for some JIF π.

Theorem 3.8. *Consider an agenda Φ. There are an aggregation rule F and a JIF π such that F is π-strategy-proof for the class of closeness-respecting preferences C, but F is not strategy-proof for C under full information.*

Proof. Consider an arbitrary formula $\psi \in \Phi$ and the rule F such that for all profiles $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ and formulas $\phi \in \Phi$, it is $\phi \in F(\boldsymbol{J})$ if $\phi \neq \psi$ and $\psi \in F(\boldsymbol{J})$ if and only if $|N_\psi^{\boldsymbol{J}}|$ is odd. Moreover, take the JIF π which denotes that each agent i is completely ignorant of the judgment of the agent $i-1$, but knows the judgments of everyone else in the group. Formally, the JIF π is such that $\pi_i(\boldsymbol{J}) = \{(J, \boldsymbol{J}_{-(i-1)}) : J \in \mathcal{J}(\Phi)\}$, where agent 0 is taken to be agent n. The aggregation rule F is not monotonic, hence we know by Dietrich and List (2007c) that it is not strategy-proof. However, it is easy to see that F is π-strategy-proof. Take an arbitrary agent i, a profile $(J_i, \boldsymbol{J}_{-i})$, a judgment set J_i^* and a preference relation $\succsim_i \in C(J_i)$, and suppose that there is a judgment set J_i^* such that $F(J_i^*, \boldsymbol{J'}_{-i}) \succ_i F(J_i, \boldsymbol{J'}_{-i})$, for some $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$. By the definition of closeness-respecting preferences, we have that there is some $\phi \in J_i$ such that $\phi \in F(J_i^*, \boldsymbol{J'}_{-i})$ and $\phi \notin F(J_i, \boldsymbol{J'}_{-i})$. This can only happen if $\phi = \psi$ and $\psi \notin J_i^*$ (which means that $\psi \neq \bot, \top$). Now, consider the following two cases (where J'_{i-1} is the coordinate that corresponds to agent $i-1$ in the partial profile $\boldsymbol{J'}_{-i}$).

Case 1: $\psi \in J'_{i-1}$. Then, $\psi \neq \top$ implies that there is some model M of the Logic such that $M \not\models \psi$. We define a new (complete and consistent) judgment set of agent $i-1$ based on the formulas that are verified by M, $J''_{i-1} := \{\phi \in \Phi : M \models \phi\}$. So, we now have that $\psi \notin J''_{i-1}$. Starting from the partial profile $\boldsymbol{J'}_{-i}$, preserving the judgments of the rest of the group and replacing the judgment of agent $i-1$, J'_{i-1}, with J''_{i-1}, we construct the new partial profile $\boldsymbol{J''}_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$. So, by definition of rule F, it holds that $F(J_i, \boldsymbol{J''}_{-i}) = F(J_i^*, \boldsymbol{J'}_{-i})$ and $F(J_i^*, \boldsymbol{J''}_{-i}) = F(J_i, \boldsymbol{J'}_{-i})$, which means that $F(J_i, \boldsymbol{J''}_{-i}) \succ_i F(J_i^*, \boldsymbol{J'}_{-i})$. To sum up, there is always a judgment set of the agent $i-1$ that agent i considers possible and that makes agent i prefer reporting her truthful judgment to reporting any untruthful judgment. We conclude that rule F is safe from π-manipulation of the agent i.

Case 2: $\psi \notin J'_{i-1}$. Then since $\psi \neq \bot$, we know that there is some model of the Logic M such that $M \models \psi$. The rest of the proof proceeds analogously to case 1. □

The proof of Theorem 3.8 brings to light a broader observation. Having a fixed aggregation rule F, suppose that an agent i holds a truthful judgment set J_i in the profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$, and there is a dishonest opinion J_i^* which results in a more desirable outcome for her. Now, further suppose that if the judgments of the rest of the group were different, it would be possible that the results by reporting J_i and J_i^* were reversed. Formally, this last assumption means that there is at least one partial profile $\boldsymbol{J'}_{-i}$ such that $F(J_i, \boldsymbol{J}_{-i}) = F(J_i^*, \boldsymbol{J'}_{-i})$ and $F(J_i^*, \boldsymbol{J}_{-i}) = F(J_i, \boldsymbol{J'}_{-i})$. If agent i deems possible one such partial profile $\boldsymbol{J'}_{-i}$ according to a JIF π, then she does not have anymore an incentive to manipulate using J_i^*. In other words, partial information of this sort can make an agent more cautious when it comes to lying. If the above holds for every profile and every agent, then partial information can guarantee immunity to manipulation. Overall, we have described a sufficient condition for

π-strategy-proofness (note that the type of the preferences is not crucial here).

3.5 Relevant Information

As we discussed in the Background, axioms play a principal role in Judgment Aggregation. First, in a direct manner, axioms allow us to determine philosophically appealing aggregation rules by identifying a set of precise mathematical properties that these rules satisfy. Second, indirectly via characterization results, axioms facilitate the categorization of aggregation rules to various families, such as the strategy-proof and the manipulable ones. Recall that Dietrich and List (2007c) showed that all aggregation rules that are immune to manipulation under full information are exactly the independent and monotonic ones (Theorem 3.9 in the Background). Can we find an analogous characterization for π-strategy-proof rules?

When an agent is missing part of the information about the truthful judgments of her peers, she has to think about multiple possible scenaria that can affect her behavior. In this case, not only the quantity but also the quality of her information can be proven essential. In general, having an aggregation rule F and a profile of judgments $\boldsymbol{J}$, there is some part of the information concerning $\boldsymbol{J}$ that is relevant for an agent's manipulation, while some other part may be of no use. For example, when the plurality rule is applied, knowing the number of agents who truthfully hold each judgment matters, but being informed about who exactly holds each judgment set is redundant. When the rule is not anonymous, however, who holds which judgment is critical for the outcome. The following definition provides the notion of a *respectful* JIF π, meaning that π contains all the relevant information for all the agents in an aggregation problem. Specifically, this happens when all the information that each agent i holds in π agrees on what her preferred behavior is (i.e., lying or not) looking at the possible results; then we can say that all relevant information is available for agent i to decide.

Definition 3.7. A JIF π *respects* a class of preferences PR with regard to F if for all agents $i \in \mathcal{N}$, all profiles $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i}) \in \mathcal{J}(\Phi)^n$, all judgment sets $J_i^* \in \mathcal{J}(\Phi)$ and all preference relations $\succsim_i \in PR(J_i)$, there is no pair of (partial) profiles $\boldsymbol{J}^1_{-i}, \boldsymbol{J}^2_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ such that $F(J_i, \boldsymbol{J}^1_{-i}) \succ_i F(J_i^*, \boldsymbol{J}^1_{-i})$ and $F(J_i^*, \boldsymbol{J}^2_{-i}) \succ_i F(J_i, \boldsymbol{J}^2_{-i})$.

Example 3.1. The profile-JIF respects any class of preferences PR with regard to any aggregation rule F. (Immediate) $\triangle$

Example 3.2. The a-profile-JIF respects any class of preferences PR with regard to all anonymous and independent (e.g., quota) rules. (Immediate) $\triangle$

Example 3.3. The js-number-JIF respects any class of preferences PR with regard to the plurality rule, paired with a lexicographic tie-breaking rule. (Immediate) $\triangle$

Example 3.4. The js-number-JIF respects any class of preferences PR with regard to the Kemeny rule, paired with a lexicographic tie-breaking rule. (Immediate) $\triangle$

Example 3.5. The a-profile-JIF does not respect the class of all closeness-respecting preferences C with regard to the plurality rule, paired with a lexicographic tie-breaking rule.

Intuitively, knowing how many agents agree with each formula in the agenda does not suffice for an agent to predict the result of the plurality rule F^{pl}, because she may still be ignorant of which exactly the truthful judgments of the group are. To illustrate, let Φ be an agenda consisting of three variables p_1, p_2, p_3 and their negations, and let 100 designate the opinion that accepts p_1 and rejects p_2 and p_3, etc. Suppose that the lexicographic tie-breaking order (of the judgment sets that will be of our interest) is as follows: $011 > 100 > 000 > 010 > 001$. Moreover, suppose that a group $\mathcal{N}$ consists of three agents 1, 2, and 3, and agent 1 strictly prefers the collective decision to be the same as her truthful opinion 100. Agent 1 also strictly prefers all the judgment sets to 011, which is totally opposite to her sincere judgment, and is indifferent about the other outcomes. Formally, agent 1 holds the closeness-respecting preference $\gtrsim_1$ such that $100 \succ_1 000 \sim_1 010 \sim_1 001 \succ_1 011$. Consider an aggregation problem where agent 1 is informed that each one of the propositions p_1, p_2 and p_3 are accepted by exactly one of the agents in the group. Since she herself accepts proposition p_1, she knows that her peers reject it. But she is completely ignorant about the rest. What if agent 2 truthfully accepts both propositions p_2 and p_3 and agent 3 rejects them? This scenario is depicted in Table 2. Then, due to the tie-breaking, the aggregated result will be 011, which is the least preferred opinion of agent 1. Hence, agent 1 would rather report insincerely that she rejects proposition p_1, submitting the untruthful judgment 000, and turn it into the collective decision, which is more desirable for her. On the other hand, what if agent 2 only accepts proposition p_2 and agent 3 only accepts proposition p_3? This scenario is depicted in Table 3. Then, agent 1's truthful opinion would win, and if she tried to lie in the aforementioned way, she would be worse off. At the end of the day, different possible scenaria suggest different strategies for agent 1, thus we can say that she does not hold all relevant information. △

	p_1	p_2	p_3
Agent 1:	Yes	No	No
Agent 2:	No	Yes	Yes
Agent 3:	No	No	No
F^{pl}	No	Yes	Yes

Table 2: If agent 1 is sincere, the result is opposite to her judgment

	p_1	p_2	p_3
Agent 1:	Yes	No	No
Agent 2:	No	Yes	No
Agent 3:	No	No	Yes
F^{pl}	Yes	No	No

Table 3: If agent 1 is sincere, the result is exactly her judgment

Example 3.6. The a-profile-JIF respects any class of preferences PR with regard to the Kemeny rule, paired with a lexicographic tie-breaking order. (Immediate) △

Example 3.7. Consider the conjunctive agenda $\Phi := \{a_1, a_2, \neg a_1, \neg a_2, c \leftrightarrow (a_1 \wedge a_2), \neg c\}$. Moreover, consider F^{pr} the premise-based procedure, and C the class of all closeness-respecting preferences. The zero-JIF does not respect C with regard to F^{pr}.

Take a group of three agents $\mathcal{N} = \{1, 2, 3\}$. We will show that there is an aggregation problem where an agent does not have all the relevant information to decide if she prefers to be truthful or to report an untruthful judgment, and hence, the zero-JIF is not respectful. Take agent $i := 3$, a judgment set $J_3 := \{\neg a_1, a_2, \neg c\}$ and a preference relation $\succsim_3 \in C(J_3)$ such that: $F^{pr}(\boldsymbol{J}) \succ_3 F^{pr}(\boldsymbol{J'})$ if $c \notin F^{pr}(\boldsymbol{J})$ and $c \in F^{pr}(\boldsymbol{J'})$; and $F^{pr}(\boldsymbol{J}) \sim_3 F^{pr}(\boldsymbol{J'})$ otherwise, for all profiles $\boldsymbol{J}, \boldsymbol{J'}$. Intuitively, agent 3 only cares about the result on the conclusion. In particular, she wants the group to reject the conclusion, in agreement with her truthful judgment. Consider a profile $(J_3, \boldsymbol{J}^1_{-3})$ as represented in Table 4, and an alternative profile $(J^*_3, \boldsymbol{J}^1_{-3})$, where agent 3 reports the untruthful judgment J^*_3 (Table 5). It is easy to see that in this case, agent 3 prefers to vote untruthfully. Consider now another partial profile $\boldsymbol{J}^2_{-3}$, where agents 1 and 2 report different judgments (Tables 6 and 7). Agent 3 considers possible the partial profile $\boldsymbol{J}^2_{-3}$ too, as the zero-JIF does not provide any information about the judgments of the rest of the group. In that scenario, agent 3 is better off by being truthful. △

	a_1	a_2	c
Agent 1:	Yes	No	
Agent 2:	Yes	Yes	
Agent 3:	No	Yes	No
F^{pr}	Yes	Yes	**Yes**

Table 4: Profile $(J_3, \boldsymbol{J}^1_{-3})$

	a_1	a_2	c
Agent 1:	Yes	No	
Agent 2:	Yes	Yes	
Agent 3:	Yes	No	No
F^{pr}	Yes	No	**No**

Table 5: Profile $(J^*_3, \boldsymbol{J}^1_{-3})$

	a_1	a_2	c
Agent 1:	Yes	Yes	
Agent 2:	No	Yes	
Agent 3:	No	Yes	No
F^{pr}	No	Yes	**No**

Table 6: Profile $(J_3, \boldsymbol{J}^2_{-3})$

	a_1	a_2	c
Agent 1:	Yes	Yes	
Agent 2:	No	Yes	
Agent 3:	Yes	No	No
	Yes	Yes	**Yes**

Table 7: Profile $(J^*_3, \boldsymbol{J}^2_{-3})$

Restricting ourselves to the respectful JIFs, we are able to provide an axiomatization for all aggregation rules that are immune to manipulation under partial information. Theorem 3.9 states that whenever a JIF π is respectful in the previous interpretation, the π-strategy-proof aggregation rules are exactly those that are independent and monotonic, or equivalently those that are strategy-proof under full information (the proof is straightforward and can be found in the Appendix).

Theorem 3.9. *For any JIF π and aggregation rule F such that π respects C with regard to F, F is π-strategy-proof for the class C of all closeness-respecting preferences if and only if F is independent and monotonic.*

Theorem 3.9 draws out a conceptual issue concerning information in Judgment Aggregation. Briefly, what matters is not the quantitative distinction between full and partial information, but the qualitative one between relevant and irrelevant information. However, we cannot yet suggest a good aggregation rule by looking only at the relevance of the information that the agents hold. This is because the conditions of Theorem 3.9 are not necessary for susceptibility to manipulation. This means that there are judgment aggregation rules for which an agent with a particular closeness-respecting preference relation, even without holding all relevant information, still has a reason to manipulate the outcome. This is shown by the next result (to be read in combination with Example 3.7).

Theorem 3.10. *For a conjunctive agenda, for $n \geqslant 3$ and for every JIF π, the premise-based procedure is π-manipulable for the class of closeness-respecting preferences.*

Proof. Since every JIF π is at least as informative as the zero-JIF, it suffices to show that the premise-based procedure is manipulable for the zero-JIF π. We sketch the proof for a conjunctive agenda. Consider an agent i with a closeness-respecting preference $\succsim_i$ that only cares about the conclusion and wants it to be rejected (see Example 3.7). Then, her best strategy is to reject all the premises, even if truthfully she accepts some of those: If the conclusion was already rejected by the rest of the group, the agent has nothing to lose; but analogously to Example 3.7, there is always a partial profile for which the conclusion is accepted in case the agent remains honest (for a detailed proof, see Lemma 5.6 in Chapter 5). □

3.6 The Premise-based Procedure: An Extended Analysis

This section revolves around a specific aggregation rule, the premise-based procedure. We examine only the most relevant agendas for it, the conjunctive agendas (and all our results hold equivalently for disjunctive agendas). The premise-based procedure on the above agendas has received noticeable attention by economists and philosophers, especially because of its significance in the area of politics and law (Chapman, 2002; Pettit, 2001). A famous argument in favor of the premise-based way of aggregating individual judgments relates to deliberative democracy (Elster, 1998), supporting the view that collective decisions on conclusions should be set by the group's opinions on the premises. Moreover, an attractive characteristic of the premise-based rule is that it guarantees consistent outcomes. However, a point that deserves further investigation is its susceptibility to manipulation.

The first results in this direction are negative. It is shown that the premise-based procedure is manipulable for the class of all closeness-respecting preferences (Dietrich and List, 2007c) and we further proved in Theorem 3.10 that partial information

does not solve the problem. On the contrary, the premise-based procedure is manipulable for the class of all closeness-respecting preferences even under total lack of information. Our next step will be to study the manipulability of the premise-based procedure by restricting our attention to agents whose preferences are special cases of closeness-respecting preferences. Our reasoning goes as follows.

As we have seen, strategy-proofness is defined with regard to a class of preferences PR. This means that different specifications of the class PR correspond to different conditions of strategy-proofness. Practically, the larger a class of preferences is, the harder it becomes to achieve immunity to manipulation. The examples in the literature that are used to show that the premise-based procedure is manipulable hinge on agents whose preferences only care about conclusion in the agenda and completely ignore the collective decision on the premises. Dietrich and List (2007c) refer to these preferences as *outcome-oriented* (we will call them *conclusion-oriented* instead) and justify them by assuming that only the conclusion and not the premises carries consequences that the individuals care about.[18] But what about agents who also care up to some degree about the premises? It is known that when the preferences are *reason-based*, in the sense that the individual only cares to obtain a collective result on the premises that matches her own judgment and is indifferent to the conclusion, then the premise-based procedure is immune to manipulation. Moreover, we now show that immunity also holds when the agents possess Hamming-distance preferences.

Theorem 3.11. *Consider a conjunctive agenda Φ. The premise-based procedure F^{pr} is immune to manipulation for the class of all Hamming-distance preferences under full information.*

Proof. Suppose, aiming for a contradiction, that there is an agent i with Hamming-distance preferences $\succsim_i$ and a profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ for which agent i has an incentive to manipulate. Then, there is a judgment set J_i^* such that $F^{pr}(J_i^*, \boldsymbol{J}_{-i}) \succ_i F^{pr}(J_i, \boldsymbol{J}_{-i})$, which by the definition of Hamming-distance preferences means that the judgment set $F^{pr}(J_i^*, \boldsymbol{J}_{-i})$ has strictly more propositions in common with J_i than the propositions that the judgment set $F^{pr}(J_i, \boldsymbol{J}_{-i})$ does. But with the premise-based procedure, if agent i switches from reporting her truthful judgment J_i to reporting the untruthful judgment J_i^*, it is not possible to obtain a collective decision that is agreeing on a premise with J_i if the initial collective judgment was not agreeing on that premise with J_i. Hence, the only way for $F^{pr}(J_i^*, \boldsymbol{J}_{-i})$ to have a proposition in common with J_i that $F^{pr}(J_i, \boldsymbol{J}_{-i})$ does not is if that proposition is the conclusion. However, in order to achieve this, J_i^* should be untruthful and change the collective judgment on at least one of the premises that J_i and $F^{pr}(J_i, \boldsymbol{J}_{-i})$ agree on. In total, $F^{pr}(J_i^*, \boldsymbol{J}_{-i})$ cannot have strictly more propositions in common with J_i than the propositions that the judgment set $F^{pr}(J_i, \boldsymbol{J}_{-i})$ does, which is a contradiction. □

[18] Formally, an agent i with truthful judgment J_i has *conclusion-oriented preferences* if her preference relation $\succsim_i$ is such that $J \succ_i J'$ if and only if $c \in J_i \cap J$ and $c \notin J_i \cap J'$ or $\neg c \in J_i \cap J$ and $\neg c \notin J_i \cap J'$; and $J \sim_i J'$ if and only if $c \in J_i \cap J$ and $c \in J_i \cap J'$ or $\neg c \in J_i \cap J$ and $\neg c \in J_i \cap J'$.

Hamming-distance preferences assume that agents consider each formula in the agenda equally important and try to maximize their agreement with the collective outcome. However, the agents may care about every proposition in the agenda to a different degree. Such preferences make sense if every premise in the agenda is connected to a reason in support of the outcome that varies in priority for the agent, and the consequences of the group's opinion in the conclusion for the individual can be compared with the significance of those reasons too. Assume that each proposition $\phi \in \Phi$ is connected to a weight w_i^ϕ that denotes how much agent i cares about proposition ϕ. We take $\sum_{\phi \in \Phi} w_i^\phi = |\Phi|$. Moreover, $w_i^\phi = w_i^{\neg\phi}$ for all agents i, that is, the agents care the same about the acceptance and the rejection of ϕ, which is reasonable under the interpretation that the weights denote the importance that agents assign to propositions. Then, the weighted Hamming-distance between two judgment sets J and J' is:

$$H_w(J, J') = |\Phi| - \sum_{\phi \in \Phi} w_i^\phi \mathbb{1}_{\phi \in (J \cap J') \cup (\overline{J} \cap \overline{J'})}$$

Let H_w be the class of all weighted Hamming-distance preferences, defined in the standard way. That is, $H_w = \{H(J_i) : J_i \in \mathcal{J}(\Phi)\}$, where $H_w(J_i) = \succsim_i$ is such that $J \succ_i J'$ if and only if $H_w(J, J_i) < H_w(J', J_i)$ and $J \sim_i J'$ if and only if $H_w(J, J_i) = H_w(J', J_i)$. Theorem 3.12 is rather intuitive (and proven in the Appendix). It shows that if an agent cares less about a premise a than she does about the conclusion c, then she has an incentive to manipulate the premise-based procedure on a conjunctive or disjunctive agenda. Intuitively, the agent will choose to lie on her judgment about the premise a if by doing so she can achieve a more desirable result on the conclusion.

Theorem 3.12. *The premise-based procedure F^{pr} is manipulable for the class of preferences H_w on a conjunctive agenda Φ if and only if there is a premise a for which some agent i cares less than she does about the conclusion c ($w_i^a < w_i^c$).*

We will now consider a special instance of the weighted Hamming-distance preferences, namely the class of conclusion-prioritizing preferences K. The preference relations in K assume that the agents give the highest priority to the result on the conclusion, and secondarily, they try to maximize the agreement on the premises. Equivalently, K corresponds to those weighted Hamming-distance preferences for which all premises have equal weight and the conclusion is assigned greater weight than all the premises together.

Definition 3.8. Let Φ be a conjunctive agenda and Φ^p the set of its premises. We call K the class of all *conclusion-prioritizing* preferences $K = \{K(J_i) : J_i \in \mathcal{J}(\Phi)\}$, where $K(J_i) = \succsim_i$ is defined as follows. For all judgment sets $J, J' \in 2^\Phi$, $J \succ_i J'$ if: If $c \in J_i$, then (1) $c \in J$ and $c \notin J'$, or (2) $c \in J$ and $c \in J'$ (or $c \notin J_i$ and $c \notin J'$) and $|\{a : a \in J \cap J_i \cap \Phi^p\}| > |\{a : a \in J' \cap J_i \cap \Phi^p\}|$. If $\neg c \in J_i$, analogously.
Moreover, $J \sim_i J'$ if: If $c \in J_i$, then $c \in J$ and $c \in J'$ and $|\{a : a \in J \cap J_i \cap \Phi^p\}| = |\{a : a \in J' \cap J_i \cap \Phi^p\}|$. If $\neg c \in J_i$, analogously.

As one could expect, similarly to conclusion-oriented preferences (see Dietrich and List, 2007c), the premise-based procedure is manipulable for conclusion-prioritizing preferences under full information. However, under partial information the balance changes. The premise-based procedure is immune to manipulation for conclusion-prioritizing preferences, while it is still manipulable for conclusion-oriented preferences. Furthermore, the amount of information that needs to be absent in order to achieve the strategy-proofness is remarkably small. Speaking informally, truthfulness is guaranteed even when the agents know almost everything about the judgments of the rest of the group, i.e., even when the uncertainty of the agents in big agendas tends to 0. Theorem 3.13 makes this claim formal.

Theorem 3.13. *Consider a conjunctive agenda Φ. The premise-based procedure F^{pr} is susceptible to manipulation for the class of all conclusion-prioritizing preferences K under full information. However, there is a family of JIFs $\{\pi_x : x \in \mathbb{N}\}$ with $\lim_{x\to\infty} U(\pi_x) = 0$, such that F^{pr} is immune to π_m-manipulation for K, where m is the size of the agenda Φ.*

Lemma 3.14. *Consider a conjunctive agenda Φ, and let a be a premise in it. The premise-based procedure F^{pr} is immune to all_but_a-manipulation for the class of all conclusion-prioritizing preferences K.*

Proof. We give an outline of the proof. For any agent i, we can always find a partial profile that agent i considers possible for which the result agrees with her judgment on the conclusion both in case she lies and in case she remains truthful, while by being insincere on a agent i will induce a collective judgment set that disagrees with her on a, and thus agrees on less premises with her truthful judgment. Hence, this possible scenario forces agent i to remain truthful. □

Proof of Theorem 3.13. The premise-based procedure F^{pr} is susceptible to manipulation for the class of preferences K because similarly to the proof of Theorem 3.10, there is a profile where an agent i can change the result on the conclusion from disagreeing with her truthful judgment to agreeing with it by lying on a premise. However, as we will see next, F^{pr} is immune to manipulation under partial information. We construct a family of JIFs $\{\pi_x : x \in \mathbb{N}\}$, where each π_x is defined based on an agenda X with size x as follows. Take an arbitrary agenda X with size x. Fixing an arbitrary premise $a \in X$, we define $\pi_x :=$ all_but_a-JIF. Then, if the size of the agenda Φ is m, by Lemma 3.14 we have that the premise-based procedure is immune to π_m-manipulation. Moreover, we will show that $\lim_{x\to\infty} U(\pi_x) = 0$. Let us consider an arbitrary agent i and a profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$. We observe that when x tends to infinity, the number of all the possible (partial) profiles on an agenda X with size x tends to infinity too, i.e., $lim_{x\to\infty}|\mathcal{J}(X)^{n-1}| = \infty$. However, the number of all the partial profiles that agent i considers possible according to π_x will be finite. Specifically, $|\mathcal{W}_i^{1,\pi_x,\boldsymbol{J}}| \leqslant 2^{n-1}$, as only the opinion of each of the other $n-1$ agents with regard to proposition a is unknown to agent i. The following holds.

$$0 \leqslant \frac{|\mathcal{W}_i^{1,\pi_x,\boldsymbol{J}'}| - 1}{|\mathcal{J}(X)^{n-1}| - 1} \leqslant \frac{2^{n-1} - 1}{|\mathcal{J}(X)^{n-1}| - 1}.$$

Moreover, it is true that

$$\lim_{x\to\infty} \frac{2^{n-1} - 1}{|\mathcal{J}(X)^{n-1}| - 1} = 0.$$

Thus,

$$\lim_{x\to\infty} U_i^{\boldsymbol{J}}(\pi_x) = \lim_{x\to\infty} \frac{|\mathcal{W}_i^{1,\pi_x,\boldsymbol{J}'}| - 1}{|\mathcal{J}(X)^{n-1}| - 1} = 0.$$

Finally, since i and $\boldsymbol{J}$ were arbitrary, we have that $\lim_{x\to\infty} U(\pi_x) = 0$. □

This section can be considered to form only a beginning, towards a more complex analysis of the agents' motivations (that take the shape of preferences) in Judgment Aggregation. For the moment, we focused on the premise-based procedure and we showed that even under very weak assumptions on the uncertainty of the agents, strategy-proofness results can be radically influenced. This observation could contribute to broader discussions about the suitability of the premise-based rule with respect to different aggregation problems in Political Science (see, e.g., Miller, 1992).

3.7 Avoiding the Impossibility

The ultimate goal of Judgment Aggregation consists of two parts: the first aims at finding aggregation rules that can produce collective outcomes which best represent the group's opinions; the second tries to ensure truthfulness by the agents, so that the application of a "good" rule in the first sense in meaningful. Both parts have been associated with several impossibility results (see the Background for more details). Specifically, Dietrich and List (2007c) showed that for a class of very common agendas, (namely the path-connected agendas), there is no judgment aggregation rule that is non-dictatorial, complete, consistent, responsive, and immune to manipulation. However, a crucial assumption of the framework of Dietrich and List is that the agents hold full information. In this section we show that under partial information, this negative result is circumvented.

Theorem 3.15. *Consider an agenda Φ and a number of agents $n \geqslant 9$. The plurality rule F^{pl} along with a lexicographic tie-breaking rule is non-dictatorial, complete, consistent, responsive and immune to zero-manipulation for the class C of all closeness-respecting preferences.*

Proof. Suppose that the number of agents n is odd, $n = 2k + 1$, for some integer $k \geqslant 3$ (the case for n even is analogous). The axioms of non-dictatorship, completeness, consistency and responsiveness are easily checked to be satisfied for the plurality rule F^{pl} together with a lexicographic tie-breaking rule. Thus, we only have

to show that F^{pl} is immune to zero-manipulation for the class C of all closeness-respecting preferences. Take an arbitrary agent i, a profile $(J_i, \boldsymbol{J}_{-i})$, and a closeness-respecting preference $\succsim_i \in C(J_i)$, and suppose that there is a judgment set J_i^* such that $F^{pl}(J_i^*, \boldsymbol{J'}_{-i}) \succ_i F^{pl}(J_i, \boldsymbol{J'}_{-i})$, for some partial profile $\boldsymbol{J'}_{-i}$ (otherwise the rule is already immune to manipulation and the proof follows). By definition of the closeness-respecting preferences and the plurality rule, this can happen only if the collective outcome $F^{pl}(J_i, \boldsymbol{J'}_{-i})$ induced by agent i's truthful judgment is some judgment set J and the manipulated result $F^{pl}(J_i^*, \boldsymbol{J'}_{-i})$ is the judgment set J_i^*, so that $J_i^* \succ_i J$. Moreover, since $J_i \succsim_i J_i^*$, by transitivity it holds that $J_i \succ_i J$. We distinguish the following two cases, with regard to the tie-breaking rule order.

Case 1: $J_i > J$ in the tie-breaking linear order. Consider the profile $\boldsymbol{J''} := (J_i, \boldsymbol{J''}_{-i})$, where k agents submit the judgment set J_i, one agent reports the opinion J_i^*, and k other agents submit judgment J. Then, $F^{pl}(J_i, \boldsymbol{J''}_{-i}) = J_i$. However, if agent i reported the insincere judgment J_i^*, there would be $k-1$ agents submitting J_i, two agents submitting J_i^*, and k agents submitting J, hence $F^{pl}(J_i^*, \boldsymbol{J''}_{-i}) = J$. We conclude that $F^{pl}(J_i, \boldsymbol{J''}_{-i}) \succ_i F^{pl}(J_i^*, \boldsymbol{J''}_{-i})$, so agent i will not be willing to manipulate by reporting the untruthful judgment J_i^*.

Case 2: The tie-breaking rule ranks J above J_i. Then, consider the profile where $k+1$ agents submit the judgment set J_i, no-one submits J_i^*, and k agents submit J. The proof proceeds as in case 1. □

The main insight is that for the plurality rule, the agents can alter the outcome if and only if their opinion is pivotal. This has been spotted already by Obraztsova et al. (2013) in Voting under full information. Moreover, since we do not impose any restriction on the structure of the agenda, Theorem 3.15 suggests a way to avoid the original impossibility result of Gibbard and Satterthwaite, which has not been explicitly mentioned in the literature to the best of our knowledge. We further wonder whether there is a lower bound on the amount of information needed for such a positive result. Interestingly, we will formulate a property of a JIF π for which the plurality rule combined with a lexicographic tie-breaking order is guaranteed to be immune to π-manipulation for all closeness-respecting preferences. This property captures that every time an agent i could potentially have a reason to manipulate using an insincere J_i^* to avoid a not attractive collective outcome J, there is a scenario she considers possible where by reporting J_i^* she would make J win instead of her most preferred truthful opinion J_i. In principle, the type of uncertainty that the following *plurality-protection property* requires can be obtained in any aggregation problem where the agents are not aware of the opinions of a big enough part of their peers.

Definition 3.9. Consider a fixed lexicographic tie-breaking rule, an agenda Φ, a JIF π, and a judgment set $J_i \in \mathcal{J}(\Phi)$. We say that π has the *plurality-protection property*

with regard to J_i if for all judgment sets $J_i^* \neq J_i, J \in \mathcal{J}(\Phi)$ such that $\phi \in J_i^*$ and $\phi \notin J$ for some formula $\phi \in J_i$, and all profiles $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i}) \in \mathcal{J}(\Phi)^n$, it holds that:

- If the tie-breaking rule ranks J_i above J, then a profile where m agents submit the judgment set J_i, $n - 2m$ agents submit the judgment set J_i^* and m agents submit the judgment set J belongs to the information set $\mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ of agent i, for some integer m such that $\frac{n}{2} \geqslant m > \frac{n}{3}$.
- Otherwise, a profile belongs to $\mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ such that $m + 1$ agents submit the judgment set J_i, $n - 2m - 1$ agents submit the judgment set J_i^* and m agents submit the judgment set J, for some integer m such that $\frac{n-1}{2} \geqslant m > \frac{n-2}{3}$.

Theorem 3.16. *If the JIF π has the plurality-protection property with regard to J_i for all judgment sets $J_i \in \mathcal{J}(\Phi)$, then the plurality rule F^{pl} along with a lexicographic tie-breaking rule is immune to π-manipulation for all closeness-respecting preferences.*

Proof. It suffices to observe that for an agent i and her truthful opinion J_i, if $\gtrsim_i$ is a closeness-respecting preference in $C(J_i)$ and J, J' are two judgment sets in 2^{Φ}, then $J \succ_i J'$ implies that there is a formula $\phi \in J_i$ such that $\phi \in J$ and $\phi \notin J'$. Then, the proof is analogous to that of Theorem 3.15. □

To illustrate, consider the board of a company deciding on its financial policy for the coming year. The board consists of the president, three members from the management department and three members from the production department. Each individual knows that she has to submit a judgment among J_1, J_2 and J_3 and that the judgment which will be submitted by the greatest number of board members will determine the decision of the company (i.e., the plurality rule is used). Each member recognizes that employees who come from the same departments (excluding herself) share common interests, but since decisions about the financial policy of the company are confidential, no-one has established beyond doubt what the opinions of her peers are. Hence, each agent considers possible every scenario where members from the same departments submit the same judgment, but without knowing which this judgment is.

Consider, without loss of generality, a member i from the management department, and suppose that she detects a profile that could be possibly declared by the board where she would be strictly better off by lying. This can happen for example when her truthful judgment is J_1, the three agents of the production department submit the judgment J_2, and the other three agents of the board (the members of her department and the president) submit the judgment J_3, where the tie-breaking rule selects J_2, but the agent i strictly prefers J_3 to J_2. If agent i possessed full information, she would be dishonest without any hesitation in the previous scenario, reporting opinion J_3. On the other hand, under partial information she has to consider an alternative case too. It is possible that the three agents of her department, together with the president, support her judgment J_1. Then, if she chooses to lie and report the opinion J_3, she will end up strictly worse off, by making the judgment J_2 win instead of her truthful judgment.

We conclude that partial information can serve as a protection from manipulation, and consequently enable the use of a strategy-proof, complete, consistent and responsive aggregation procedure that is impossible to have under full information. However, the plurality rule is not absolutely appealing from a normative perspective, because it completely ignores the internal structure of the individual judgments. Especially in situations with many possible opinions and small groups, it leaves lot of power to the tie-breaking rule. We know, though, that all the representative-voter rules (see the Background) defined by Endriss and Grandi (2014) also satisfy completeness, consistency and responsiveness. Hence, a valid question is whether some of them are strategy-proof under partial information too. We investigate the average-voter rule, and we show that unfortunately it is manipulable, even under zero information.

Theorem 3.17. *There is an agenda Φ and a group $\mathcal{N}$ for which the average-voter rule F^{av} together with a lexicographic tie-breaking order is susceptible to manipulation under zero information for the class C of all closeness-respecting preferences.*

The detailed proof of Theorem 3.17 is given in the Appendix. The idea is that there is an aggregation problem with an agent i, whose only undesirable outcome evaluates all the propositions in the exact opposite way than a dishonest judgment J of hers. In this case, by submitting the untruthful judgment J, the Hamming distance between the collective decision and the agent's not wanted outcome never becomes smaller. Thus, lying by reporting J can never harm agent i when the average-voter rule is applied, while it can still make her better off in some possible scenario.

3.8 Concluding Remarks

We have seen that when we want to aggregate the judgments of a group and we assume that the agents are fully informed about the opinions of the others, finding a "good" rule that cannot be manipulated is too idealistic and not possible to achieve. However, under reasonable conditions of uncertainty, there is a rule that guarantees truthfulness, without having to sacrifice the completeness of the collective result, or even worse, its consistency or responsiveness. Hence, a stance in favor of accounting for partial information in Judgment Aggregation is defensible from various perspectives. Inside the framework we presented, we are able to uniformly capture realistic scenaria of individual reasoning with regard to collective procedures and study their impact on the agents' decisions as far as lying is concerned. In this chapter we started off by examining carefully the premise-based procedure and the plurality rule, but extensive research is required in order to map more aggregation rules to specific situations of partial information, where immunity to manipulation is achievable. We leave this challenge for further work, together with one conjecture: that the plurality rule is the only aggregation rule for which the impossibility theorem of Dietrich and List (2007c) is circumvented.

Chapter 4

Higher-level Strategic Manipulation

Hitherto we have been analyzing the level-1 reasoning of agents in aggregation problems, that is, we have been making the implicit assumption that the only parameter that affects their strategic behavior is the information they hold about the *truthful* judgments of the rest of the group. However, the members of a group are very likely to realize that their peers may reason strategically too, and thereby choose the best course of action in the light of their own information. This observation brings level-2 reasoning into the picture, which is triggered by an individual's uncertainty about the uncertainty of others. Now, the behavior of the agents is not merely depending on a passive environment. We will delve deeper into this account for *sophisticated individuals*. It is then natural to look at an agent's higher-order reasoning, which is materialized when the individual becomes aware of the fact that her peers reflect on her uncertainty about their uncertainty, and so on. Our basic intuitions stem from the fields of Epistemic Game Theory (e.g., Perea, 2012) and Epistemic Logic (e.g., Chopra et al., 2004; Halpern, 2005; Hendricks, 2006; Van Ditmarsch et al., 2012). This chapter elucidates the various levels of reasoning that take place in an agent's mind prior to making a final decision about which judgment to submit, and investigates the manipulability of aggregation rules accordingly, in a one-step procedure.[19]

As an illustration, consider a simple aggregation scenario that arises directly from the context of Game Theory. Two friends, Alice and Bob, have to decide as a group on whether to order pizza for dinner (p). The third friend in the company, Chris, plays a little game with them, telling them that they can order pizza if and only if exactly one of the two says that he or she wants it. Assume now that it is common knowledge that Alice loves pizza, while Bob is on a diet and does not want to have an unhealthy dinner. At the beginning Alice may think that, since Bob does not want pizza, she can

[19]The idea of modelling sophisticated agents in a Social Choice context was introduced by Farquharson (1969). In his pioneering work, he employed the method of *iterated elimination of dominated strategies* to decide the "rational" actions of higher-level reasoners, in a game-theoretical interpretation of Voting. However, little has been done since then regarding the study of the connections between interactive reasoning and the manipulability of aggregation procedures. In this thesis, we wish to bridge this gap in the literature, focusing on the framework of Judgment Aggregation.

be truthful, because she will be the only one accepting the proposition p and thus they will order her preferred meal for dinner (*first-order reasoning*). However, Alice could also think that, since Bob knows that she wants pizza, he has an incentive to lie by saying that he would also like to have some, so that Chris does not to allow them to make the order (*second-order reasoning*). In this case, an incentive for Alice to lie is created. Continuing this reasoning algorithm, Alice may think that Bob has already followed the previous reasoning in his mind, making the decision to tell the truth that he does not want pizza as he expects her to lie, and therefore it would be better for Alice to actually tell the truth, and so on (*higher-order reasoning*).

We realize that it is not clear how to determine at which level the reasoning of an agent terminates. Theoretically, the interactive reasoning of the agents in a group could proceed indefinitely. The question about which level of reasoning can be expected in practice by rational agents is addressed by behavioral scientists (e.g., Camerer et al., 2004; Costa-Gomes and Crawford, 2006; Costa-Gomes et al., 2001), whose empirical results are often not able to provide a categorical global answer. Despite of the limitations that the identification of the exact computational abilities of human beings presents, it is generally accepted that in common real-life strategic situations agents engage in thinking of at most three levels (Arad and Rubinstein, 2012; Camerer et al., 2004; Stahl and Wilson, 1995). Thus, we will restrict our focus on finite levels of interactive reasoning. Under these new assumptions of higher-level reasoning, we will explore basic Judgment Aggregation problems. The main issue that we wish to address is to what degree higher-level reasoning (possibly combined with partial information) can protect a rule from being susceptible to manipulation.

The remainder of this chapter is structured as follows. Section 4.1 begins with discussing and formalizing level-2 reasoning, and building on that, Section 4.2 defines the manipulability of aggregation rules. We continue with connecting former results of the literature and of Chapter 3 with our ongoing study (Section 4.3). In particular, we show that if an aggregation rule is immune to manipulation for reasoners of the first level under any kind of partial information, then the rule will preserve its immunity to manipulation also for second-level reasoners (and this theorem holds in general for higher-level reasoners). We provide some further intuitive examples to illustrate our model in Section 4.4. Subsequently, Section 4.5 designs the formal basis to model more sophisticated agents, whose reasoning goes beyond level-2, and their incentives for manipulation. We follow the concept of *level-k* reasoning, first introduced by Nagel (1995) and Stahl and Wilson (1995).[20] Then, a deeper insight on the relevant assumptions concerning the knowledge of the agents with respect to the preferences of their peers is provided in Section 4.6. Finally, the main result of this chapter is proven in Section 4.7, stating that any aggregation rule which is manipula-

[20]In experimental Voting Theory, the level-k model has been recently used by Bassi (2015). She showed that this model is relevant for the understanding of the agents' strategic choices when common rules, like the plurality rule, are applied. Our approaches can be said to be complementary rather than overlapping. The author conducts laboratory experiments in order to test human reasoning and behavior, while we are interested in the pure theoretical properties of aggregation rules.

ble by less sophisticated reasoners of the first level will also be manipulable by very sophisticated agents who reason in arbitrarily high levels. Hence, we conclude that, roughly speaking, higher-level reasoning cannot guarantee immunity to manipulation. We review and debate our findings in Section 4.8.

4.1 Information about the Information of Others

Consider an aggregation rule F and an agenda Φ. We assume that the information of the agents about the truthful opinions of the others is described by a JIF $\pi : \mathcal{J}(\Phi)^n \rightarrow \mathcal{I}$, as defined in Chapter 3. We will analyze situations where the agents perform up-to level-2 reasoning, that is, they reflect on the information that the others hold about the truthful profile of judgments and strategize accordingly. In our model all the agents are aware of the *type of information* that the rest of the group holds, which does not necessarily mean that they know the exact information of the others in a specific aggregation problem, but rather how that information is derived by the truthful profile, whatever that profile may be. More formally, we only assume that the JIF π is common knowledge among the agents.

The previous assumption makes sense in multiple aggregation scenaria. For instance, consider a social network whose structure is known to everyone in it. An example can be the board of a company, consisting of employees from different departments. Suppose that the board has to make a collective decision by aggregating the judgments of its members, and that several meetings in the different departments precede the final reporting of judgments. It is then practicable to assume that everyone knows the truthful opinions of the employees in her own department, and this is common knowledge. However, the agents cannot know what the truthful opinion of everyone else is, hence they lack the information about what exactly the others know about their colleagues; had they known more about the truthful profile, they could reconstruct the information that everyone holds. For the moment, what they know is the type, but not the full content of the group's information.

Apart from the truthful judgment that an agent holds, a key factor of her behavior in an aggregation problem is her preference relation over the possible collective outcomes. As we have seen, agents with the same truthful opinion may have an incentive to manipulate or not, depending on their different preferences. Hence, when examining the interactive reasoning of the members of a group, the assumptions considering the knowledge of the agents about the preferences of the others can be proven to be essential. In particular, when an agent reasons about the reasoning of another agent, there is a point where she has to wonder about the other agent's preferences. We will follow a basic intuition here, which prescribes that the preferences of the agents, in a different manner than their opinions, are not possible to be revealed to others. The judgments of the agents is what an aggregation procedure asks for. Thus, we will assume that this is also what the agents may have learned about via discussions about the aggregation situation. The preferences on the other hand are intrapersonal. They play

a role when an agent has to choose about lying or telling the truth, but this operation is never shared. A safe assumption is only that the agents know that everyone prefers results that match her own truthful opinion up to a degree. So, we will say that it is common knowledge that the preferences of the group belong to some specific class PR, and in practice this class will usually be taken to be the class C of all preferences that are closeness-respecting (recall the Background). Finally, we will assume that it is common knowledge that nothing more considering the preferences of the agents is common knowledge.[21]

Making the above formal, given a truthful profile of judgments $\boldsymbol{J}$ and a JIF π, an agent i's information about the truthful judgments of the rest of the group is given by $\pi_i(\boldsymbol{J})$ (recall Chapter 3). This information induces the set $\mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ of (partial) profiles that agent i considers possible to be the truthful ones, or in other words, the different scenaria about the judgments of the group that are compatible with her information and level-1 reasoning. However, after reflecting on the information that her peers hold, agent i may consider different profiles possible to be reported by the agents. Precisely, an agent performs second-order reasoning when she thinks that the other agents all reason in the first level and apply their best strategies accordingly. The set $\mathcal{W}_i^{2,\pi,\boldsymbol{J}}$ includes the partial profiles that agent i considers possible to be submitted by the group after she engages in level-2 reasoning. It may be the case that according to agent i's second-order reasoning, some other agent, say agent j, has an incentive to manipulate and report an untruthful opinion (following agent j's level-1 reasoning). Then, agent i will not consider the scenario where agent j is truthful possible anymore; on the contrary, the relevant cases for her will be those where agent j lies.

Definition 4.1. Consider an aggregation rule F, a truthful profile of judgments $\boldsymbol{J}$, a class of preferences PR, and a JIF π. Let $\mathcal{W}_i^{1,\pi,\boldsymbol{J}} := \{\boldsymbol{J}_{-i}^1, \ldots, \boldsymbol{J}_{-i}^r\}$ be the set of partial profiles that agent i considers possible that the group truthfully holds according to π. Then, for each such partial profile $\boldsymbol{J}_{-i}^v = (J_1^v, \ldots, J_{i-1}^v, J_{i+1}^v, \ldots, J_n^v) \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ and for each possible profile of preference relations $(\succsim_1, \ldots, \succsim_n)$ in class PR, we define a new set of partial profiles $\widetilde{\mathcal{W}}_i^{2,\pi,\boldsymbol{J}}(\boldsymbol{J}_{-i}^v, (\succsim_1, \ldots, \succsim_n))$ that agent i considers rational, that is, where her peers reason in the first level and report one of their best strategies when their truthful opinions are in $\boldsymbol{J}_{-i}^v$. Formally,

$$\widetilde{\mathcal{W}}_i^{2,\pi,\boldsymbol{J}}(\boldsymbol{J}_{-i}^v, (\succsim_1, \ldots, \succsim_n)) := \bigtimes_{j \neq i} S_j^F(\mathcal{W}_j^{1,\pi,(J_i,\boldsymbol{J}_{-i}^v)}, \succsim_j, J_j^v)$$

Finally, by taking the union of all the sets of rational partial profiles induced by any partial profile that agent i considers possible and any combination of preferences in the class PR for the group, we define the set

$$\mathcal{W}_i^{2,\pi,\boldsymbol{J}} := \bigcup_{v \in \{1,\ldots,r\}} \bigcup_{(\succsim_1,\ldots,\succsim_n) \in PR^n} \widetilde{\mathcal{W}}_i^{2,\pi,\boldsymbol{J}}(\boldsymbol{J}_{-i}^v, (\succsim_1, \ldots, \succsim_n))$$

[21] To better understand our assumptions with regard to the knowledge of the agents concerning the preferences of the others, consult Section 4.6.

of all partial profiles that are compatible with agent i's second-order reasoning.

When agent i reasons about the strategic reasoning of her peers, she deems them *rational*, meaning that she expects them to use their best strategies. Loosely speaking, we suppose that agent i treats the reasoning of the other agents it as if it were her own, in the sense that she assumes that the rest of the group compute their best reply following the same procedure as she does (the procedure of Definition 3.5). We then say that the agent i *believes in the other agents' rationality*. The assumption about *common belief in rationality* will be more evident in our analysis of Section 4.5, where the agents that we will study engage in higher-order reasoning.[22]

A useful remark is that $\widetilde{\mathcal{W}}_i^{2,\pi,\boldsymbol{J}}(\boldsymbol{J}^v_{-i}, (\succsim_1, \ldots, \succsim_n)) = \{\boldsymbol{J}^v_{-i}\}$ if and only if, considering the profile $\boldsymbol{J}^v_{-i}$, it is rational for all the agents to remain truthful. Following this observation, we see that if the aggregation rule is π-strategy-proof, that is, if no agent has an incentive to π-manipulate independently of the preferences she holds, then for every agent i it is true that $\mathcal{W}_i^{1,\pi,\boldsymbol{J}} = \mathcal{W}_i^{2,\pi,\boldsymbol{J}}$.

Note that for the set of partial profiles $\mathcal{W}_i^{2,\pi,\boldsymbol{J}}$, the axioms of reflexivity (REF), symmetry (SYM) and transitivity (TRANS), as defined in Section 3.1, are not valid in general. For example, we can see that the reflexivity axiom is violated in a situation as the following one: if agent i engages in second-order reasoning and deduces that agent j has an incentive to submit an untruthful judgment independently of what exactly agent j's preferences are, then agent i will not consider the truthful profile possible to be submitted anymore.[23]

4.2 Strategizing with Second-level Reasoning

In this section we define the incentives of an agent to manipulate an aggregation rule when she engages in second-order reasoning. As before, the absence of any incentive fosters strategy-proofness. An agent does not have an incentive to manipulate an aggregation rule under the information she holds if and only if, considering all the partial profiles that are compatible with her level-2 reasoning, her truthful judgment is her only best strategy (recall Section 3.3).

Definition 4.2. An aggregation rule is π-*manipulable under level-2 reasoning* for a class of preferences PR if and only if there are a profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ and an agent i who has preferences $\succsim_i \in PR(J_i)$, such that agent i, considering possible the set of

[22]Note that here we are using the word "belief" without paying attention to its philosophical or formal semantic meaning, that distinguishes it from "knowledge" and is studied in Epistemology (see, e.g., Armstrong, 1973; Hintikka, 1962). This makes sense when we look at the agents' epistemic attitudes only as triggering events of choice behavior, rather than as objects in their own right. Within this view, the *belief* of an agent can be considered a strong enough factor that determines her behavior, equivalently to her *knowledge*.

[23]As Lemma 4.7 shows however, the reflexivity axiom is always satisfied for the class of closeness-respecting preferences.

(partial) profiles $\mathcal{W}_i^{2,\pi,\boldsymbol{J}}$, has an incentive to manipulate on $\boldsymbol{J}$. That is, if and only if there is a judgment set $J_i^* \neq J_i$ such that $J_i^* \in S_i^F(\mathcal{W}_i^{2,\pi,\boldsymbol{J}}, \succsim_i, J_i)$.

As usual, an aggregation rule is strategy-proof if and only if it is not manipulable. Making use of Lemma 3.1, the following definition is straightforward.

Definition 4.3. An aggregation rule is *π-strategy-proof under level-2 reasoning* for a class of preferences PR if and only if for all profiles $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ and all agents i holding any preferences $\succsim_i \in PR(J_i)$, it holds that $S_i^F(\mathcal{W}_i^{2,\pi,\boldsymbol{J}}, \succsim_i, J_i) = \{J_i\}$.

At this point, a clarification of the terminology is required. The reader has probably realized that so far we have not made any explicit assumption about whether the agents of the groups that we examine are all reasoning in the same level. To be precise, when we argue that an aggregation rule is susceptible to manipulation under second-order reasoning, what would be more accurate to say is that the aggregation rule is manipulable whenever there is *at least one* agent in the group who is able to perform second-order reasoning. However, in order to claim that an aggregation rule is immune to manipulation under second-order reasoning, we have to refer to groups where *all* the agents reason in the second level. Intuitively, manipulability can be caused by the reasoning of only one more sophisticated agent, while strategy-proofness requires everyone to be at the same level (or more generally at a level that does not provide incentives for manipulation).

4.3 The Interplay between First-level and Second-level Reasoning

Our next topic of interest is the logical relation between first-order and second-order reasoning within the scope of strategic behavior. We first provide a basic result (Theorem 4.1), which specifies that all the independent and monotonic rules, besides being strategy-proof under first-order reasoning (Dietrich and List, 2007c), are also strategy-proof under second-order reasoning and any kind of partial information.

Theorem 4.1. *For every JIF π, if an aggregation rule F is independent and monotonic, then F is π-strategy-proof under level-2 reasoning for the class C of all closeness-respecting preferences.*

Proof. Analogous to the proof of Theorem 3.2. □

The fact that quota rules are independent and monotonic (Dietrich and List, 2007b) in combination with Theorem 4.1, implies the next Corollary.

Corollary 4.2. *Quota rules are immune to π-manipulation under level-2 reasoning for every JIF π, for the class C of all closeness-respecting preferences.*

Next, Theorem 4.3 asserts that immunity to manipulation under full information guarantees immunity to manipulation in all cases of partial information and second-order reasoning.

Theorem 4.3. *Consider the class C of all closeness-respecting preferences. Every aggregation rule that is immune to profile-manipulation under level-1 reasoning for C is also π-strategy-proof under level-2 reasoning for C, for every JIF π.*

Proof. We know by Dietrich and List (2007c) (recall Theorem 2.2) that every strategy-proof aggregation rule is independent and monotonic. Moreover, Theorem 4.1 states that every independent and monotonic aggregation rule is also π-strategy-proof under second-order reasoning, for any JIF π. □

The proceeding analysis concerns aggregation rules that are immune to manipulation under partial information, yet maybe susceptible to manipulation under full information (recall Theorem 3.8 for an instance of such a case). Does immunity to π-manipulation for first-level reasoners imply immunity to π-manipulation for reasoners of the second level? Theorem 4.4 answers positively.

Theorem 4.4. *Consider a class of preferences PR. For every JIF π and every aggregation rule F, if F is π-strategy-proof under level-1 reasoning for PR, then F is π-strategy-proof under level-2 reasoning for PR too.*

Proof. This is a special case of Theorem 4.11, which we prove in Section 4.7. □

Corollary 4.5. *Consider a class of preferences PR. For every JIF π and every aggregation rule F, if F is π-manipulable under level-2 reasoning for PR, then F is π-manipulable under level-1 reasoning for PR too.*

We have thus far partially studied the logical relation between manipulability under first-order and second-order reasoning. We know that the latter implies the former. A compelling further question is whether the opposite direction also holds. The answer is given by Theorem 4.6 (proven in the Appendix), which demonstrates an aggregation problem where an aggregation rule F is not strategy-proof under a natural information function when some agent reasons within level 1, but F is strategy-proof under that information function when everyone reasons in level 2.

Theorem 4.6. *Let F^{pl} be the plurality rule along with a lexicographic tie-breaking rule, and C the class of all closeness-respecting preferences.*

(a) F^{pl} is susceptible to winner-manipulation for C under first-level reasoning;

(b) F^{pl} is immune to winner-manipulation for C under second-level reasoning.

Hence, second-order reasoning is never harmful, and sometimes it is even beneficial for strategy-proofness.

4.4 Examples

To clarify the previous analysis, we reexamine two aggregation problems that have already been presented in this thesis, aiming our attention at the effects of second-order reasoning on them. We study in detail two extreme cases: one where a rule that is strategy-proof under first-order reasoning remains strategy-proof under reasoning of the second level, and another one where a susceptible to manipulation rule for first-order reasoners remains manipulable for reasoners of the second level.

Example 4.1 discusses the parity-type aggregation rule F that was defined at the proof of Proposition 3.8. Regarding level-1 reasoners, the rule F was shown to be susceptible to manipulation under full information, but strategy-proofness was achieved under relatively small uncertainty. Strategy-proofness remains obtainable under this kind of uncertainty in situations where the agents engage in level-2 reasoning too; Theorem 4.4 proves it in general, and Example 4.1 presents it explicitly for F.

Example 4.1. Consider an agenda Φ, an arbitrary formula $\psi \in \Phi$ and the rule F such that for all profiles $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ and formulas $\phi \in \Phi$, it is $\phi \in F(\boldsymbol{J})$ if $\phi \neq \psi$ and $\psi \in F(\boldsymbol{J})$ if and only if $|N_\psi^{\boldsymbol{J}}|$ is odd. Moreover, consider the JIF π, which is common knowledge and denotes that each agent i is completely unaware of the judgment of agent $i-1$, but knows the judgments of everyone else in the group. The aggregation rule F is not monotonic, hence we know by Dietrich and List (2007c) that it is not strategy-proof under full information and first-order reasoning. However, we saw in Proposition 3.8 that F is π-strategy-proof under first-order reasoning. Moreover, it is easy to see that F is π-strategy-proof under second-order reasoning. Take an arbitrary agent i, who reflects on the incentives of the other agents to manipulate. Agent i deduces that the present uncertainty about the truthful profile makes all the agents of the group remain truthful (recall that a second-order reasoner thinks that everyone else reasons in the first level); therefore she knows that agent $i-1$ will be truthful too. However, agent i is still ignorant of what the truthful opinion of agent $i-1$ is. The uncertainty agent i has after reasoning in the second level is reduced to exactly the same uncertainty she would have if she reasoned only within the first level, and this uncertainty guarantees that she does not have an incentive to manipulate. $\triangle$

Example 4.2 considers the plurality rule, which was proven to be central in Section 3.7 for avoiding the main impossibility result in Judgment Aggregation with respect to strategic manipulation (Dietrich and List, 2007c). We proved that by depriving first-order reasoners of some information about the truthful profile of judgments, we can achieve truthfulness for the (also consistent, complete and responsive) plurality rule, which on the contrary is susceptible to manipulation under full information. Our question now is: Can we reach strategy-proofness following a different direction, that is, accounting for agents who reason in higher levels? To that end, we investigate whether incentives for manipulation for fully informed, second-order reasoners when the plurality rule is applied disappear. Our answer, unfortunately, is negative.

Example 4.2. We claim that the plurality rule F^{pl} combined with a lexicographic tie-breaking rule is manipulable under second-order reasoning and full information. We will sketch the idea of the proof via an aggregation problem with eight agents that have three complete and consistent judgments J_1, J_2 and J_3 to choose from, where the tie-breaking order is $J_1 > J_2 > J_3$. Recall that if some agent i reasons within level-2, then she thinks that her peers all reason in level-1, i.e., that they only reflect on their own information about the truthful profile of the group. Consider a profile $\boldsymbol{J}'$ that is compatible with agent i's second-order reasoning, where three agents submit the judgment set J_1, two agents report J_2, and other three agents submit J_3. So, J_1 is the resulting collective decision. Suppose now that agent i's truthful judgment is J_2 and that she strictly prefers the judgment set J_3 to the judgment set J_1. Then, she has an incentive to lie by submitting opinion J_3 and making it win, unless there is another profile compatible with her second-order reasoning where she would end-up strictly worse off by doing so. This "safety"-profile should be exactly a profile where three agents submit J_1, four agents submit J_2, and one agent reports J_3 (there, agent i's truthful profile J_2 wins, but if she switched to J_3, she would make J_1 win). We will show that this profile is not compatible with agent i's second-order reasoning. First, checking all the possible cases, one can verify the profile $\boldsymbol{J}'$ is possible from the perspective of the fully informed, second-order reasoner agent i, only if it coincides with the truthful profile. But in the profile $\boldsymbol{J}'$, only agents who submit the judgment J_2 would be able to change the result and thus have an incentive to manipulate under first-order reasoning. This means that, according to agent i's level-2 reasoning, all the three members of the group who submit the judgment J_3 in the truthful profile should keep their choice fixed. Then, a profile where only one agent submits J_3 is not compatible with agent i's second-order reasoning. We conclude that there is no "safety"-profile for agent i, so she has an incentive to lie under second-level reasoning and full information, making the rule susceptible to manipulation. △

Table 8 collects all the results about the manipulability of the plurality rule that we have proven so far.[24,25]

	full information	winner information	zero information
first-level reasoning	×	×	✓
second-level reasoning	×	✓	✓

Table 8: Manipulability of the plurality rule.
" ✓" denotes immunity and " × " susceptibility to manipulation.

[24]The positive results hold for top-respecting or more restricted classes of preferences; the negative results hold for any kind of preferences, as long as some agent can strictly order two dishonest opinions.

[25]For readability reasons, Table 8 does not mention JIFs with the *plurality-protection property* (recall Definition 3.9). Note that the plurality rule is immune to manipulation under first-order reasoning (and hence also second-order reasoning) under this kind of information too.

4.5 Strategizing with Higher-level Reasoning

In this section we continue with the modeling of higher levels of reasoning, generalizing our previous study. We design our framework along the lines of the *level-k reasoning model* (Nagel, 1995; Stahl and Wilson, 1995). Recall that level-1 reasoners only speculate about their own information about the possible truthful judgments of the rest of the group, while level-2 reasoners give further thought to the information that the others hold about the truthful profile. Level-k agents, now, are able to apply exactly k levels of this reasoning operation; they reason about what the other agents know about what the other agents know about ... what the other agents know about the truthful judgments of the group. In other words, level-k agents think that everyone else reasons within level-$(k-1)$.

In order to locate level-k reasoners' incentives to lie, we will refer to the set of profiles that are compatible with their higher-level reasoning. Definition 4.4 builds inductively on the analogous definitions that concern level-1 and level-2 reasoning (of Sections 3.2 and 4.2 respectively).

Definition 4.4. Consider an aggregation rule F, a class of preferences PR, and a JIF π. We will define the set of (partial) profiles $\mathcal{W}_i^{k,\pi,\boldsymbol{J}}$ that agent i considers possible to be submitted by the group, when she engages in level-k reasoning and the truthful profile of judgments is $\boldsymbol{J}$:

- As in Section 3.1, we define $\mathcal{W}_j^{1,\pi,\boldsymbol{J'}}$ the set of partial profiles that agent j would consider possible to be the *truthful* ones, if the actual truthful profile was $\boldsymbol{J'}$. Let $\mathcal{W}^{1,\pi,\boldsymbol{J}} := \{J_1, \ldots, J_r\}$.

- Having defined the set of partial profiles $\mathcal{W}_j^{k-1,\pi,\boldsymbol{J'}}$ that an agent j considers possible to be submitted by the group when she engages in level-$(k-1)$ reasoning and the truthful profile is $\boldsymbol{J'}$, we proceed as follows.

  For every partial profile $\boldsymbol{J}_{-i}^v \in \mathcal{W}^{1,\pi,\boldsymbol{J}}$ that agent i considers possible to be the truthful one, and for each possible profile of preference relations $(\succsim_1, \ldots, \succsim_n)$ in PR, we define a new set of partial profiles $\widetilde{\mathcal{W}}_i^{k,\pi,\boldsymbol{J}}(\boldsymbol{J}_{-i}^v, (\succsim_1, \ldots, \succsim_n))$ that agent i considers rational, that is, where her peers reason in level-$(k-1)$ and report one of their best strategies when their truthful opinions are in $\boldsymbol{J}_{-i}^v$. Formally,

$$\widetilde{\mathcal{W}}_i^{k,\pi,\boldsymbol{J}}(\boldsymbol{J}_{-i}^v, (\succsim_1, \ldots, \succsim_n)) := \bigtimes_{j \neq i} S_j^F(\mathcal{W}_j^{k-1,\pi,(J_i,\boldsymbol{J}_{-i}^v)}, \succsim_j, J_j^v)$$

  Finally, by taking the union of all the sets of rational partial profiles induced by any partial profile that agent i considers possible to be the truthful one and any combination of preferences in the class PR for the group, we define the set

$$\mathcal{W}_i^{k,\pi,\boldsymbol{J}} := \bigcup_{v \in \{1,\ldots,r\}} \bigcup_{(\succsim_1,\ldots,\succsim_n) \in PR^n} \widetilde{\mathcal{W}}_i^{k,\pi,\boldsymbol{J}}(\boldsymbol{J}_{-i}^v, (\succsim_1, \ldots, \succsim_n))$$

Similarly to previous sections concerning reasoners of lower levels, an agent who engages in level-k reasoning is assumed to believe in the rationality of her peers, in the sense that she expects them to compute their best strategy — using the same procedure as she does (recall Definition 3.5) — and behave accordingly. Moreover, a level-k reasoner believes that everyone else in the group believes that everyone else is rational, that everyone believes that everyone believes that everyone else is rational, and so on. We call this assumption *common belief in rationality* and it is a crucial parameter in the investigation of this chapter.[26]

Definition 4.5. An aggregation rule is *π-manipulable under level-k reasoning* for a class of preferences PR if and only if there are a profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ and an agent i holding preferences $\succsim_i \in PR(J_i)$, such that agent i has an incentive to manipulate on $\boldsymbol{J}$, considering possible the set of (partial) profiles $\mathcal{W}_i^{k,\pi,\boldsymbol{J}}$ to be submitted by the group. That is, if and only if $S_i^F(\mathcal{W}_i^{k,\pi,\boldsymbol{J}}, \succsim_i, J_i) \neq \{J_i\}$ (recall Definition 3.5).

An aggregation rule F is π-strategy-proof under level-k reasoning if and only if F is not π-manipulable under level-k reasoning.

Definition 4.6. An aggregation rule is *π-strategy-proof under level-k reasoning* for a class of preferences PR if and only if for all profiles $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ and all agents i holding any preferences $\succsim_i \in PR(J_i)$, it holds that $S_i^F(\mathcal{W}_i^{k,\pi,\boldsymbol{J}}, \succsim_i, J_i) = \{J_i\}$.

Once more, we should stress the fact that an aggregation rule is guaranteed to be immune to manipulation under level-k reasoning only if *all* the agents in the group reason in level-k exactly. If there exists some less or more sophisticated agent in the group, then strategy-proofness should be established for her level of reasoning too.

4.6 Common Knowledge on Preferences — An Example

As we discussed at the beginning of this chapter, we assume that only a wide class, where the agents' preferences belong to, is common knowledge to the members of a group (this class is in general designated by PR, but in practice it is often taken to be the class C of all closeness-respecting preferences). In order to better understand how common knowledge on the class of the agents' preferences affects their higher-level

[26] Indeed, most of the main results of Section 4.7 would change completely if we assumed that the agents do not hold the same rationality assumptions about their peers. Even though one could argue that this assumption constitutes an oversimplification of the actual reasoning of the agents, it is common to make it in theoretical disciplines that deal with the interactive reasoning of individuals in strategic situations, such as Epistemic Game Theory (e.g., Perea, 2012) and Epistemic Logic (e.g., Bacharach et al., 2012). Intuitively, we wish to study the interaction of rational individuals that apply the same method in deciding whether to lie or tell the truth and they consider it rational for others to do so, too. The reader is encouraged to reflect further on this issue.

reasoning procedure, we will revisit the example of Alice and Bob. So, Alice and Bob want to decide on whether to order pizza (p) for dinner. Their friend Chris lets them know that the aggregation rule that he is going to use to make a decision for them is the odd-parity rule F, which prescribes that p is accepted by the group if and only if an odd number of individuals accepts it. Alice and Bob can thus order pizza if and only if exactly one of them admits that he or she wants to have pizza for dinner. Alice likes pizza a lot ($J_a = \{p\}$), while Bob avoids it most of the time, because he is on a diet ($J_b = \{\neg p\}$). The truthful judgments of the two agents are common knowledge. We distinguish two cases for our illustration. In the first case, both Alice's and Bob's preferences are directly related to their judgments, i.e., Alice strictly prefers to order pizza for dinner than eating something different ($\{p\} \succ_a \{\neg p\}$), while for Bob having pizza is the strictly worst scenario $\{\neg p\} \succ_b \{p\}$), and this is common knowledge. This means that it is common knowledge that Alice and Bob have Hamming-distance preferences. In the second case, even if Alice and Bob actually have Hamming-distance preferences as in the first case, it is common knowledge that their preferences are closeness-respecting, but they are both unaware of the exact preference relation of the other. This means that Alice considers possible the scenario where Bob is indifferent about whether they will order pizza or not for this specific dinner, and the same holds for Bob with regard to Alice. The results of the various levels of reasoning of the two agents are depicted in Tables 9 and 10. Level 0 represents the truthful judgments of the agents. In every level of reasoning k, the written profile is the one that would be submitted if both agents reasoned in level-k. Moreover, the judgments that are in bold in each level denote that the agents who hold them have an incentive to manipulate when they reason in that level. In order to see the incentives for manipulation at a level k, we have to compare the collective result of rule F in that level with the truthful judgment of the agent in question. There are two reasons that can lead to manipulation. If the collective result is in favor of the agent and the agent is lying to achieve it, then he or she will keep lying; if the result is making the agent worse off while he or she is truthful, then he or she will choose to lie instead.

	p	p	p	p	p	
Alice:	Yes	Yes	**No**	**No**	Yes	...
Bob:	No	**Yes**	**Yes**	No	No	...
F	Yes	No	Yes	No	Yes	...
levels of reasoning	0	1	2	3	4	

Table 9: Common knowledge on Hamming-distance preferences.

We observe that in the first case, where it is common knowledge that the agents possess Hamming-distance preferences, neither Alice nor Bob have an incentive to manipulate when they reason in the third level. Hence, they submit their truthful

judgments and the profile that appears in level 4 is the same as the truthful profile of level 0 (see Table 9 above). Afterwards, the sequence of the submitted profiles for higher levels of reasoning will be repeated. Generalizing the above example and analyzing in a similar way all the possible initial truthful profiles, we can see that the aggregation rule that Chris suggested is strategy-proof when the two agents reason at any level-k with $k \equiv 0 \mod 4$.

	p	p	p	p	p	
Alice:	Yes	Yes	Yes	Yes	Yes	...
Bob:	No	**Yes**	**Yes**	**Yes**	**Yes**	...
F	Yes	No	No	No	No	...
levels of reasoning	0	1	2	3	4	

Table 10: Common knowledge on closeness-respecting preferences.

In the second case, where it is common knowledge only that the preferences of the agents are closeness-respecting, however, the status of the rule changes. Specifically, it is the case that in every level of reasoning there is one agent (namely Bob) that has an incentive to manipulate. The difference in comparison to the first case is due to the following simple fact: although when Bob reflects on the truthful profile (depicted in level 0) has an incentive to manipulate according to his Hammning-distance preferences, Alice is not certain about that. In reality, Alice still considers possible the case where Bob remains truthful, because she thinks that maybe he is indifferent about the collective outcome. But if Bob remained truthful, then a lie by Alice would make her strictly worse off. This uncertainty leads Alice to tell the truth at level 1. Now, Bob, who can follow this reasoning and knows that Alice will tell the truth, manages to achieve the desirable result for him by lying. Since Alice — no matter her level of reasoning — will never be safe enough to manipulate, Bob will always have an incentive to do so (see Table 10 above). We conclude that in this scenario the rule that Chris suggested is susceptible to manipulation under any level of reasoning.

This example emphasizes an important aspect of the research direction that we pursue here. It is now clear that the preferences of the agents, as well as the knowledge of the group about those preferences, play a principal role with respect to the strategy-proofness of a rule in Judgment Aggregation. Contrary to classical brunches of Social Choice Theory, such as Voting Theory and Preference Aggregation, in Judgment Aggregation there are numerous reasonable ways to generate individual preferences from individual opinions, which implies that our relevant assumptions can be critical for the theoretical results. Notably, we just saw that when uncertainty about the exact preferences of the others increases, then an aggregation rule can be transformed from strategy-proof under some level of reasoning to manipulable.

4.7 The Interplay between Lower and Higher Levels of Reasoning

We now investigate the logical connections between first-order and higher-order reasoning with regard to the strategy-proofness of aggregation rules (all the proofs of this section follow from our definitions and are explained in the Appendix).

To begin, we show that for every aggregation rule F, if an agent has an incentive to manipulate F under zero information when engaging in first-order reasoning, then she will still have an incentive to manipulate F by reasoning in any higher level (Theorem 4.8). Lemma 4.7 is essential to establish the above. Keeping in mind that we work with closeness-respecting preferences, when an agent i is not aware of the precise preference relations of her peers, but she is only certain that these preferences are closeness-respecting, then all the (partial) profiles that she considers possible to be the truthful ones will be compatible with her higher-order reasoning. Intuitively, this holds because according to agent i, the other agents in the group may be indifferent between all the alternative outcomes of the aggregation; and in such a case, they will all choose to remain truthful.

Lemma 4.7. *Suppose that the agents have closeness-respecting preferences in the class C and this is common knowledge. However, the agents are not aware of the exact preferences of the others, and this is common knowledge too. Then, for every JIF π, for every agent i, profile $\boldsymbol{J}$ and level of reasoning k, it is the case that $\mathcal{W}_i^{1,\pi,\boldsymbol{J}} \subseteq \mathcal{W}_i^{k,\pi,\boldsymbol{J}}$.*

Theorem 4.8. *Consider an aggregation rule F and the zero-JIF π. If F is zero-manipulable under level-1 reasoning for the class of all closeness-respecting preferences C, then F will also be zero-manipulable for C under level-k reasoning, for every level k.*

Theorem 4.8 is further instantiated by two corollaries. Corollary 4.9 focuses on the average-voter rule, which, considering first-level reasoners, is known to be susceptible to manipulation under full information (since the average-voter rule is not independent, this follows by Dietrich and List (2007c) and their Theorem 2.2 formulated in the Background). Moreover, the average-voter rule is also susceptible to manipulation under zero information for at least some specific agenda (see Theorem 3.17). So, even for reasoners of higher levels, there will always be that agenda for which the average-voter rule is susceptible to manipulation under zero information.

Corollary 4.9. *There are an agenda Φ and a set of agents $\mathcal{N}$ for which the average-voter rule combined with a specific lexicographic tie-breaking rule F^{av} is susceptible to zero-manipulation under level-k reasoning for any natural number k, for the class C of all closeness-respecting preferences.*

Corollary 4.10 refers to the premise-based procedure, which is susceptible to manipulation under full information (Dietrich and List, 2007c), and is also manipulable

under zero information, as we discuss in Section 3.6 (see Theorem 3.10). We now know that even if we restrict our analysis to more sophisticated agents, the premise-based procedure will still be susceptible to manipulation under zero information.

Corollary 4.10. *The premise-based procedure is susceptible to zero-manipulation under level-k reasoning for every level k, for the class C of all closeness-respecting preferences.*

Interestingly, despite the fact that higher-order reasoning is not always able to protect an aggregation rule from potential manipulation, we will show that it is never damaging with respect to strategy-proofness. That is, independently of the information about the truthful judgments of the group that is available to the agents, if an aggregation rule is immune to manipulation for agents who only engage in first-level reasoning, then the rule will be immune to manipulation for higher-order reasoners too (Theorem 4.11). But under which circumstances can agents who perform higher-order reasoning guarantee that a rule which was susceptible to manipulation for less sophisticated agents turns to be immune to manipulation? Having an aggregation procedure that is manipulable under some partial information when the agents of the group reason in the first-level, a desirable result would establish that if all the agents engaged in reasoning of at least r levels for some natural number r, then the rule would become π-strategy-proof. Unfortunately, we will show that this is never the case. On the contrary, consider any aggregation rule that is manipulable under first-order reasoning. Even if we may be able to identify a natural number r for which the rule is strategy-proof for groups consisting only of level-r reasoners, if there is at least one agent who can potentially go one step further and reason in level-$(r+1)$, this will cause the manipulability of the rule.

Theorem 4.11. *Consider any class of preferences PR, any judgment aggregation rule F, and any JIF π. If F is immune to π-manipulation under level-1 reasoning for PR, then F is immune to π-manipulation under level-k reasoning for PR, for every k.*

Theorem 4.12. *Consider any class of preferences PR, any judgment aggregation rule F, and any JIF π. If F is susceptible to π-manipulation under first-level reasoning for PR, then even if F is immune to π-manipulation for PR under level-k reasoning for some k, it will be susceptible to π-manipulation under level-$(k+1)$ reasoning.*

After all, we are able to fully address how higher-level reasoning affects strategy-proofness in our model. Theorem 4.12 states that if a rule is susceptible to manipulation when the group reasons in the first level, then the rule can never be strategy-proof for two consecutive levels of reasoning. This result makes it impossible to argue that higher-level reasoning can prevent manipulation in a global manner.

On the one hand, behavioral experiments only provide evidence that in strategic problems of real-life, people reason within an interval of levels (a common approach is to attempt to obtain a probability distribution over reasoning levels. See for instance the recent work by Penczynski, 2016). Hence, every aggregation rule that is

manipulable under first-level reasoning can in practice be considered manipulable under higher levels of reasoning too, conditioning on our assumptions for human beings (note for instance that we do not consider complexity issues here). Roughly speaking, as convergence to strategy-proofness via higher-level speculations is never guaranteed, first-level reasoning determines whether a rule can be considered manipulable or not (all the above holds independently of the information available to the agents).

On the other hand, our results can be more promising regarding agents of artificial intelligence, for example in the context of Multiagent Systems. Within such a framework, an agent may be programmed to reason in a fixed level, and this level can be chosen from the modeler to be such that ensures strategy-proofness.

A more theoretical observation may appeal to the reader who is keen on the pure mathematical properties of our reasoning model: The manipulability status of every aggregation rule is characterized by an elegant periodicity. We shall first notice informally that when an aggregation rule is strategy-proof under a level k, then all the agents who perform reasoning in the next level $k+1$ believe that everyone else will be truthful, hence the scenaria they consider possible are exactly the same as the ones compatible with first-level reasoning. Intuitively, this is the reason why after a strategy-proof reasoning-level k, the reasoning of the agents from the modeler's perspective reduces to level 1. More generally and formally, this implies that the following three cases constitute a partition of the set of all aggregation rules, or, in other words, any aggregation rule F belongs to exactly one of these categories (see Table 11):

1. F is strategy-proof for level-k reasoning, for every natural number k.
2. F is manipulable for level-k reasoning, for every natural number k.
3. F is strategy-proof for level-k reasoning, for every natural number k such that $k \equiv 0 \mod r$, where $r \neq 1$ is some natural number (that may differ for different rules), and F is manipulable for level-k' reasoning, for every natural number k' such that $k' \not\equiv 0 \mod r$.

reasoning levels:	1	2	...	r	$r+1$	$r+2$	...	$2r$	$2r+1$	...
Case 1:	✓	✓	✓	✓	✓	✓	✓	✓	✓	...
Case 2:	×	×	×	×	×	×	×	×	×	...
Case 3:	×	×	×	✓	×	×	×	✓	×	...

Table 11: Possible manipulability categories for any aggregation rule F. " ✓" denotes strategy-proofness and " × " manipulability.

4.8 Concluding Remarks

In this chapter we pursued the study of agents who perform advanced interactive reasoning, that is, agents who attempt to reason about the strategic reasoning of their peers, inside a formal framework which extends the one of the previous chapter. We provided the toolbox to uniformly incorporate partial information and higher-level reasoning in Judgment Aggregation, thereby enriching the current literature in the area. Our investigation revolved around one main hope: that agents who are able to and willing to give deeper thought to the intentions of their peers with respect to manipulation would eventually find it more worthy to remain truthful themselves. Sadly, this hope was rather refuted. No matter which aggregation procedure we may choose to use, if we cannot achieve truthfulness for uncomplicated reasoners of the first level, then there will always be an arbitrarily high level of reasoning for which our rule will still be susceptible to manipulation. Nevertheless, our analysis could also be considered fruitful from a more philosophical point of view, because it illuminates some facts about interactive strategic reasoning that a technically-oriented mind may tend to forget. Sophisticated speculations about the reasoning of other people in an agent's environment, even when simplified by convenient theoretical assumptions, are costly with regard to an individual's time and mental energy. The challenge posed by interactive reasoning is twofold. First, computing the final outcome of an aggregation rule (taking certainty over the truthful judgments of the group for granted) is often intractable (Endriss and de Haan, 2015; Endriss et al., 2012; Hemaspaandra et al., 2005). Hence, one can simply imagine the difficulties that an agent faces at the moment when she realizes that she has to compute, not only the outcomes compatible with her information, but also those that the rest of the group may consider possible according to the information each possesses, etc. Many of these difficulties are evident in the simple examples presented in this chapter. Second, the mental complications of sophisticated reasoning are also notable up to a different degree. Even if a human being could have supernatural powers as far as computation is concerned, in order to perform higher-level reasoning she should additionally *conceive* what it means for other persons to reason about other persons' reasoning about others' reasoning with regard to the group's reasoning, and so on. Consider now an agent who can actually follow this procedure for, say, three steps. The agent knows that it makes sense to keep reasoning in this way, but she just loses track of the steps after a while. Moreover, the agent understands her limitations and she accepts them as she recognizes that most people in her group will be restricted in a similar way too. However, imagine a case where the various levels of reasoning change completely the perspective of the agent with regard to the rational choices of her peers (recall for instance the example in Section 4.6). Since the agent knows that she *should* be able to reason in the fourth level but she is only able to reach the third, it seems quite unreasonable to expect that she will choose her best strategy of the third level. Which strategy would she then choose? We will not try to provide an answer in the context of this thesis; but, we wish to high-

light the multiple types of uncertainty that higher-level reasoning may imply for an agent. Given that this way of thinking seems to be an unpredictable burden, a decision theorist that accepts the truth-bias assumption in Judgment Aggregation could argue that agents who are smart enough to estimate the complexity of the situation will end up submitting their truthful opinion, going after the only option that looks safe. We conclude this chapter by inviting further interdisciplinary research, including the collaboration of scholars from the areas of Social Choice Theory and Cognitive Science (among others), towards the direction of discovering further insights about agents' strategizing in aggregation scenaria, or connecting the dots between various research domains that have so far been developed separately.

Chapter 5

Iterative Judgment Aggregation

So far we have been presupposing that the manipulability of aggregation procedures is unwelcome. However, total truthfulness seems to be an ideal that is very hard to attain in practice. Therefore, we now shift our research focus and ask: What if we stop fighting manipulation? In order to investigate this new question, we will introduce a framework of *Iterative Judgment Aggregation*, inspired by Iterative Voting (e.g., Airiau and Endriss, 2009; Grandi et al., 2013; Lev and Rosenschein, 2012; Meir et al., 2010; Reijngoud and Endriss, 2012); see Meir (2017) for an overview. Our model proceeds in rounds. In the beginning, every agent submits an opinion and is (maybe partially) informed about the judgments submitted by the others. Then, one of the agents, who thinks strategically and wishes to submit a new opinion, proceeds with doing so, and her peers are (maybe partially) informed about the caused changes. In the next round another agent has the opportunity to realize her strategic behavior, and so on. We restrict our attention to agents who always reason in level-1, due to our results of Section 4.7, which, loosely speaking, suggest that the untruthful acts of higher-level reasoners are reducible to the status of an aggregation rule in the first level. Focusing on two manipulable rules that have been consistently analyzed throughout this thesis, the *premise-based procedure* and the *plurality rule*, we tackle the following points:

(a) Does an iterative process terminate, and under which conditions?

(b) If a convergence state is reached, how fast does this happen?

(c) Are the collective decisions that are obtained by enabling iterative manipulation valuable for the group as a whole?

The remainder of this chapter is structured as follows. Section 5.1 introduces the main framework of Iterative Judgment Aggregation for fully informed agents. Based on that framework, the premise-based procedure is studied in Section 5.2 and notably, it is shown to always converge. Section 5.3 proceeds with incorporating partial information in iterative procedures of aggregation. After some standard definitions are generalized, a further analysis of the premise-based procedure, this time under zero

information, is pursued in Section 5.4. Luckily, we are able to show that the iteration will still be guaranteed to reach a terminal state, and interestingly, the convergence speed is faster, in comparison to the case of full information, if and only if the agents are initially insincere. Next, the plurality rule is examined thoroughly in Section 5.5, and we see that higher uncertainty does not always function in favor of convergence. The chapter continues with Section 5.6. Issues concerning the social benefits of truthfulness are addressed, and our results are rather disapointing. Specifically, we show that being sincere when the premise-based procedure is applied can be infinitely detrimental for the group as a whole, and also that the damage caused by truthfulness in the plurality rule increases linearly with regard to the number of possible judgments that the agents may hold. Our final point in question is whether strategic behavior is able to smoothen the aforhementioned results. For the former case, the news are greatly positive: the iteration of the premise-based procedure always achieves the optimal social outcome. However, this is not true for the plurality rule. We summarize and open some routes for further research in Section 5.7.

5.1 The Model under Full Information

Let F be an aggregation rule. We consider an iteration of the aggregation procedure, where in each round the rule F receives a profile of the agents' opinions as input and prescribes a (temporary) collective outcome. To start off, we assume that in every round t all the agents are fully informed about the group's profile $\boldsymbol{J}_t$ (this assumption is to be dropped in later sections). Then, we look whether there is an agent who wants to alter her previously submitted judgment aiming for a more desirable collective result for her. If this is the case, one such agent is randomly selected and has the chance to re-submit her judgment. What is crucial moreover, is that in every round the agents choose a reaction with regard to the information that is available to them in that specific round only, that is, they are *memory-less*. In addition, they are *myopic*, in the sense that they only aim at a better outcome in the next round, said differently, they treat every round as if it were the last one.[27]

5.1.1 Improvement Dynamics

An agent will change her previously submitted judgment if she can be better off by doing so. But what does "better off" precisely mean inside the iterative framework? First, when an agent is fully informed about the judgments of her peers, it may be the case that she is able to reach a strictly preferable outcome for her in the next round by switching to a new judgment. However, an agent will not always be able to affect the outcome. Then, she has two options: either to keep her current judgment (even

[27]The assumptions of memory-less and myopic agents are standard in Iterative Voting. See for example the first work in the area by Meir et al. (2010).

if it is an untruthful one) or to change her judgment and report her truthful opinion (in case she was not doing that before). The former type of individual overcomes her truth-bias, trying not to be herself the reason for which the aggregation procedure will not terminate; hence, we call her *inertia-friendly*. This behavior is reasonable to expect when agents are tired of a time- and energy-consuming iterative procedure and weight changing their judgment as rarely as possible more than being truthful. On the opposite, *inertia-averse* individuals value truth more than stagnation.[28] Consider a profile $\boldsymbol{J}_t = (J_{i,t}, \boldsymbol{J}_{-i,t})$ submitted in round t of the iterative procedure for rule F.

Definition 5.1. Consider an inertia-friendly agent i, with truthful judgment J_i and preferences $\succsim_i$. Agent i, under full information, has an opportunity to perform an *improvement step* in round t, using the judgment $J_{i,t+1}$, if

$$F(J_{i,t+1}, \boldsymbol{J}_{-i,t}) \succ_i F(J_{i,t}, \boldsymbol{J}_{-i,t})$$

Definition 5.2. Consider an inertia-averse agent i, with truthful judgment J_i and preferences $\succsim_i$. Agent i, under full information, has an opportunity to perform an *improvement step* in round t, using the judgment $J_{i,t+1}$, if

$$F(J_{i,t+1}, \boldsymbol{J}_{-i,t}) \succ_i F(J_{i,t}, \boldsymbol{J}_{-i,t})$$

or if there is no $J'_{i,t+1} \in \mathcal{J}(\Phi)$ such that $F(J'_{i,t+1}, \boldsymbol{J}_{-i,t}) \succ_i F(J_{i,t}, \boldsymbol{J}_{-i,t})$, but it is $J_{i,t+1} = J_i \neq J_{i,t}$, and

$$F(J_{i,t+1}, \boldsymbol{J}_{-i,t}) \succsim_i F(J_{i,t}, \boldsymbol{J}_{-i,t})$$

5.1.2 Best Improvement Steps

It will usually be the case that an agent has different ways available to perform an improvement step in a round. Then, it is reasonable to assume that she will restrict her options to those that are the best, that is, in game-theoretical terms, *undominated*. Consider a round t of the iterative procedure for rule F, where the declared profile of the group is $\boldsymbol{J}_t = (J_{i,t}, \boldsymbol{J}_{-i,t})$, and agent i has preferences $\succsim_i$. A judgment set $J \in \mathcal{J}(\Phi)$ is undominated if there is no other judgment set J' such that $F(J', \boldsymbol{J}_{-i,t}) \succ_i F(J, \boldsymbol{J}_{-i,t})$. We now define the set $BI_{i,t}$ of the judgments that can be used by agent i for a *best improvement step* in round t as follows.[29]

[28] Meir et al. (2010) refer to inertia-averse individuals as *truth-biased*. However, in our framework of Chapters 3 and 4 that dealt with strategic manipulation in a single round, we have characterized as truth-biased the agents who choose to tell the truth in case this choice gives them at least as good an outcome as lying in all possible scenarios consistent with their information. In those terms, we could say that inertia-averse agents (or truth-biased agents in the terminology of Meir et al., 2010) demonstrate one extra layer of truth-bias, choosing truth over stability. Nonetheless, to avoid overlaps with our previously used terminology we will stick to the term *inertia*, which specifically concerns iterative procedures.

[29] $BI_{i,t}$ depends on more parameters, such as the rule F, the agent i's preferences $\succsim_i$, as well as her inertia-type, which are omitted in order to lighten the notation.

- Since the agents are truth-biased, if agent i's truthful judgment J_i is undominated and can be used by her for an improvement step in round t, then this will be her only best improvement step. That is, $BI_{i,t} := \{J_i\}$.
- Otherwise, $BI_{i,t} := \{J \in \mathcal{J}(\Phi) : J \text{ is undominated and can be used by agent } i \text{ for an improvement step in round } t\}$.

It is easy to see that if the set of judgments that can be used by an agent i for an improvement step in round t is non-empty, then the set of judgments that correspond to agent i's best improvement steps will be a non-empty set too.

5.1.3 Response Policies

An agent's *response policy* captures some additional natural assumptions that can restrict her choices regarding her best improvement steps even further. Consider agent i that has the preference relation $\gtrsim_i$ and holds the truthful judgment J_i, while the partial profile of the rest of the group in round t is $\boldsymbol{J}_{-i,t}$.

Definition 5.3. We distinguish four conditions with respect to the judgment set $J_{i,t+1} \in \mathcal{J}(\Phi)$ (chosen randomly if there are more than one) that the agent may submit in the next round.

- If the agent is *outcome-focused*: The Hamming-distance between what the agent submits in round $t+1$ and the collective outcome is minimized, i.e., $J_{i,t+1} \in \arg\min_{J \in BI_{i,t}} H(J, F(J, \boldsymbol{J}_{-i,t}))$.
- If the agent is *round-focused*: The Hamming-distance between what the agent submits in round $t+1$ and what she was submitting in round t is minimized: $J_{i,t+1} \in \arg\min_{J \in BI_{i,t}} H(J, J_{i,t})$.
- If the agent is *truth-focused*: The Hamming-distance between what the agent submits and her truthful judgment is minimized, i.e., $J_{i,t+1} \in \arg\min_{J \in BI_{i,t}} H(J, J_i)$.
- If the agent is *unrestricted*: $J_{i,t+1} \in BI_{i,t}$ and no further restriction is imposed.

Depending on the profile that the agents submit in the first round, and the order in which they are modifying their opinions, different *improvement paths* may be created. We say that the iterative procedure *converges to a stable state* (or an *equilibrium*) if every improvement path terminates after a finite number of rounds. In other words, an equilibrium, as in standard Game Theory, is a profile where no agent can profit by a unilateral deviation, or where there is no improvement step available for any agent.[30]

[30] Some scholars have studied iterative voting procedures with respect to the convergence in stable *outcomes* (e.g., Endriss et al., 2016b). Notice that we do not follow this direction here.

Obviously, if an iteration procedure terminates for unrestricted agents, then it also always reaches a convergence state for agents who belong in any other of the three categories. This is easy to see thinking of the contrapositive case: if there is an improvement path that leads to a cycle for agents who use one of the more restricting policies, then the same path may happen to be followed by unrestricted agents too; thus, non-convergence in the restricting cases implies non-convergence for unrestricted agents.

Moreover, strategy-proof aggregation rules can be vacuously said to converge after zero rounds, as their iterative procedure never commences. We know that all independent and monotonic rules are immune to manipulation (see Theorem 2.2 in the Background, by Dietrich and List, 2007c), and moreover that the quota rules are independent and monotonic (Theorem 2.1 in the Background, formulated by Dietrich and List, 2007b). Hence, the following corollary holds.

Corollary 5.1. *Every profile $\boldsymbol{J}$ is an equilibrium profile with respect to the quota rules.*

On the other hand, for some alternative rules it is easy to see that their iteration may never terminate.

Example 5.1. Consider the odd-parity rule F according to which a formula is accepted by the group if and only if an odd number of agents accepts it. Let Φ be an agenda consisting of only one formula ϕ and its negation $\neg\phi$, and let the group $\mathcal{N}$ contain only two agents, Alice (A) and Bob (B). Suppose that the opinions of A and B on ϕ differ, so that A accepts ϕ and B rejects it. In addition, both A and B strictly prefer the collective decision to agree with their individual judgment rather than to disagree with it. Then, whatever judgments the two agents submit, and independently of whether they are inertia-friendly or inertia-averse and which policy they follow, one of them will always have an opportunity to perform an improvement step. Specifically, if in round t rule F accepts ϕ, then Bob will be better off by modifying his judgment in round $t+1$, conditionally that Alice keeps her submitted opinion fixed. If in round t rule F rejects ϕ, the same holds for Alice. This means that the agents will never be able to reach an equilibrium. △

5.2 Iterative Premise-based Procedure under Full Information

The various incentives of agents to misrepresent their truthful opinions when the premise-based procedure is applied are discussed extensively in Section 3.6. Now, considering a sequence of judgment aggregation rounds under the premise-based procedure, the two central questions are whether and how fast agents will agree on a collective decision from which no-one has an incentive to deviate. We work with a

conjunctive agenda Φ (but note that all our results hold equivalently for a disjunctive agenda too), whose non-negated part is $\Phi^+ := \{a_1, \ldots, a_k, c\}$, where $a_1, \ldots, a_k$ are propositional variables, and $c \leftrightarrow (a_1 \wedge \ldots \wedge a_k)$. Building on the insights of Section 3.6, which demonstrates that substantial reasons for manipulability arise mainly when the agents have conclusion-oriented preferences, we will treat this case.[31,32]

We begin with studying the iterative procedure when in the first round all the agents submit their truthful opinions. This assumption makes sense for agents that rely on the fact that if the collective outcome is not satisfactory, then in the future they will have the chance to change their judgments. Then, the iteration is guaranteed to terminate after at most one round.

Theorem 5.2. *Consider a conjunctive agenda Φ, and fully informed agents with conclusion-oriented preferences. Independently of other assumptions on the agents, the premise-based procedure F^{pr} converges from the truthful profile in at most one round.*

Proof. Call A the set of agents who accept the conclusion and R the set of agents who reject it. We assume that all the agents have conclusion-oriented preferences, hence in the first round they have an opportunity for an improvement step if and only if the collective decision on c disagrees with their own judgment on c and they can change that. The agents in A should accept all the premises in order to accept the conclusion and have no way to manipulate. Suppose now that an agent i in R has an incentive to manipulate, which means that the collective result accepts the conclusion in round 1 and agent i can make the group reject the conclusion in round 2. In the second round, agents in A have still no way to manipulate, and all the agents in R are happy because they obtain their desirable outcome. Hence, everyone who is truthful has no reason to manipulate and the only untruthful agent cannot return to her truthful judgment, as that would make her worse off. We conclude that the procedure stops in one round. □

We continue with the investigation of agents who may not submit their truthful judgment at the beginning of the iterative procedure. This assumption makes sense for example when considering individuals who may get involved in more sophisticated planning and try to guide the procedure towards the most desirable result for them in the long-term. Theorems 5.4 and 5.5 indicate that the premise-based procedure is still guaranteed to reach an equilibrium, but it may take a linear number of rounds with respect to the number of the agents n. Lemma 5.3 is quite intuitive, and plays an important role in the sequel. It is formally proven in the Appendix.

[31] Recall that an agent i with truthful judgment J_i has *conclusion-oriented preferences* if her preference relation $\succsim_i$ is such that $J \succ_i J'$ if and only if $c \in J_i \cap J$ and $c \notin J_i \cap J'$ or $\neg c \in J_i \cap J$ and $\neg c \notin J_i \cap J'$; and $J \sim_i J'$ if and only if $c \in J_i \cap J$ and $c \in J_i \cap J'$ or $\neg c \in J_i \cap J$ and $\neg c \in J_i \cap J'$.

[32] Since conclusion-oriented agents are indifferent about the collective decision on all the premises, it is meaningless to consider them outcome-focused. Thus, one could assume that they follow some of the other policies, but this will not affect our results by any means.

Lemma 5.3. *Consider a conjunctive agenda Φ, and a conclusion-oriented agent i, who truthfully accepts the conclusion in Φ. If agent i has an opportunity to perform an improvement step in round t of the iterative premise-based procedure, then she is untruthful in round t, and her unique best improvement step is performed by being truthful in round $t+1$, (i.e., accepting all the premises and the conclusion).*

Theorem 5.4. *Consider a conjunctive agenda Φ, and fully informed, inertia-friendly agents with conclusion-oriented preferences, who follow any policy. The premise-based procedure F^{pr} converges from any initial profile in at most $2n$ rounds.*

Proof. Call A the set of agents who accept the conclusion and R the set of agents who reject it. By Lemma 5.3, in any round t, an agent $i \in A$ has an opportunity to make a best improvement step if and only if she moves to her truthful opinion from an insincere one. This move can be realized at most once for every agent in A. On the other hand, an agent $j \in R$ has an opportunity to perform an improvement step in round t if she can choose a judgment which makes a previously accepted conclusion be rejected by the group. How many times would agents in R have the chance to flip the result on the conclusion? If the conclusion is collectively rejected, only an agent $i \in A$ can turn it to be accepted again in the future, by submitting her truthful judgment that she was not submitting before. This can happen at most $|A|$ times, because as we saw, every agent in A will move at most once. In total, no-one will have an available improvement step after at most $|A| + |A| \leqslant 2n$ rounds. □

Theorem 5.5. *Consider a conjunctive agenda Φ, and fully informed, inertia-averse agents with conclusion-oriented preferences, who follow any policy. The premise-based procedure F^{pr} converges from any initial profile in at most $3n$ rounds.*

The main insight of Theorem 5.5 is that inertia-averse agents who reject the conclusion may have an opportunity to perform an improvement step, by being truthful, also in case the group already rejects the conclusion. This fact adds at most n rounds in the result of Theorem 5.4. The detailed proof can be found in the Appendix.

5.3 Adding Partial Information

In this section we focus on iterative procedures where in every round the agents are partially informed about the submitted profile of their peers and the information they hold is described by some JIF π (see Chapter 3, Section 3.1). Now, the agents make improvement steps that take their information into account.

5.3.1 Improvement Dynamics under Partial Information

We consider risk-averse agents in accordance with our analysis in the previous chapters, and we say that an agent will be better off by submitting a new judgment if (1)

there is a scenario consistent with her information about the current round under which she is able to achieve a strictly preferable outcome for her in the next round and (2) there is no scenario under which her new judgment will lead her to a strictly less desirable outcome. Analogously to the case of full information, when the agent is not able to alter the collective outcome, she has two options: either to keep her current judgment (even if it is an untruthful one), or to change her judgment and report her truthful opinion (in case she was not doing that before). These two types of individuals are called *inertia-friendly* and *inertia-avert* respectively, as before. Consider a profile $\boldsymbol{J}_t = (J_{i,t}, \boldsymbol{J}_{-i,t})$ declared in round t of the aggregation procedure for rule F.

Definition 5.4. Consider an inertia-friendly agent i, with truthful judgment J_i and preferences $\succsim_i$. Agent i, under the information described by the JIF π, has an opportunity to perform an *improvement step* in round t using the judgment $J_{i,t+1}$, if

1. $F(J_{i,t+1}, \boldsymbol{J'}_{-i,t}) \succ_i F(J_{i,t}, \boldsymbol{J'}_{-i,t})$, for some $\boldsymbol{J'}_{-i,t} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}_t}$ and
2. $F(J_{i,t+1}, \boldsymbol{J''}_{-i,t}) \succsim_i F(J_{i,t}, \boldsymbol{J''}_{-i,t})$, for all other $\boldsymbol{J''}_{-i,t} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}_t}$

Definition 5.5. Consider an inertia-averse agent i, with truthful judgment J_i and preferences $\succsim_i$. Agent i, under the information described by the JIF π, has an opportunity to perform an *improvement step* in round t using the judgment $J_{i,t+1}$, if

1. $F(J_{i,t+1}, \boldsymbol{J'}_{-i,t}) \succ_i F(J_{i,t}, \boldsymbol{J'}_{-i,t})$, for some $\boldsymbol{J'}_{-i,t} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}_t}$ and
2. $F(J_{i,t+1}, \boldsymbol{J''}_{-i,t}) \succsim_i F(J_{i,t}, \boldsymbol{J''}_{-i,t})$, for all other $\boldsymbol{J''}_{-i,t} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}_t}$

or if there is no $J'_{i,t+1} \in \mathcal{J}(\Phi)$ such that the above conditions hold, but it is $J_{i,t+1} = J_i \neq J_{i,t}$, and $F(J_{i,t+1}, \boldsymbol{J'}_{-i,t}) \succsim_i F(J_{i,t}, \boldsymbol{J'}_{-i,t})$, for all $\boldsymbol{J'}_{-i,t} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}_t}$.

5.3.2 Best Improvement Steps under Partial Information

Similarly to the case of full information, we assume that partially informed agents are also able to distinguish their *best improvement steps* and, if they have the opportunity, perform one of them. Consider a round t of the iterative procedure for rule F, where the declared profile of the group is $\boldsymbol{J}_t = (J_{i,t}, \boldsymbol{J}_{-i,t})$, and agent i holds the preferences $\succsim_i$ and her information is described by the JIF π. Then, a judgment set $J \in \mathcal{J}(\Phi)$ is *undominated* in the standard game-theoretical sense if there is no other opinion J' such that (1) $F(J', \boldsymbol{J'}_{-i,t}) \succ_i F(J, \boldsymbol{J'}_{-i,t})$, for some $\boldsymbol{J'}_{-i,t} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}_t}$ and (2) $F(J', \boldsymbol{J''}_{-i,t}) \succsim_i F(J, \boldsymbol{J''}_{-i,t})$, for all other $\boldsymbol{J''}_{-i,t} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}_t}$ (see Definition 3.5). Then, we define the set $BIP_{i,t}$ of the judgments that can be used by agent i for a *best improvement step* under her partial information in round t:[33]

[33] $BIP_{i,t}$ depends on more parameters, such as the rule F, the agent i's preferences $\succsim_i$, as well as her information π_i and her inertia-type, which are omitted to lighten the notation.

- Since the agents are truth-biased, if agent i's truthful judgment J_i is undominated and can be used for an improvement step in round t, then this will be the agent's only best improvement step. That is, $BIP_{i,t} := \{J_i\}$.
- Otherwise, $BIP_{i,t} := \{J \in \mathcal{J}(\Phi) : J \text{ is undominated and can be used for an improvement step by agent } i \text{ in round } t\}$.

5.3.3 Response Policies under Partial Information

We now need to refine the *response policy* notion for agents under partial information. With regard to round-focused, truth-focused and unrestricted agents, the definitions of their policies under full and partial information are totally analogous. This is the case because these policies are not concerned with the (partially unknown) collective outcome; they only look at the agents' previously submitted judgment, their truthful judgment, or nothing at all, respectively. Consider agent i that has the preference relation $\succsim_i$ and holds the truthful judgment J_i, while the partial profile of the rest of the group in round t is $\boldsymbol{J}_{-i,t}$, and agent i is partially informed about it, by the JIF π.

Definition 5.6. The agent chooses among her available best improvement steps under the partial information she holds (randomly if there are more than one). Moreover,

- If the agent is *round-focused*: The Hamming-distance between what the agent submits in round $t+1$ and what she was submitting in round t is minimized: $J_{i,t+1} \in \underset{J \in BIP_{i,t}}{\arg\min}\, H(J, J_{i,t})$.
- If the agent is *truth-focused*: The Hamming-distance between what the agent submits and her truthful judgment is minimized, i.e., $J_{i,t+1} \in \underset{J \in BIP_{i,t}}{\arg\min}\, H(J, J_i)$.
- If the agent is *unrestricted*: $J_{i,t+1} \in BIP_{i,t}$ and no further restriction is imposed.

However, adapting outcome-focused agents to the framework of partial information requires some further consideration; according to their policy, they try to submit a judgment close to the collective decision, which they may not be able to fully predict. Recall Definitions 5.4 and 5.5. Suppose that judgment $J_{i,t+1}$ offers an opportunity for an improvement step to agent i. We define the judgment set $J^A_{i,t+1}$, which is the most desirable collective outcome that agent i aims for, induced by the judgment set $J_{i,t+1}$:

1. If all possible outcomes that the agent can achieve by changing her judgment are equally valuable for her, then she has an improvement step only if she is inertia-averse, and that step uses her truthful judgment J_i. So, $J^A_{i,t+1} := F(J_i, \boldsymbol{J}_{i,t})$;

2. Otherwise, there exists some $\boldsymbol{J}'_{-i,t} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}_t}$, such that (1) $F(J_{i,t+1}, \boldsymbol{J}'_{-i,t}) \succ_i F(J_{i,t}, \boldsymbol{J}'_{-i,t})$ and (2) $F(J_{i,t+1}, \boldsymbol{J}''_{-i,t}) \succsim_i F(J_{i,t}, \boldsymbol{J}''_{-i,t})$, for all $\boldsymbol{J}''_{-i,t} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}_t}$ (call this condition for $\boldsymbol{J}'_{-i,t}$ $(\star)$). Then,

$$J^A_{i,t+1} := \max_{\boldsymbol{J}'_{-i,t} \text{ s.t. } (\star)} F(J_{i,t+1}, \boldsymbol{J}'_{-i,t})$$

Finally, we can say that an outcome-focused agent tries to minimize the distance with one of her targeted, most desirable collective decisions.

- If the agent is *outcome-focused*: The Hamming-distance between what the agent submits and her most desirable collective outcome she aims for is minimized, i.e., $J_{i,t+1} \in \arg\min_{J \in BIP_{i,t}} H(J, J^A)$.

The preceding definition is simple to understand via an example for the case of the plurality rule (together with a lexicographic tie-breaking rule), when the agents are only informed about the currently winning judgment set in every round. Suppose that an agent strictly prefers a judgment set J' over J and she is indifferent between all the other judgment sets and J. Nevertheless, suppose that it happened and the agent is currently submitting judgment J, which is winning. The agent cannot be worse off by changing her judgment, so she considers making an improvement step and targets the win of her more desirable judgment set J'. Then, she has at least two options to choose from: she can either withdraw her support from judgment J and submit a different judgment J'' (looking at the scenario she considers possible where judgment J' is already close enough to win after J loses one vote), or she can directly submit J'. An agent who is outcome-focused will choose the second option.

5.4 Iterative Premise-based Procedure under Zero Information

Recall that we are studying premise-based aggregation on a conjunctive agenda (and that the reasoning for a disjunctive agenda is equivalent), and we focus on conclusion-oriented individuals. The crucial observation that determines our results is twofold: on the one hand, any agent who wants the conclusion of the agenda to be accepted truthfully accepts all the premises, and she can never manipulate the outcome by lying; on the other hand, an agent who rejects the conclusion can manipulate by lying on some of the premises (by rejecting them even if she truthfully accepts them), and moreover, when she is completely ignorant of the submitted profile, then her unique best move is to reject all the premises (Lemma 5.6, proven in the Appendix).

Lemma 5.6. *Consider a conjunctive agenda Φ, the premise-based procedure F^{pr}, and a zero-informed agent i who has conclusion-oriented preferences and rejects the*

conclusion in Φ. Then, agent i's option to reject all the premises dominates all her other options.

Thanks to Lemma 5.6, we are able to compute the convergence speed of the premise-based procedure under zero information.

Theorem 5.7. *Consider a conjunctive agenda Φ, and zero-informed agents who have conclusion-oriented preferences. Independently of other assumptions on the agents, the premise-based procedure F^{pr} converges from any initial profile (including the truthful one) in at most n rounds.*

Proof. The agents who accept the conclusion truthfully accept all the premises and have no way to manipulate by lying, no matter the information they possess. So, being truthful can potentially be the only best improvement step for them. Moreover, by Lemma 5.6, every agent who rejects the conclusion can potentially make a unique improvement step, by rejecting all the premises. Thus, all the agents may perform an improvement step at most once. Thus, the iterative procedure stops in at most n (the number of agents) rounds. □

Recalling the results of Section 5.2 on full information, we reach some interesting observations about the effects of partial information on the convergence speed of the premise-based procedure. When the agents are completely ignorant of the judgments that their peers submit, the iteration can take up to n rounds to converge, no matter whether the agents initially submit their truthful profile or not. However, when we deal with agents under full information, things change. Then, starting from the truth guarantees termination in one round, while submitting non-truthful initial opinions may lead to convergence after $3n$ rounds. So, withholding information from the agents can be both damaging and beneficial, depending on whether sincerity in the first round is to be expected or not. This makes evident that partial information is an essential parameter in the study of Iterative Judgment Aggregation. Our results are summarized in Table 12.

	full information	zero information
inertia-friendly from truth	1	n
inertia-averse from truth	1	n
inertia-friendly from any profile	$2n$	n
inertia-averse from any profile	$3n$	n

Table 12: Iterative premise-based procedure: convergence speed.

5.5 Iterative Plurality

In Voting Theory, the properties of the plurality rule in iteration under full information have been extensively studied by Meir et al. (2010). Nonetheless, the agents' pref-

erences over the possibly interconnected judgment sets in Judgment Aggregation is what theoretically may distinguish the applications of the plurality rule here, in comparison to Voting. Hence, it is worth examining to what degree the existing formal results of the literature in Voting are still valid in our framework.

5.5.1 Iterative Plurality under Full Information

Outcome-focused agents capture a reasonable assumption: whenever they can directly submit a preferable judgment set and turn it into the collective outcome, then they will do so, instead of manipulating the result indirectly. Theorem 5.8 (corresponding to Meir et al., 2010, Theorem 3) establishes that when the agents are outcome-focused and inertia-friendly, the iterative plurality procedure is guaranteed to converge.

Theorem 5.8. *The plurality rule F^{pl} paired with a lexicographic tie-breaking rule converges from any initial profile for inertia-friendly, outcome-focused agents with closeness-respecting preferences, under full information.*

Proof. First, observe that when the plurality rule is used for individuals that are inertia-friendly and outcome-focused, the winning judgment sets will always have a non-decreasing number of supporters in future rounds (we call this fact $(\star)$).

Let $M(t)$ be the set of judgment sets that may be held in round t by some agent who has an improvement step to make. We will show that $M(t) \supseteq M(t+1)$ for every round t, and that the set of rounds $\{t : M(t) = M(t+1)\}$ is finite. Denote with $n^t(J)$ the number of voters who submit the judgment set J in round t. If an inertia-friendly and outcome-focused agent i has an opportunity for an improvement step in round t, it means that she will submit a judgment set $J_{i,t+1}$ that is preferable to her, and make it win in round $t+1$. Suppose that the winning judgment set in round t was J_t. Then, it must be $n^{t+1}(J_{i,t+1}) \geqslant n^t(J_t)$. We will show that the judgment set $J_{i,t}$ will only lose supporters in the future. We distinguish two cases.

Case 1: $n^t(J_{i,t+1}) = n^t(J_t) - 1$ and the tie-breaking rule ranks J_i^{t+1} above J_t. Then, after the move of agent i, we have that $n^{t+1}(J_{i,t+1}) = n^t(J_t)$.
Sub-case $1a$**:** If the tie-breaking rule ranks $J_{i,t}$ above $J_{i,t+1}$ (and hence above J_t), then $n^t(J_{i,t}) \leqslant n^t(J_t) - 1$, so $n^{t+1}(J_{i,t}) \leqslant n^t(J_t) - 2 \leqslant n^{t+1}(J_{i,t+1}) - 2$. Thus, judgment $J_{i,t}$ cannot win in round $t+1$ or in any later round (because of fact $(\star)$), and hence no new agent will have an incentive to submit judgment $J_{i,t}$ in the future.
Sub-case $1b$**:** If the tie-breaking rule ranks $J_{i,t}$ below $J_{i,t+1}$, then no new agent has a reason to submit judgment $J_{i,t}$ in the future, because $n^{t+1}(J_{i,t}) \leqslant n^{t+1}(J_{i,t+1}) - 1$ (and fact $(\star)$).

Case 2: $n^t(J_{i,t+1}) = n^t(J_t)$. Then, after the move of agent i, it will hold that $n^{t+1}(J_{i,t+1}) = n^t(J_t) + 1$. Since $n^t(J_{i,t}) \leqslant n^t(J_t)$, we have that $n^{t+1}(J_{i,t}) \leqslant$

$n^t(J_t) - 1 \leqslant n^{t+1}(J_{i,t+1}) - 2$. Thus, no agent will have an incentive to submit judgment $J_{i,t}$ in the future, because of fact $(\star)$.

Overall, if an agent who has an opportunity for an improvement step holds a judgment set $J \in M(t)$ in an arbitrary round t, then we know that J's supporters will strictly decrease in round $t+1$. This means that in the future, potential manipulators may hold judgment J at most a finite number of times. Equivalently, the set of rounds $\{t : J \in M(t)\}$ is finite. This holds for every judgment set in $M(t)$. We conclude that $M(t) \supseteq M(t+1)$ and that the set of rounds $\{t : M(t) = M(t+1)\}$ is finite. □

Our proof of Theorem 5.8 is comparable to the one given by Reyhani and Wilson (2012) in the framework of Voting, and can be related to the more general condition of *set monotonicity* presented by Obraztsova et al. (2015). Along the lines of these authors' reasoning, it should be the case that the sets of potentially winning judgments in every round are characterized by monotonic set inclusion. We present a dual idea above, showing that during the iterative procedure the judgments that can be dropped by an agent in a round in order to manipulate the collective outcome in the next round, decrease in a set-theoretic sense. Hence, in the higher-level, our proof-method coincides with that of Reyhani and Wilson (2012), but is applied in a different set domain.

In Voting Theory, it is shown in the Master's thesis of Reijngoud (2011) that Theorem 5.8 can be generalized for agents who follow any different policy which does not restrict the manner they may manipulate the result, provided that the initial profile is truthful. The corresponding result in the framework of Judgment Aggregation is stated in Theorem 5.9. The proof is omitted, since the translation from Voting is immediate.

Theorem 5.9. *The plurality rule F^{pl} paired with a lexicographic tie-breaking order always converges from the truthful profile when the agents are inertia-friendly, unrestricted, and they have closeness-respecting preferences under full information.*

Most importantly, Theorem 5.10 (proven in the Appendix) identifies a necessary condition for the convergence of the plurality rule, namely that the agents are inertia-friendly.[34]

Theorem 5.10. *The plurality rule F^{pl} together with a lexicographic tie-breaking order may never converge from the truthful or any other profile, when the agents are inertia-averse and they hold full information, independently of the policy they follow.*

5.5.2 Iterative Plurality under Partial Information

Theorem 5.8 states that the plurality rule F^{pl} together with a lexicographic tie-breaking order always converges from the truthful profile when the agents are inertia-friendly,

[34]The corresponding fact in Voting is mentioned — without its proof — by Meir et al. (2010).

outcome-focused, and they have closeness-respecting preferences under full information. We further know from Section 3.7 that the plurality rule is immune to manipulation under zero information. These two results establish that the iterative procedure of the plurality rule always terminates after a finite number of rounds under the above conditions, when the agents are in one of the two extremes of the information-spectrum. Thus, we wonder whether convergence is guaranteed for any type of intermediate information that the agents may hold. Notably, we show that under a very realistic JIF which informs the members of the group only about the current winner in every round, the plurality rule may not converge even for inertia-friendly and outcome-focused agents who initially submit their truthful profile.

Theorem 5.11. *The plurality rule F^{pl} paired with a lexicographic tie-breaking order may never converge from the truthful profile when the agents are inertia-friendly, outcome focused, and they have closeness-respecting preferences under the winner-JIF.*

Proof. Consider the agenda Φ with the judgment sets $J_1, J_2, J_3, J_4 \in \mathcal{J}(\Phi)$ and a group with five agents. The agents 3 and 4 have truthful opinions J_3 and J_4 respectively (while agent 2's truthful judgment is J_2, and agents 1 and 5 both truthfully hold J_1). Suppose moreover that the agents 3 and 4 have closeness-respecting preferences $\succsim_3$ and $\succsim_4$ such that $J_3 \sim_3 J_4 \sim_3 J_2 \succ_3 J_1$ and $J_3 \sim_4 J_4 \sim_4 J_2 \succ_4 J_1$ (by choosing a specific agenda and judgment sets, the preference relations can be easily shown to be closeness-respecting). Suppose that the lexicographic tie-breaking order is $J_1 \succ J_2 \succ J_3 \succ J_4$. Consider the procedure depicted in Table 13.

	J_1	J_2	J_3	J_4	
round 1:	$\underline{2}$	1	1	1	a_3
round 2:	$\underline{2}$	2	0	1	a_4
round 3:	$\underline{2}$	2	1	0	a_3
round 4:	$\underline{2}$	1	2	0	a_4
round 5:	$\underline{2}$	1	1	1	a_3
	...				

Table 13: Iterative plurality procedure under winner-information.

The numbers underneath each judgment set represent the amount of voters that submit it in the specific round. The underlined numbers denote that the respective judgment set wins the round and is the (temporary) collective decision of the group. At the right side of each row we see the agent who makes an improvement step in the specific round. In the first round all the agents submit their truthful judgments and judgment J_1 wins. Agent 3, seeing his least preferred opinion win, makes an improvement step by submitting the judgment set J_2 in the second round, aiming for a scenario where J_2 would need only her support to win, and having nothing to lose otherwise. However, J_1 remains the winner of the second round. Following the same

reasoning, agent 4 untruthfully submits judgment J_3 in the third round. The result does not change, and J_1 still represents the collective outcome. In the fourth round agent 3 hopes that her truthful judgment may now have enough support to win and moves back to it, and agent 4 acts similarly in the fifth round. Hence, a cycle is created. □

By inspecting the proof of Theorem 5.11, an interesting phenomenon may be brought to the reader's attention. A profile where an equilibrium is not reached can still induce a stable collective decision. We omit delving deeper into this kind of situation for the purposes of this thesis, but related questions are open for future research.

5.6 Is Strategic Behavior Socially Profitable?

Within the scope of our work strategic manipulation is a selfish act, employed by members of a group who merely try to obtain a better outcome for themselves. However, while identifying successful insincere behavior becomes more intricate in various dimensions for an individual, a question abounds concerning the social aspects of it.

The general basis of this idea has been introduced in Algorithmic Game Theory by Koutsoupias and Papadimitriou (1999) (see also Christodoulou and Koutsoupias, 2005), who talk about the social damage that strategic behavior may cause in a game. These authors established the famous notion of the *Price of Anarchy*. According to it, the social inefficiency of a game is computed by looking at the ratio between the (sum of the agents' utilities in the) optimal outcome and the (sum of the agents' utilities in the) worst Nash equilibrium that can be reached. Some years later, Brânzei et al. (2013) used a similar intuition in Voting Theory.[35] Referring to elections, the objects of comparison are now the score of the winner in the truthful profile (instead of the optimal outcome of the game) and the score of the winner in all possible equilibria. Considering the greatest of the respective ratios, Brânzei et al. (2013) define the *Dynamic Price of Anarchy*.

In this section we initiate the discussion about *social welfare* in the field of Judgment Aggregation. First, we wonder how good telling the truth actually is for a group of agents *en masse*. The measure that we define and use to this end is the *Price of Truth* (PoT). Then, in order to investigate to what extent strategic thinking is able to improve the outcome in social terms, we define the *Dynamic Price of Anarchy* (DPoA) in Judgment Aggregation.

5.6.1 The Price of Truth

Ideally, truthfulness by the whole group should maximize a kind of utilitarian social welfare, defined relatively to the framework of Judgment Aggregation, or at least approximate such a maximum. We assume that each agent has a *desired judgment*

[35] However, the first to investigate the welfare consequences of strategic behavior in Voting was Lehtinen (2007).

set, i.e., a judgment set that she mostly cares about, which is a subset of her truthful judgment.[36] For agent i holding the truthful judgment J_i, we will denote her desired judgment set as $J_i^\heartsuit$, so that $J_i^\heartsuit \subseteq J_i$. Consider an aggregation rule F and a truthful profile $\boldsymbol{J}$. Having a group of agents $\mathcal{N} = \{1, \ldots, n\}$, the optimal social outcome is achieved when the collective decision maximizes the *proportional* agreement with the agents' desired sets, viz., the *social welfare*. Formally, this happens when $F(\boldsymbol{J}) \in \underset{\boldsymbol{J'} \in \mathcal{J}(\Phi)^n}{\arg\max} \sum_{i \in \mathcal{N}} \frac{|F(\boldsymbol{J'}) \cap J_i^\heartsuit|}{|J_i^\heartsuit|}$. We define the *Price of Truth of F over profile $\boldsymbol{J}$*:[37]

Definition 5.7.

$$PoT(F, \boldsymbol{J}) := \frac{\max\limits_{\boldsymbol{J'} \in \mathcal{J}(\Phi)^n} \sum\limits_{i \in \mathcal{N}} \frac{|F(\boldsymbol{J'}) \cap J_i^\heartsuit|}{|J_i^\heartsuit|}}{\sum\limits_{i \in \mathcal{N}} \frac{|F(\boldsymbol{J}) \cap J_i^\heartsuit|}{|J_i^\heartsuit|}}$$

The *Price of Truth* of an aggregation rule F is its maximum Price of Truth for all possible initial profiles.

Definition 5.8.

$$PoT(F) := \max_{\boldsymbol{J} \in \mathcal{J}(\Phi)^n} PoT(F, \boldsymbol{J})$$

If the PoT of an aggregation rule is 1, then being sincere is always the best option for the group as a whole. The bigger the value of the PoT is, the more damage truthfulness can cause to the group. Example 5.2 highlights a disappointing fact: The PoT can in particular be infinitely high.

Example 5.2. Consider a conjunctive agenda Φ. The PoT of the premise-based procedure F^{pr} when the agents only care about the conclusion (i.e., $J_i^\heartsuit \in \{\{c\}, \{\neg c\}\}$ for every agent i) is infinite. Indeed, there is a truthful profile $\boldsymbol{J}$ where all the agents reject the conclusion, but the rule accepts it (see Table 14). So, $\sum\limits_{i \in \mathcal{N}} |F(\boldsymbol{J}) \cap J_i^\heartsuit| = 0$. △

Example 5.3. The PoT of the majority rule when the agents have Hamming-distance preferences and their desired judgment sets are the same as their truthful judgments is exactly 1 (this follows immediately from the definition of the majority rule). △

Even though we saw that sincerity does not always provide a way to reach the socially optimal outcome, we ponder whether this is the case for some aggregation rules of special interest. Such a rule is plurality, which was recruited in Section 3.7 in order to circumvent the impossibility result of Dietrich and List (2007c), by showing immunity

[36] The idea to consider desired judgment sets is not arbitrary. Some authors have pursued this direction as a principal way to characterize the agents' preferences in Judgment Aggregation. For instance, de Haan (2017) defines the notion of the *subset-based preferences*, according to which an agent ranks the possible collective outcomes with regard to some specific subset of issues that are important to her.

[37] Definition 5.7 can be extended by accounting for different *weights* that an agent may assign to the different formulas in her desired judgment set.

	a_1	a_2	a_3	c
Agent 1:	Yes	Yes	No	**No**
Agent 2:	Yes	No	Yes	**No**
Agent 3:	No	Yes	Yes	**No**
F^{pr}	Yes	Yes	Yes	**Yes**

Table 14: Profile with infinite Price of Truth for the premise-based procedure.

to manipulation under partial information. Is then truthfulness, that we can in this case guarantee, good in social terms? Theorem 5.12 points out that the answer depends on the number of the admissible (i.e., complete and consistent) subsets of the agenda.

Theorem 5.12. *Consider an agenda Φ with $|\mathcal{J}(\Phi)| = k$. If $J_i^{\heartsuit} = J_i$ for every agent i in the group, in every truthful profile $\boldsymbol{J} = (J_1, \ldots, J_n)$, then the PoT of the plurality rule F^{pl}, paired with a lexicographic tie-breaking rule, is $\Theta(k)$.*

Proof. We first construct an instance of a truthful profile $\boldsymbol{J}$ on a specific agenda, for which the PoT is linear in the number of admissible judgment sets. Hence, we establish that the PoT of the plurality rule F^{pl}, along with a lexicographic tie-breaking rule, is at least linear in k. Let $|\Phi| = \{\phi_1, \ldots, \phi_m, \sim\phi_1, \ldots, \sim\phi_m\}$, so $|J| = m$ for every $J \in \mathcal{J}(\Phi)$. To ease the notation, we will write $100\ldots0$ for the judgment set that accepts formula ϕ_1 and rejects all the other formulas, etc. We take an agenda Φ with $m \geqslant k$ and a sufficiently large set of agents $\mathcal{N}$, where the following profile $\boldsymbol{J} := (J_1, \ldots, J_n)$ contains all the possible complete and consistent subsets of Φ.[38,39]

$$
\begin{array}{lll}
J_1 = J_{k+1} & = \ldots = J_{(x-1)k+1} & := 11\ldots1\ldots1 \\
J_2 = J_{k+2} & = \ldots = J_{(x-1)k+2} & := 00\ldots0\ldots0 \\
J_3 = J_{k+3} & = \ldots = J_{(x-1)k+3} & := 10\ldots0\ldots0 \\
J_4 = J_{k+4} & = \ldots = J_{(x-1)k+4} & := 01\ldots0\ldots0 \\
 & \ldots & \\
J_k = J_{2k} & = \ldots = J_{xk} & := 00\ldots1\ldots0
\end{array}
$$

The optimal social welfare for the profile $\boldsymbol{J}$ is obtained when the collective decision is the judgment set $00\ldots0\ldots0$, which can be realized when all the agents submit the

[38] Such an agenda can be constructed by moving to the framework of Judgment Aggregation with integrity constraints (Grandi and Endriss, 2011), considering the set of atomic propositions $\{p_1, \ldots, p_m\}$, imposing the appropriate integrity constraint so that only the relevant ballots are rational, and then moving back to the formula-based Judgment Aggregation framework (Dokow and Holzman, 2009).

[39] Here, to ease the demonstration, we assume that n is divisible by k, but this restriction is not critical; the proof is valid in general, as long as $n > k$.

same judgment J_2. Recall that k is the number of all the alternative judgment sets that the agents submit, and m is their size. Then,

$$\max_{\boldsymbol{J'}\in\mathcal{J}(\Phi)^n} \sum_{i\in\mathcal{N}} \frac{|F(\boldsymbol{J'}) \cap J_i|}{m} = \frac{1}{m}(x\cdot 0 + x\cdot m + x(k-2)\cdot(m-1)) = \frac{x}{m}((k-1)m+2-k)$$

Analogously, we can calculate the welfare produced by the truthful outcome of the plurality rule, assuming that the lexicographic tie-breaking rule ranks J_1 at the top, in which case J_1 will be the outcome.

$$\sum_{i\in\mathcal{N}} \frac{|F(\boldsymbol{J}) \cap J_i|}{m} = \frac{1}{m}(x \cdot m + x \cdot 0 + x(k-2)\cdot 1) = \frac{x}{m}(m+k-2)$$

Thus, the PoT of the plurality rule for profile $\boldsymbol{J}$ is $\frac{(k-1)m+2-k}{m+k-2}$, which is linear with regard to k as $m \geqslant k$. This means that $PoT(F^{pl})$ is $\Omega(k)$.

We will further prove that this bound is tight. We start with calculating an upper bound for the optimal social welfare, considering all possible profiles $\boldsymbol{J}$ in an agenda Φ with $|\mathcal{J}(\Phi)| = k$, such that every admissible profile in $\mathcal{J}(\Phi)$ appears in $\boldsymbol{J}$. In the best-case scenario, we will have that $n-(k-1)$ of the agents have the same truthful judgment, and they get completely satisfied with the collective decision, while the rest $k-1$ agents have different truthful judgments that only disagree with the collective outcome on one formula. This means that

$$\max_{\boldsymbol{J'}\in\mathcal{J}(\Phi)^n} \sum_{i\in\mathcal{N}} \frac{|F(\boldsymbol{J'}) \cap J_i|}{m} \leqslant \frac{1}{m}((n-k+1)\cdot m + (k-1)\cdot(m-1)) = \frac{1}{m}(nm-k+1)$$

Afterwards, we calculate a lower bound for the social welfare obtained by the truthful profile of the agents, again considering all possible profiles $\boldsymbol{J}$ such that every admissible profile in $\mathcal{J}(\Phi)$ appears in $\boldsymbol{J}$. The winning judgment set of the plurality rule should be truthfully held by at least $\frac{n}{k}$ of the agents. These agents get fully happy with the outcome. However, in the worst case scenario all but $k-2$ of the agents truthfully support an opinion that has no formula in common with the winning judgment. The remaining $k-2$ agents have to truthfully hold different judgments, which in the worst case only share one formula with the collective decision. That is

$$\sum_{i\in\mathcal{N}} \frac{|F(\boldsymbol{J}) \cap J_i|}{m} \geqslant \frac{1}{m}(\frac{n}{k} \cdot m + 0 + (k-2)\cdot 1) = \frac{1}{m}(\frac{nm}{k} + k - 2)$$

In total, the PoT of the plurality rule for any profile $\boldsymbol{J}$ where the agents submit all k different complete and consistent subsets of the agenda Φ can be at most $\frac{knm-k^2+k}{nm+k^2-2k}$, which is $O(k)$. In a completely analogous manner we can show that the PoT for any profile $\boldsymbol{J}$ where the agents submit less than k different judgments is still at most $O(k)$.

We conclude that the PoT of the plurality rule is $\Theta(k)$. □

5.6.2 The Dynamic Price of Anarchy

Next, we wonder whether encouraging the iteration of an aggregation rule that is guaranteed to converge can make a significant part of the agents "happier". By drawing connections between the relevant work in Voting (Brânzei et al., 2013) and Game Theory (Koutsoupias and Papadimitriou, 1999), we propose an interpretation of the *Price of Anarchy* notion in our framework, based on the agents' preferences. In order to compare our findings with the corresponding ones regarding the Price of Truth, we will restrict our attention to iterative procedures that start from an initial truthful profile. A bit more formally, having an aggregation rule and an initial truthful profile, we will take into account all the iterations where at least one improvement step takes place, and for each one of these iterations we will calculate the proportion of the formulas in the agents' desired judgment sets that agree with the outcome of the achieved equilibrium profile. The minimum for all possible equilibria will constitute the social welfare value of the iteration procedure for the given initial profile. Consider an aggregation rule F, a truthful profile $\boldsymbol{J}$, and agents i with desired judgment sets $J_i^{\heartsuit}$. We define the *Dynamic Price of Anarchy of F over profile $\boldsymbol{J}$*:[40,41]

Definition 5.9.

$$DPoA(F, \boldsymbol{J}) := \frac{\max\limits_{\boldsymbol{J'} \in \mathcal{J}(\Phi)^n} \sum\limits_{i \in \mathcal{N}} \frac{|F(\boldsymbol{J'}) \cap J_i^{\heartsuit}|}{|J_i^{\heartsuit}|}}{\min\limits_{\boldsymbol{J''} \in EQ} \sum\limits_{i \in \mathcal{N}} \frac{|F(\boldsymbol{J''}) \cap J_i^{\heartsuit}|}{|J_i^{\heartsuit}|}},$$

where EQ is the set of all equilibrium profiles $\boldsymbol{J''}$, such that there is a path of improvements steps (of length at least one) starting from profile $\boldsymbol{J}$ and reaching $\boldsymbol{J''}$.[42]

Considering the maximum value of the DPoA of a rule over all initial truthful profiles, we define its *Dynamic Price of Anarchy*.

$$DPoA(F) := \max_{\boldsymbol{J} \in \mathcal{J}(\Phi)^n} DPoA(F, \boldsymbol{J})$$

The closer to 1 the DpoA of an aggregation rule is, the more socially profitable this rule is in iteration.

The Premise-based Procedure

We exemplify Definition 5.9 by focusing on the premise-based procedure. We consider conjunctive agendas and agents who have conclusion-oriented preferences (see

[40] The DPoA depends on more parameters, such as the agents' preferences, their information, their inertia-type and their policies, which are omitted to lighten the notation.

[41] In general, a realistic constraint on the type of the preferences of the agents should be imposed concerning their desired judgment sets; caring only about a specific subset of their truthful judgment should be reflected by the agents' preferences. We don't include this restriction in our definition, but it will be taken care of in the results of this section.

[42] The DPoA over a profile $\boldsymbol{J}$ is defined only when the aggregation rule provides the agents with some improvement step, because only then our question about the benefits of iteration is meaningful.

Sections 3.6 and 5.2 for a motivation). Since an agent i in this scenario only cares about the conclusion c in the agenda, we take her desired judgment set $J_i^{\heartsuit}$ to be either $\{c\}$ or $\{\neg c\}$, depending on whether she accepts or rejects the conclusion respectively. Suppose that $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ is the truthful profile of the group. Then, the DPoA of the premise-based procedure F^{pr} will be:

$$DPoA(F^{pr}) = \max_{\boldsymbol{J} \in \mathcal{J}(\Phi)^n} \frac{\max\limits_{\boldsymbol{J}' \in \mathcal{J}(\Phi)^n} \sum\limits_{i \in \mathcal{N}} \mathbb{1}_{c \in (J_i \cap F(\boldsymbol{J}')) \cup (\overline{J_i} \cap \overline{F(\boldsymbol{J}')})}}{\min\limits_{\boldsymbol{J}'' \in EQ} \sum\limits_{i \in \mathcal{N}} \mathbb{1}_{c \in (J_i \cap F(\boldsymbol{J}'')) \cup (\overline{J_i} \cap \overline{F(\boldsymbol{J}'')})}},$$

where EQ is defined as before.

Theorem 5.13. *The DPoA of the premise-based procedure F^{pr} for fully informed, conclusion-oriented agents, that follow any policy and are of any inertia type, is* 1.

Proof. By Theorem 5.2, every non-trivial iteration of the premise-based procedure on a conjunctive agenda converges after exactly one round under full information. Moreover, when the initial profile is truthful (that we assume here), the only improvement step can be made by an agent who rejects the conclusion and sees that at least half of her peers also reject it. In the end, the result on the conclusion will have to agree with the majority's desired sets at least, and this is the optimal collective decision. □

For the case where agents hold zero information, we provide an analogous result in Theorem 5.14.

Theorem 5.14. *The DPoA of the premise-based procedure F^{pr} for zero-informed, conclusion-oriented agents, that follow any policy and are of any inertia type, is* 1.

Proof. We give an outline of the proof. It must be the case that at least half of the agents truthfully accept the conclusion, or at least half of them truthfully reject it. Moreover, the dominant option of the agents who accept the conclusion under zero information on a conjunctive agenda is to be truthful, and the dominant option of the agents who reject the conclusion is to reject all the premises (Lemma 5.6). Hence, once a convergence state is reached, either the majority of the agents will accept all the premises, making the conclusion accepted, or the majority of the agents will reject them, making the conclusion rejected. In either case, at least half of the agents will satisfy their desired judgment set, and this is the optimal collective decision. □

If providing an aggregation rule F we can show that $PoT > DPoA$, it means that strategic behavior benefits the group more than sincerity when F is applied. Theorems 5.13 and 5.14 (which confirm the efficiency of the premise-based procedure in iteration), together with Example 5.2 (according to which the Price of Truth of the premise-based procedure can be infinite), imply the following corollaries.

Corollary 5.15. *Take a conjunctive agenda Φ, any conclusion-oriented agents, and any truthful profile $\boldsymbol{J}$ which induces at least one round of iteration. Then, the PoT of the premise-based procedure F^{pr} for $\boldsymbol{J}$ is strictly higher than the DPoA of F^{pr} for $\boldsymbol{J}$.*

Corollary 5.16. *Take a conjunctive agenda Φ and conclusion-oriented agents. Then, the PoT of the premise-based procedure F^{pr} is strictly higher than the DPoA of F^{pr}.*

Recall the conclusion-based procedure defined in the Background, where the collective decision on a conjunctive (or disjunctive) agenda only concerns the conclusion and is taken using the majority rule on it. Obviously, the conclusion-based procedure is non-manipulable and returns the optimal decision for conclusion-oriented individuals. Theorems 5.13 and 5.14 imply that despite the differences of the premise-based and the conclusion-based procedure, the collective judgment on the conclusion produced by those two rules in iteration is the same.[43] Dietrich and List (2007c) have already noticed this effect in a slightly different context, and they call the two procedures *strategically equivalent*. Our framework, however, provides more refined insights into the precise influence of the agents' information and preferences with regard to the strategies they choose to apply, and as a consequence it can distinguish between the various equilibria that may be reached.

The Plurality Rule

First, we demonstrate in Theorem 5.17 that for agendas with only three admissible subsets, the iteration of the plurality rule is beneficial for the group as a whole.

Theorem 5.17. *Consider an agenda Φ with $\mathcal{J}(\Phi) = 3$, a sufficiently large group of fully-informed agents, and the plurality rule F^{pl} together with a lexicographic tie-breaking order. If for any truthful profile $\boldsymbol{J}$ it holds that $J_i^{\heartsuit} = J_i$ for all agents i, the agents have Hamming-distance preferences, they follow any policy and they are of any inertia-type, then the DPoA of F^{pl} for $\boldsymbol{J}$ is strictly smaller the PoT of F^{pl} for $\boldsymbol{J}$.*

Proof. Let $\mathcal{J}(\Phi) = \{J_1, J_2, J_3\}$ and $|J_1| = |J_2| = |J_3| = m$. Denote $|J_1 \cap J_2| = x_{12}$, $|J_1 \cap J_3| = x_{13}$ and $|J_2 \cap J_3| = x_{23}$. Since we only look at truthful profiles $\boldsymbol{J}$ where at least one agent has an opportunity to perform an improvement step, the following two cases partition the set of all truthful profiles that are of interest to us.

Case 1: There are (at least) two judgment sets, say J_1, J_2 without loss of generality, that are submitted by the same number of agents σn, where σ is a positive rational number. Moreover, the tie-breaking rule selects J_1, while an agent i who truthfully submits J_3 prefers J_2 to J_1. Then, we have that:

[43] Note, though, that in order to obtain a non-trivial iteration of the premise-based procedure under full information, there has to be a significant alignment of the agents' judgments with regard to some premise.

$$PoT(F, \boldsymbol{J}) = \frac{\max\limits_{\boldsymbol{J'} \in \mathcal{J}(\Phi)^n} \sum\limits_{i \in \mathcal{N}} |F(\boldsymbol{J'}) \cap J_i|}{\sigma n \cdot m + \sigma n \cdot x_{12} + (n - 2\sigma n) \cdot x_{13}}.$$

After agent i makes her improvement step, J_2 becomes the collective outcome according to plurality, and the procedure terminates. We have that:

$$DPoA(F, \boldsymbol{J}) = \frac{\max\limits_{\boldsymbol{J'} \in \mathcal{J}(\Phi)^n} \sum\limits_{i \in \mathcal{N}} |F(\boldsymbol{J'}) \cap J_i|}{\sigma n \cdot x_{12} + \sigma n \cdot m + (n - 2\sigma n) \cdot x_{23}}.$$

As agent i truthfully holds the judgment J_3 and having Hamming-distance preferences prefers J_2 to J_1, it holds that $x_{13} < x_{23}$, so $DPoA(F, \boldsymbol{J}) < PoT(F, \boldsymbol{J})$.

Case 2: The winning judgment set is J_1 without loss of generality, having been submitted by σn agents, where σ is a positive rational number. Moreover, profile J_2 is truthfully held by $\sigma n - 1$ agents and the tie-breaking rule ranks J_2 above J_1, while an agent i who truthfully submits J_3 prefers J_2 to J_1. Then, we have that:

$$PoT(F, \boldsymbol{J}) = \frac{\max\limits_{\boldsymbol{J'} \in \mathcal{J}(\Phi)^n} \sum\limits_{i \in \mathcal{N}} |F(\boldsymbol{J'}) \cap J_i|}{\sigma n \cdot m + (\sigma n - 1) \cdot x_{12} + (n - 2\sigma n + 1) \cdot x_{13}}.$$

After agent i makes her improvement step, J_2 becomes the collective outcome according to plurality, and the procedure terminates. We have that:

$$DPoA(F, \boldsymbol{J}) = \frac{\max\limits_{\boldsymbol{J'} \in \mathcal{J}(\Phi)^n} \sum\limits_{i \in \mathcal{N}} |F(\boldsymbol{J'}) \cap J_i|}{\sigma n \cdot x_{12} + (\sigma n - 1) \cdot m + (n - 2\sigma n + 1) \cdot x_{23}}.$$

Since agent i truthfully holds the judgment J_3 and having Hamming-distance preferences prefers J_2 to J_1, it holds that $x_{13} < x_{23}$. Moreover, we have that $\sigma < \frac{1}{2}$, because there is at least one agent who holds each one of the three judgment sets and the numbers of supporters of the first and the second judgment differ only for one agent (hence it cannot be the case that half of the agents or more submit the first judgment). We have that:

$$\frac{DPoA(F, \boldsymbol{J})}{PoT(F, \boldsymbol{J})} = \frac{\sigma n \cdot m + (\sigma n - 1) \cdot x_{12} + (n - 2\sigma n + 1) \cdot x_{13}}{\sigma n \cdot x_{12} + (\sigma n - 1) \cdot m + (n - 2\sigma n + 1) \cdot x_{23}},$$

which as n increases is

$$\frac{\sigma \cdot m + \sigma \cdot x_{12} + (1 - 2\sigma) \cdot x_{13}}{\sigma \cdot x_{12} + \sigma \cdot m + (1 - 2\sigma) \cdot x_{23}} < 1.$$

□

However, we know that the PoT is already small enough for agendas with only three admissible subsets (Theorem 5.12). So, we now wish to investigate the relation between the PoT and the DPoA for a more general class of agendas. We settle in Theorem 5.18 that the DPoA of the plurality rule, similarly to the PoT, is linear in the number of the admissible subsets of the agenda. This means that iteration is as inefficient as truth in the worst case scenario. However, even if the worst case scenaria for the two cases coincide, for specific profiles on richer agendas, iteration can cause greater social damage than sticking to the truth, independently of how large the number of the agents is (Theorem 5.19), and *vice versa* (Theorem 5.20). All the proofs are given in the Appendix.

Theorem 5.18. *Consider an agenda Φ with $|\mathcal{J}(\Phi)| = k > 3$. If the agents are fully informed and they have Hamming-distance preferences, and if $J_i^{\heartsuit} = J_i$ for every agent i in every truthful profile $\boldsymbol{J} = (J_1, \ldots, J_n)$, then the DPoA of the plurality rule F^{pl}, paired with a lexicographic tie-breaking rule, is $\Theta(k)$.*

Theorem 5.19. *Consider an agenda Φ with $|\mathcal{J}(\Phi)| = k > 3$. If the agents are of any inertia-type, fully informed, they have Hamming-distance preferences and they follow any policy, and if $J_i^{\heartsuit} = J_i$ for every agent i in every truthful profile $\boldsymbol{J} = (J_1, \ldots, J_n)$, then there is a profile $\boldsymbol{J}$ such that the DPoA of the plurality rule F^{pl} for $\boldsymbol{J}$, paired with a lexicographic tie-breaking rule, is strictly higher than the PoT of F^{pl} for $\boldsymbol{J}$.*

Theorem 5.20. *Consider an agenda Φ with $|\mathcal{J}(\Phi)| = k > 3$. If the agents are of any inertia-type, fully informed, they have Hamming-distance preferences and they follow any policy, and if $J_i^{\heartsuit} = J_i$ for every agent i in every truthful profile $\boldsymbol{J} = (J_1, \ldots, J_n)$, then there is a profile $\boldsymbol{J}$ such that the DPoA of the plurality rule F^{pl} for $\boldsymbol{J}$, paired with a lexicographic tie-breaking rule, is strictly smaller than the PoT of F^{pl} for $\boldsymbol{J}$.*

5.7 Concluding Remarks

Several aspects, both of conceptual and of practical nature, of the agents' strategic behavior in the framework of Judgment Aggregation have been explored in this chapter. We pondered whether the use of manipulable rules in iteration is a reasonable concept, meaning first that it can promise the achievement of an equilibrium state, and second that this equilibrium can lead to a beneficial outcome for the group. The answer goes hand in hand with the specific aggregation rules in question, as well as with other contextual parameters, such as the individual preferences, information, and desires. Intriguingly, we can remark that most of the times a trade-off is required, between asking for truthfulness and approaching the optimal social welfare. In the end, the choice of a suitable procedure directly depends on our desiderata.

That being said, there are many directions that future research could expand upon. First, we shall notice that all our examples of non- or slow convergence were based on the construction of ad-hoc agendas and groups of agents. Nevertheless, we believe

that an extended study of the connections between particular agenda-structures and the convergence of the iteration of aggregation rules could be very fruitful. Moreover, in this chapter we restricted our work to conclusion-oriented and Hamming-distance preferences, in order to obtain unequivocally realistic results for the premise-based procedure and the plurality rule respectively. A generalization to a larger class of preferences could be a next step (note that our negative results immediately hold for any wider class of preferences; the positive ones do not). Of course, one could investigate the behavior in iteration of more aggregation rules too. Due to the complexity of this kind of analytical work, we would suggest the use of simulation methods, which are broadly applied for similar purposes in Voting Theory (e.g., Koolyk et al., 2016; Meir et al., 2014; Reijngoud and Endriss, 2012). Finally, the framework we developed deals only with myopic and memory-less agents that change their judgments sequentially. All these assumptions can be dropped in various ways. For example, it would be natural to study situations where the agents are able to plan long-term strategies, or where they may use the data collected by previous rounds to compute their next move (for the latter, an appropriate method that is highly appreciated in the field of Machine Learning, but also more and more exercised in Game Theory, Economics, and the area of Multiagent Systems, is *reinforcement learning*. See for instance Jaksch et al., 2010; Nowé et al., 2012; Sutton and Barto, 1998). In addition, instead of restricting our focus to agents who move one by one, we could allow for simultaneous changes (this design has been exploited in Voting Theory by Meir, 2015).

Chapter 6

Conclusion

In this chapter we review the main accomplishments of this thesis (Section 6.1), and we discuss several ideas for future work (Section 6.2).

6.1 Summary of the Results

The goal of this thesis was to unravel the formal assumptions that govern individual strategic behavior in collective decision making, and specifically in Judgment Aggregation. To that end, we extended the basic framework of the literature in three ways.

First, contrary to the reductionist approach that has been followed so far in Judgment Aggregation, there are numerous situations where agents do not know everything about the opinions of all the other members of their group. For example, consider the reviewers of a scientific article. They may be asked by the editor of a journal to report their judgments on different criteria regarding the article's quality, without knowing the exact judgments of the other reviewers, whose opinions are going to be aggregated with their own in order for the editor to decide whether to accept the article for publication or not. To make the informational status of the agents in different aggregation procedures explicit, we developed a model of *partial information* in Judgment Aggregation. Based on that model, we studied various aggregation rules and we were able to identify cases where some individual had a reason to report an insincere opinion aiming for a more desirable collective choice for her, that is, cases of *manipulation*. We showed that when the uncertainty of an agent regarding the opinions of her peers grows larger, her incentives for manipulation never increase, and often decrease. In that sense, withholding information from a group can only favor strategy-proofness. Combining the different types of information with the different types of preferences that the agents may hold, we also conducted a detailed analysis of the manipulability of the *premise-based procedure*: a method of aggregation that plays a central role among political scientists, in discussions about deliberative democracy. Ultimately, the main achievement of our model was to single out an aggregation rule that overcomes the impossibility result of Dietrich and List (2007c), being non-dictatorial and

immune to manipulation under partial information, viz., the *plurality rule*.

Second, even though overlooked by social choice theorists, *interactive reasoning* is a principal element of every strategic situation that concerns intelligent agents. In a Judgment Aggregation setting, consider again the reviewers of an article, whose aggregated opinions are common knowledge and suggest rejection. Suppose that one of the reviewers has a personal interest to get the article accepted and is willing to lie in order to achieve her goal. What if another reviewer, that is better off by being honest if and only if everyone else is also honest, figures out that a manipulation act is possible? Then, a new incentive for manipulation will be created, triggered by an agent's reasoning about the reasoning of others. In order to account for this kind of situation, we designed a model of *higher-level reasoning*, tailored to the framework of Judgment Aggregation. From our study, it became evident that not only what the agents know about the opinions of their peers, but also their information about the preferences of the others, crucially affect their strategic behavior. We showed that if non-sophisticated agents do not have incentives to manipulate an aggregation procedure, then being sincere is what makes more sophisticated agents better-off too. However, our main result was the following: If a non-sophisticated agent has an incentive to manipulate, then for any arbitrary level, there exists a higher level of interactive reasoning such that an agent who engages in it will also have a reason to manipulate.

Third, agents often participate in aggregating procedures that run for more than one round. To capture these situations, we built a model of *Iterative Judgment Aggregation*. Accounting for various motivations regarding the agents' strategic behavior, we extensively investigated two aggregation rules that have been fundamental in our work, namely the *plurality rule* and the *premise-based procedure*. Our analysis revolved around the convergence of an iterative process, that is, the reach of a terminal state where no agent wishes to change her submitted opinion. We proved that having agents who are not "extremely truth-biased" (meaning that they may choose to remain untruthful if they cannot influence the collective decision in some round) is a necessary condition for the convergence of the plurality rule. As far as the premise-based procedure is concerned, convergence is always guaranteed, sooner or later, depending on the amount of information that the agents hold, and their submitted opinions in the first round. Finally, by allowing the agents to behave strategically sequentially, we were able to measure the social effects of telling the truth on the one hand, and of acting strategically on the other. We did so by defining and comparing two values: the Price of Truth, and the Dynamic Price of Anarchy respectively. In the plurality rule, in cases where the members of a group had a large number m of judgments available to submit, both truthfulness and strategic behavior could be detrimental for the social welfare; in particular, the Price of Truth and the Dynamic Price of Anarchy could be linear in m. However, for the premise-based procedure our results indicated undeniably the benefits of strategic behavior, that was showed to always ensure the optimal social outcome. This last observation could provide an insightful argument to the supporters of the premise-based procedure in Political Science debates.

6.2 Future Research

We believe that the most intriguing feature of our work is its intersection with a number of different research areas, ranging from Philosophy and the Social Sciences to Artificial Intelligence. Hence, besides the possible topical extensions that have been discussed at the end of each chapter of this thesis separately, we aspire to a future work that builds more fundamental connections between the various disciplines that peruse agents' strategic reasoning and (collective) behavior. We give some suggestions in the following lines.

The framework developed in this thesis was adjusted to the specific needs of Judgment Aggregation, and this choice was made in order to facilitate our reasoning in the targeted examples and proofs. Nonetheless, looking for modelling attempts that are close to ours, the first candidates are found in the fields of *Game Theory* and *(Dynamic) Epistemic Logic*. Now that a first step is made towards the understanding of strategic behavior in Judgment Aggregation, a formal embedding of our framework into the existing and well-established models of those two areas can be very fruitful for the study of collective decision making.

Furthermore, in different parts of this thesis we have referred to the *experimental work* concerning the investigation of the agents' actual behavior in strategic situations, but, to the best of our knowledge, there does not exist related research focused wholly on Judgment Aggregation. Besides the resources that behavioral results (would) provide to test our formal models, the benefits between the two approaches (the theoretical and the experimental one) could also work *vice versa*. Thanks to the rigorous mathematical tools employed in this thesis, numerous — previously hidden — assumptions that directly influence human and artificial agents' reasoning are now made explicit, and are waiting for further behavioral analysis.

Last but not least, all our models could be investigated from a *computational complexity* point of view. We conjecture that the richer the environment that an agent has to take into account when she reasons strategically is (e.g., when there is higher uncertainty about the truthful opinions and the possible manipulations of her peers), the harder it is for the agent to decide whether it is worth being dishonest or not.

Appendix

Proof of Theorem 2.3. We only show the non-trivial direction, which says that if an aggregation rule F is strategy-proof for all reflexive, transitive and complete closeness-respecting preferences, then F is independent and monotonic.

We will need an intermediate result by Dietrich and List (2007c). In particular, these authors define a preference-free notion of strategy-proofness. They say that a rule F is (preference-free) manipulable at the profile of judgments $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ by agent i if there is a formula ϕ in the agenda Φ such that $F(J_i, \boldsymbol{J}_{-i})$ disagrees with J_i on ϕ, but $F(J_i^*, \boldsymbol{J}_{-i})$ agrees with J_i on ϕ, for some untruthful opinion J_i^*. Based on this definition, Dietrich and List prove that every aggregation rule is immune to (preference-free) manipulation if and only if it is independent and monotonic.

Back to our proof, it now suffices to demonstrate that if an aggregation rule F is strategy-proof for all reflexive, transitive and complete closeness-respecting preferences, then F is immune to (preference-free) manipulation. So, assume that strategy-proofness is the case. To show immunity to (preference-free) manipulation, consider a formula $\phi \in \Phi$, an agent $i \in \mathcal{N}$, and a truthful profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ such that $F(J_i, \boldsymbol{J}_{-i})$ disagrees with J_i on ϕ. We need to prove that $F(J_i^*, \boldsymbol{J}_{-i})$ still disagrees with J_i on ϕ, for every dishonest judgment J_i^*. We define a preference relation $\succsim_i$ over all possible collective outcomes such that $J \succsim_i J'$ if and only if J_i agrees on ϕ with J but not with J', or it agrees on ϕ with both, or it disagrees with both (intuitively, this would be the case if agent i only cares about the issue ϕ in the agenda). It is easy to verify that $\succsim_i$ is reflexive, transitive, complete, and closeness-respecting. Hence, by strategy-proofness it will be $F(J_i, \boldsymbol{J}_{-i}) \succsim_i F(J_i^*, \boldsymbol{J}_{-i})$ for all untruthful judgments J_i^*. But as $F(J_i, \boldsymbol{J}_{-i})$ disagrees with J_i on ϕ, the definition of $\succsim_i$ implies that $F(J_i^*, \boldsymbol{J}_{-i})$ also disagrees with J_i on ϕ, for every dishonest judgment J_i^*. □

Proof of Lemma 3.6. Consider a JIF π that is at least as informative as another JIF σ and an aggregation rule F that is σ-manipulable. We will show that F is also π-manipulable. By σ-manipulability we know that there are an agent $i \in \mathcal{N}$, a profile $\boldsymbol{J} = (J_1, \ldots, J_n) \in \mathcal{J}(\Phi)^n$, a preference relation $\succsim_i \in PR(J_i)$ and an insincere opinion J_i^* such that $F(J_i^*, \boldsymbol{J'}_{-i}) \succ_i F(J_i, \boldsymbol{J'}_{-i})$, for some $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{1,\sigma,\boldsymbol{J}}$, and $F(J_i^*, \boldsymbol{J''}_{-i}) \succsim_i F(J_i, \boldsymbol{J''}_{-i})$, for all other $\boldsymbol{J''}_{-i} \in \mathcal{W}_i^{\sigma_i(\boldsymbol{J})}$. Now, consider the profile $\boldsymbol{J'} = (J_i, \boldsymbol{J'}_{-i})$. We will show that if the truthful profile is $\boldsymbol{J'}$, then agent i has an incentive to manipulate by reporting the untruthful judgment set J_i^*. It suffices to

show two things: (a) that $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ (which holds because reflexivity (REF) of $\mathcal{W}$ implies that $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{1,\pi,(J_i,\boldsymbol{J'}_{-i})}$) and (b) that for all other partial profiles $\boldsymbol{J''}_{-i}$ such that $\boldsymbol{J''}_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J'}}$, it also holds that $\boldsymbol{J''}_{-i} \in \mathcal{W}_i^{1,\sigma,\boldsymbol{J}}$, i.e., $\mathcal{W}_i^{1,\pi,\boldsymbol{J'}} \subseteq \mathcal{W}_i^{1,\sigma,\boldsymbol{J}}$. For (b), by symmetry (SYM) and transitivity (TRANS) of $\mathcal{W}$, we have that $\mathcal{W}_i^{1,\pi,\boldsymbol{J}} = \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$ and $\mathcal{W}_i^{1,\sigma,\boldsymbol{J}} = \mathcal{W}_i^{1,\sigma,\boldsymbol{J'}}$, and since π is at least as informative as σ, it is the case that $\mathcal{W}_i^{1,\pi,\boldsymbol{J}} \subseteq \mathcal{W}_i^{1,\sigma,\boldsymbol{J'}}$. □

Proof of Theorem 3.9. From Theorem 3.2 we have that if an aggregation rule F is independent and monotonic, then F is π-strategy-proof for all JIFs π, so it is also π-strategy-proof for all JIFs π that respect C with regard to F. Hence, we only need to show the left to right direction, which says that if F is π-strategy-proof for the class C, then F is independent and monotonic. We show the contrapositive. That is, we prove that if F is not both independent and monotonic, then it is π-manipulable for C. Suppose that F is not both independent and monotonic. Then, F is manipulable for C (Dietrich and List, 2007c). This means that there are an agent i, a profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$, a judgment set J_i^* and a closeness-respecting preference relation $\succsim_i$ such that $F(J_i^*, \boldsymbol{J}_{-i}) \succ_i F(J_i, \boldsymbol{J}_{-i})$. As π respects C with regard to F, together with completeness of $\succsim_i$, this implies that $F(J_i^*, \boldsymbol{J'}_{-i}) \succsim_i F(J_i, \boldsymbol{J'}_{-i})$ for all partial profiles $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$. Hence, F is π-manipulable. □

Proof of Theorem 3.12. To illustrate the idea of the proof, consider a conjunctive agenda with two premises a_1, a_2 and a conclusion $c \leftrightarrow (a_1 \wedge a_2)$. For the one direction, consider the profile depicted in Table 15 and suppose that for agent 2 it is $w_2^{a_1} < w_2^c$.

	a_1	a_2	$c \leftrightarrow (a_1 \wedge a_2)$
Agent 1:	Yes	Yes	Yes
Agent 2:	Yes	No	No
Agent 3:	No	Yes	No
F^{pr}:	Yes	Yes	Yes

Table 15: Incentive for agent 2 to manipulate if $w_2^{a_1} < w_2^c$.

In this scenario, agent 2 initially accepts proposition a_1 (and the group agrees with her), but she rejects the conclusion c (and the group disagrees with her). Considering the judgments of agents 1 and 3, Agent 2 has the option to be dishonest on proposition a_1 and manipulate the rule to obtain a rejection of the conclusion. Hence, agent 2 can achieve a collective judgment that agrees with her on the conclusion and disagrees with her on premise a_1. This collective judgment will be closer to her truthful judgment by the value $w_2^c - w_2^{a_1} > 0$, so she strictly prefers it as the result. This means that agent 2 has an incentive to lie and the premise-based procedure is susceptible to manipulation. For the other direction, it is easy to see following the intuition of the

proof of Theorem 3.11 that if the agent cares about each premise at least as much as she does about the conclusion, then she can never be better off by lying. □

Proof of Theorem 3.17. Consider a set of three agents $\mathcal{N} := \{1, 2, 3\}$ and an agenda $\Phi := \{\phi_1, \phi_2, \phi_3, \neg\phi_1, \neg\phi_2, \neg\phi_3\}$ such that $\mathcal{J}(\Phi) = \{\{\phi_1, \phi_2, \neg\phi_3\}, \{\phi_1, \neg\phi_2, \neg\phi_3\}, \{\phi_1, \neg\phi_2, \phi_3\}, \{\neg\phi_1, \phi_2, \phi_3\}\}$.[44] Moreover, consider the tie-breaking rule with the linear order $\{\neg\phi_1, \phi_2, \phi_3\} > \{\phi_1, \phi_2, \neg\phi_3\} > \{\phi_1, \neg\phi_2, \phi_3\} > \{\phi_1, \neg\phi_2, \neg\phi_3\}$. Then, take agent $i := 1$, with truthful judgment $J_1 := \{\phi_1, \phi_2, \neg\phi_3\}$ and closeness-respecting preference relation $\gtrsim_1$ such that $\{\phi_1, \phi_2, \neg\phi_3\} \sim_1 \{\phi_1, \neg\phi_2, \neg\phi_3\} \sim_1 \{\phi_1, \neg\phi_2, \phi_3\} >_1 \{\neg\phi_1, \phi_2, \phi_3\}$. Now, consider the truthful profile $\boldsymbol{J} = (J_1, J_2, J_3)$ depicted in Table 16. We will show that there is an insincere judgment $J_1^* \in \mathcal{J}(\Phi)$ that gives agent 1 an incentive to manipulate the aggregation procedure. Consider the profile $\boldsymbol{J}^* = (J_1^*, J_2, J_3)$, depicted in Table 17.

	ϕ_1	ϕ_2	ϕ_3
Agent 1:	Yes	Yes	No
Agent 2:	No	Yes	Yes
Agent 3:	Yes	No	Yes
F^{av}	No	Yes	Yes

Table 16: Profile (J_1, J_2, J_3)

	ϕ_1	ϕ_2	ϕ_3
Agent 1:	Yes	No	No
Agent 2:	No	Yes	Yes
Agent 3:	Yes	No	Yes
F^{av}	Yes	No	Yes

Table 17: Profile (J_1^*, J_2, J_3)

We can see that $F^{av}(J_1^*, J_2, J_3) >_1 F^{av}(J_1, J_2, J_3)$. Moreover, we claim that $F^{av}(J_1^*, J_2', J_3') \gtrsim_1 F^{av}(J_1, J_2', J_3')$, for all other partial profiles $(J_2', J_3') \in \mathcal{J}(\Phi)^2$. We will prove this claim by contradiction. Recalling that preference $\gtrsim_1$ is complete, suppose that there is some partial profile $(J_2', J_3') \in \mathcal{J}(\Phi)^2$ such that $F^{av}(J_1, J_2', J_3') >_1 F^{av}(J_1^*, J_2', J_3')$, so $F^{av}(J_1^*, J_2', J_3') = \{\neg\phi_1, \phi_2, \phi_3\}$. Let us call L the judgment set $\{\neg\phi_1, \phi_2, \phi_3\}$. Then, $F^{av}(J_1^*, J_2', J_3') = L$ and $F^{av}(J_1, J_2', J_3') = L' \neq L$, for some $L' \in \mathcal{J}(\Phi)$. One of the following two cases holds:

Case 1: $H(L, (J_1^*, J_2', J_3')) = H(L', (J_1^*, J_2', J_3'))$ and the ties are broken in favor of L. Then, for any $L' \in \mathcal{J}(\Phi)$, making the calculations we have that $3 + H(L, J_2') + H(L, J_3') \leqslant 1 + H(L', J_2') + H(J_1, J_3')$, or equivalently $2 + H(L, J_2') + H(L, J_3') \leqslant H(L', J_2') + H(L', J_3')$. This implies that $H(L, (J_1, J_2', J_3')) \leqslant H(L', (J_1, J_2', J_3'))$, and hence $F^{av}(J_1, J_2', J_3') \neq L'$, which is a contradiction.

Case 2: $H(L, (J_i^*, J_2', J_3')) < H(L', (J_i^*, J_2', J_3'))$. Analogous to case 1. □

[44] Such an agenda can be constructed by moving to the framework of Judgment Aggregation with integrity constraints (Grandi and Endriss, 2011), considering the set of atomic propositions $\{p, q, r\}$, imposing the appropriate integrity constraint so that only the ballots $\{(110), (100), (011), (101)\}$ are rational, and then moving back to the formula-based Judgment Aggregation framework (Dokow and Holzman, 2009). See the Background for more details.

Proof of Theorem 4.6. Consider an agenda Φ with $|\mathcal{J}(\Phi)| \geqslant 3$.

(a) Take the judgment sets $J_1, J_2, J_3 \in \mathcal{J}(\Phi)$. Consider agent 1 with truthful opinion J_1 and preference relation $\succsim_1 \in C(J_1)$, such that $J_1 \sim_1 J_2 \succ_1 J_3$, without loss of generality. Imagine an aggregation scenario where agent 1 is informed that the winning collective decision of the truthful profile is J_3. Then, as she has nothing to lose, she will try to manipulate the result by dishonestly reporting J_2, hoping that her vote is pivotal.

(b) Take an arbitrary agent i, a truthful profile $\boldsymbol{J} = (J_i, \boldsymbol{J}_{-i})$ with $F^{pl}(\boldsymbol{J}) = J$, and a closeness-respecting preference $\succsim_i \in C(J_i)$. Suppose that there is a judgment set J_i^* such that $F^{pl}(J_i^*, \boldsymbol{J'}_{-i}) \succ_i F^{pl}(J_i, \boldsymbol{J'}_{-i})$, for some partial profile $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{2,\text{winner-JIF},\boldsymbol{J}}$. By definition of the closeness-respecting preferences and the plurality rule, this can happen only if the the manipulated result is the judgment set J_i^* and $J_i^* \succ_i J$. By definition of closeness-respecting preferences, we have that there is a formula $\phi \in \Phi$ such that $\phi \in J_i \cap J_i^*$ and $\phi \notin J$. Fix this formula ϕ and imagine that agent i reasons as follows. Since she does not know what each agent's truthful opinion is, it is possible for her that agent 2 sincerely holds judgment J_i^*. Moreover, it is also possible for her that agent 2 only cares about proposition ϕ in her truthful judgment, so she holds a closeness-respecting preference relation $\succsim_2$ such that $J_i \succ_2 J$. But it is common knowledge that judgment J is the collective decision on the truthful profile. Hence, agent 2 who, according to agent 1, engages in fist-level reasoning, may try to manipulate the result and be better off by dishonestly reporting J_i, because she has nothing to lose. In case J_i was pivotal in the truthful profile, this manipulation can indeed make it win. On the other hand, if agent i tries to manipulate too, then she will miss the opportunity to see her truthful opinion wining, and she will be worse off. We conclude that it is risky for agent i to manipulate, so she will avoid doing so. □

Proof of Lemma 4.7. Suppose that $\mathcal{W}_i^{1,\pi,\boldsymbol{J}} = \{\boldsymbol{J}^1_{-i}, \ldots, \boldsymbol{J}^r_{-i}\}$. Consider the possible case where every agent j holds the closeness-respecting preference relation $\succsim'_j$ such that $J \sim'_j J'$ for all $J, J' \in \mathcal{J}(\Phi)$. By Definition 3.5, since the agents do not have an incentive to manipulate when they have those preferences, we have that for all $v \in \{1, \ldots, r\}$

$$S_j^F(\mathcal{W}_j^{k-1,\pi,(J_i,\boldsymbol{J}^v_{-i})}, \succsim'_j, J_j^v) = \{J_j^v\}$$

Now, recall Definition 4.1. We have that

$$\mathcal{W}_i^{k,\pi,\boldsymbol{J}} := \bigcup_{v\in\{1,\ldots,r\}} \bigcup_{(\succsim_1,\ldots,\succsim_n):\succsim_j\in C\ \forall j} \widetilde{\mathcal{W}}_i^{k,\pi,\boldsymbol{J}}(\boldsymbol{J}^v_{-i}, (\succsim_1, \ldots, \succsim_n))$$

where

$$\widetilde{\mathcal{W}}_i^{k,\pi,\boldsymbol{J}}(\boldsymbol{J}_{-i}^v,(\succsim_1,\ldots,\succsim_n)) := \bigtimes_{j\neq i} S_j^F(\mathcal{W}_j^{k-1,\pi,(J_i,\boldsymbol{J}_{-i}^v)},\succsim_j,J_j^v)$$

Hence, it will be

$$\mathcal{W}_i^{k,\pi,\boldsymbol{J}} \supseteq \bigcup_{v\in\{1,\ldots,r\}} \widetilde{\mathcal{W}}_i^{k,\pi,\boldsymbol{J}}(\boldsymbol{J}_{-i}^v,(\succsim_1',\ldots,\succsim_n'))$$

which means that

$$\mathcal{W}_i^{k,\pi,\boldsymbol{J}} \supseteq \bigcup_{v\in\{1,\ldots,r\}} \bigtimes_{j\neq i} \{J_j^v\}$$

or equivalently

$$\mathcal{W}_i^{k,\pi,\boldsymbol{J}} \supseteq \bigcup_{v\in\{1,\ldots,r\}} \boldsymbol{J}_{-i}^v = \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$$

□

Proof of Theorem 4.8. Since the rule F is zero-manipulable under first-order reasoning, there is a first-order reasoner i, which considers all scenaria possible to be the truthful ones, i.e., $\mathcal{W}_i^{1,\pi,\boldsymbol{J}} = \mathcal{J}(\Phi)^{n-1}$, and she has an incentive to manipulate. We know by Lemma 4.7 that $\mathcal{W}_i^{1,\pi\boldsymbol{J}} \subseteq \mathcal{W}_i^{2,\pi,\boldsymbol{J}}$, so it will be $\mathcal{J}(\Phi)^{n-1} = \mathcal{W}_i^{1,\pi\boldsymbol{J}} \subseteq \mathcal{W}_i^{2,\pi,\boldsymbol{J}} \subseteq \mathcal{J}(\Phi)^{n-1}$, which means that $\mathcal{W}_i^{2,\pi,\boldsymbol{J}} = \mathcal{W}_i^{1,\pi,\boldsymbol{J}} = \mathcal{J}(\Phi)^{n-1}$. We conclude that agent i will have an incentive to manipulate under second-order reasoning too. □

Proof of Theorem 4.11. We give a proof by induction.

- **Induction Basis.** We have that F is immune to π-manipulation under first-level reasoning, by the hypothesis.
- **Induction Hypothesis.** Suppose that F is immune to π-manipulation under level-$(k-1)$ reasoning.
- **Induction Step.** We will show that F is immune to π-manipulation under level-k reasoning. Consider an arbitrary agent $i \in \mathcal{N}$ and a profile $\boldsymbol{J} \in \mathcal{J}(\Phi)$. The set of partial profiles that agent i considers possible after engaging in level-k reasoning when the actual profile is $\boldsymbol{J}$ is $\mathcal{W}_i^{k,\pi,\boldsymbol{J}}$, as defined in Definition 4.4. But since F is immune to π-manipulation under reasoning of level $k-1$, it is the case that no agent has an incentive to lie under level-$(k-1)$ reasoning. In other words this means that the only best strategy of each agent in every possible scenario is her truthful strategy. Specifically, for all $\boldsymbol{J}'_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$, all agents j

and all preference relations $\succsim_j$, it will be $S_j^F(\mathcal{W}_j^{k-1,\pi,(J_i,\boldsymbol{J'}_{-i})}, \succsim_j, J'_j) = \{J'_j\}$. Then, Definition 4.4 implies that $\mathcal{W}_i^{\pi_i^k(\boldsymbol{J})} = \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$, as follows:

$$\mathcal{W}_i^{k,\pi,\boldsymbol{J}} = \bigcup_{v\in\{1,\dots,r\}} \bigcup_{(\succsim_1,\dots,\succsim_n):\succsim_j\in C\ \forall j} \bigtimes_{j\neq i} S_j^F(\mathcal{W}_j^{k-1,\pi,(J_i,\boldsymbol{J}^v_{-i})}, \succsim_j, J_j^v)$$

$$= \bigcup_{v\in\{1,\dots,r\}} \bigcup_{(\succsim_1,\dots,\succsim_n):\succsim_j\in C\ \forall j} \bigtimes_{j\neq i} \{J_j^v\}$$

$$= \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$$

Thus, since agent i does not have an incentive to manipulate under level-1 reasoning (we know that by the hypothesis), she will not have an incentive to manipulate under level-k reasoning either. □

Proof of Theorem 4.12. Suppose that F is susceptible to π-manipulation under first-level reasoning and immune to π-manipulation under level-k reasoning for some k. We will show that F is susceptible to π-manipulation under level-$(k+1)$ reasoning. Since F is susceptible to π-manipulation under level-1 reasoning, there is an agent $i \in \mathcal{N}$ and a profile $\boldsymbol{J} \in \mathcal{J}(\Phi)$ such that agent i, holding the information $\mathcal{W}_i^{\pi_i^1(\boldsymbol{J})}$ after level-1 reasoning, has an incentive to manipulate. Now, the set of partial profiles that agent i considers possible after engaging in level-$(k+1)$ reasoning, when the truthful profile is $\boldsymbol{J}$, is $\mathcal{W}_i^{k,\pi,\boldsymbol{J}}$, as defined in Definition 4.4. But since F is immune to π-manipulation under level-k reasoning, it is the case that no agent has an incentive to lie under level-k reasoning. In other words, this means that the only best strategy of each agent in every possible scenario is her truthful strategy. Specifically, for all $\boldsymbol{J'}_{-i} \in \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$, all agents j and all preference relations $\succsim_j$, we have that $S_j^F(\mathcal{W}_j^{k,\pi,(J_i,\boldsymbol{J'}_{-i})}, \succsim_j, J'_j) = \{J'_j\}$. Then, Definition 4.4 implies that $\mathcal{W}_i^{k+1,\pi,\boldsymbol{J}} = \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$, as follows:

$$\mathcal{W}_i^{k+1,\pi,\boldsymbol{J}} = \bigcup_{v\in\{1,\dots,r\}} \bigcup_{(\succsim_1,\dots,\succsim_n):\succsim_j\in C\ \forall j} \bigtimes_{j\neq i} S_j^F(\mathcal{W}_j^{k,\pi,(J_i,\boldsymbol{J}^v_{-i})}, \succsim_j, J_j^v)$$

$$= \bigcup_{v\in\{1,\dots,r\}} \bigcup_{(\succsim_1,\dots,\succsim_n):\succsim_j\in C\ \forall j} \bigtimes_{j\neq i} \{J_j^v\}$$

$$= \mathcal{W}_i^{1,\pi,\boldsymbol{J}}$$

Hence, since agent i has an incentive to manipulate under level-1 reasoning, she will have an incentive to manipulate under level-$(k+1)$ reasoning too. □

Proof of Lemma 5.3. First, we need to show that if the agent is sincere in round t, then she has no opportunity to make an improvement step. Indeed, if the conclusion is already accepted by the group, this is the most desirable result for agent i. Otherwise, a rejected conclusion by the group means that some premise is rejected by the majority of the agents. But since agent i is already truthfully accepting all the premises, she could not give extra support to the rejected premise in order to make it accepted, hence there is no improvement step she could make.

So, suppose that agent i is insincere in round t. If the conclusion is already accepted, then she only has an opportunity for an improvement step if she is inertia-averse and she moves to her truthful opinion; hence, the proof follows. Now, aiming for a contradiction, suppose that the group rejects the conclusion in round t, and agent i can turn it into being accepted by making a best improvement step using $J_i^* \neq J_i$, which accepts some proper subset of the premises. As the group will accept the conclusion in round $t+1$, the majority will have to accept all the premises. But if agent i reported her truthful opinion, adding some support to the rest of the premises she was rejecting by submitting J_i^*, it would still be the case that the majority accepts all the premises, and thus, the conclusion too. This means that agent i's truthful judgment induces the same result as her insincere judgment J_i^*. But then, since agent i is truth-biased, J_i^* cannot be used as a best improvement step, which is a contradiction. □

Proof of Theorem 5.5. Call A the set of agents who accept the conclusion and R the set of agents who reject it. By Lemma 5.3, in any round t, an agent $i \in A$ has an opportunity to make a best improvement step if and only if she moves to her truthful opinion from an insincere one. This move can be realized at most once for every agent in A. On the other hand, an agent $j \in R$ has an opportunity to perform an improvement step in round t if (a') she can lie (choosing a judgment according to her policy) in order to make a previously accepted conclusion be rejected by the group or if (b') she can move to being truthful without causing a less desirable collective conclusion. But if the conclusion is collectively rejected, only an agent $i \in A$ can turn it to be accepted again in the future, by submitting her truthful judgment that she was not submitting before. This can happen at most $|A|$ times. Hence, agents in R may perform an improvement step using condition (a') at most $|A|$ times. In addition, they may make an improvement step using condition (b') at most $|R| + |A|$ times (we have to add $|A|$ because maybe an agent who has already returned to her truthful judgment manipulates again to "fix" a new move of an agent in A and later she returns back to the truth). In total, no-one will have an available opportunity to perform an improvement step after at most $|A| + |A| + |R| + |A| = 2|A| + n \leqslant 3n$ rounds. □

Proof of Lemma 5.6. Suppose that agent i chooses to reject some proper subset X of the premises. Take an arbitrary premise $a_i^* \notin X$. Then, we can construct a possible partial profile where not lying on a_i^* is dominated by lying on a_i^*; this is the case when no manipulation on the premises in X is able to flip the collective deci-

sion on the outcome, but manipulation on a_i^* can do so. See below an example with three agents and three premises, where agent 1 is asked to (re-)submit a judgment in round t. Agent 1 chooses to reject premises a_2 and a_3, but not premise a_1. The result is depicted in Table 18. However, in that possible scenario, had the agent chosen to reject premise a_1 too, she would obtain collective rejection of the conclusion (Table 19), which is preferable to her. The example can be easily generalized. □

	a_1	a_2	a_3	c
Agent 1:	**Yes**	No	No	No
Agent 2:	No	Yes	Yes	Yes
Agent 3:	Yes	Yes	Yes	Yes
F^{pr}	Yes	Yes	Yes	**Yes**

Table 18: Agent 1 chooses to lie on premise a_2, but the conclusion is still collectively accepted.

	a_1	a_2	a_3	c
Agent 1:	**No**	No	No	No
Agent 2:	No	Yes	Yes	Yes
Agent 3:	Yes	Yes	Yes	Yes
F^{pr}	No	Yes	Yes	**No**

Table 19: Agent 1 chooses to lie on premise a_1 too, and this makes the conclusion collectively rejected.

Proof of Theorem 5.10. We show the case where agents are outcome-focused (the other cases are analogous). Consider the agenda Φ with judgment sets $J_1, J_2, J_3, J_4 \in \mathcal{J}(\Phi)$, and two agents 3 and 4 with truthful opinions J_3 and J_4 respectively. Suppose moreover that the agents have closeness-respecting preferences $\gtrsim_3$ and $\gtrsim_4$ such that $J_3 \sim_3 J_4 \sim_3 J_1 \succ_3 J_2$ and $J_3 \sim_4 J_4 \sim_4 J_2 \succ_4 J_1$ (by constructing a specific agenda and judgment sets, the preference relations can be shown to indeed be closeness-respecting; we omit this part of the proof to ease the demonstration). Suppose that $n = 6$ and that the lexicographic tie-breaking order is $J_1 \succ J_2 \succ J_3 \succ J_4$. Consider the procedure depicted in Table 20. The numbers underneath each judgment

	J_1	J_2	J_3	J_4	
round 1:	$\underline{2}$	2	1	1	a_4
round 2:	2	$\underline{3}$	1	0	a_3
round 3:	$\underline{3}$	3	0	0	a_4
round 4:	$\underline{3}$	2	0	1	a_3
round 5:	$\underline{2}$	2	1	1	a_4
	...				

Table 20: Iterative plurality procedure with inertia-averse agents.

set represent the amount of voters that submit it in the specific round. The underlined numbers denote that the respective judgment set wins the round and is the (temporary) collective decision of the group. At the right side of each row we see the agent who makes an improvement step in the specific round. In the first round all the agents

submit their truthful judgments and judgment J_1 wins. Agent 4 manipulates making her preferred judgment set J_2 win in the second round, while afterwards agent 3 untruthfully submits judgment J_1 and makes it the winner of the third round. In the fourth round agent 4 cannot influence the result anymore and therefore she prefers to be truthful, and the same holds for agent 3 in the fifth round. The profile of the fifth round is the same as the profile of the first round and hence a cycle is created. □

Proof of Theorem 5.18. We first construct an instance of a truthful profile $\boldsymbol{J}$ on a specific agenda where the DPoA is linear with respect to the admissible judgment sets, hence we establish that the DPoA of the plurality rule F^{pl}, together with a lexicographic tie-breaking order, is at least linear in k. Let $\Phi = \{\phi_1, \ldots, \phi_m, \sim\phi_1, \ldots, \sim\phi_m\}$, so $|J| = m$ for every $J \in \mathcal{J}(\Phi)$. To ease the notation, we write $100\ldots0$ for the judgment set that accepts formula ϕ_1 and rejects all the other formulas, etc. We take an agenda Φ with $m \geqslant k$ and a sufficiently big set of agents $\mathcal{N}$, where the following profile $\boldsymbol{J} := (J_1, \ldots, J_n)$ contains all the possible admissible subsets of Φ.

$$
\begin{array}{lll}
J_1 = J_{k+1} & = \ldots = J_{(x-1)k+1} & := 11111\ldots1\ldots1111 \\
J_2 = J_{k+2} & = \ldots = J_{(x-1)k+2} & := 01111\ldots1\ldots1111 \\
J_3 = J_{k+3} & = \ldots = J_{(x-1)k+3} & := 11111\ldots1\ldots1100 \\
J_4 = J_{k+4} & = \ldots = J_{(x-1)k+4} & := 11111\ldots1\ldots1000 \\
J_5 = J_{k+5} & = \ldots = J_{(x-1)k+5} & := 00000\ldots0\ldots0000 \\
J_6 = J_{k+6} & = \ldots = J_{(x-1)k+6} & := 10000\ldots0\ldots0000 \\
J_7 = J_{k+7} & = \ldots = J_{(x-1)k+7} & := 01000\ldots0\ldots0000 \\
J_8 = J_{k+8} & = \ldots = J_{(x-1)k+8} & := 00100\ldots0\ldots0000 \\
 & \ldots & \\
J_k = J_{2k} & = \ldots = J_{xk} & := 00000\ldots1\ldots0000
\end{array}
$$

The optimal social welfare for the profile $\boldsymbol{J}$ is achieved when all the agents submit the judgment set J_5, so the collective decision is $00000\ldots0\ldots0000$. Then,

$$
\begin{aligned}
\max_{\boldsymbol{J'} \in \mathcal{J}(\Phi)^n} \sum_{i \in \mathcal{N}} |F(\boldsymbol{J'}) \cap J_i| &= x \cdot 0 + x \cdot 1 + x \cdot 2 + x \cdot 3 + x(k-5) \cdot (m-1) + x \cdot m \\
&= x((k-4)m + 11 - k)
\end{aligned}
$$

Notice that all the judgment sets receive equal numbers of supporters, and suppose that the tie-breaking ranks J_i above J_j for every $i < j$, so it makes J_1 win. Now consider the following iterative procedure. At the beginning, an agent who truthfully holds J_4 switches to J_3 (which is preferable to her according to Hamming-distance preferences). In round 2, an agent whose truthful opinion is J_1 sees that J_3 wins and moves to J_2 that she prefers. Now both J_2 and J_3 have two more supporters than all the other judgment sets, so no other judgment can become the collective decision in the future by a unilateral deviation of an agent. It is easy to see that more agents prefer

J_2 to J_3, so the iteration will terminate with J_3 winning. We can calculate the welfare produced by this equilibrium.

$$\begin{aligned}\sum_{i\in\mathcal{N}} |J_3 \cap J_i| &= x\cdot(m-2) + x\cdot(m-3) + x\cdot m \\ &\quad + x\cdot(m-1) + x\cdot 2 + x(k-5)\cdot 3 \\ &= x(4m+3k-19)\end{aligned}$$

Thus, the DPoA of plurality rule for profile $\boldsymbol{J}$ is $\frac{(k-4)m+11-k}{4m+3k-19}$, which is linear with regard to k as $m \geqslant k$. So, the DPoA of the plurality rule is $\Omega(k)$.

We will further prove that this bound is tight. From this point the proof proceeds identically to the proof of Theorem 5.12. There, we calculated an upper bound for the optimal social welfare, considering all possible profiles $\boldsymbol{J}$ in an agenda Φ with $|\mathcal{J}(\Phi)| = k$. Afterwards, we found a lower bound for the social welfare obtained by the truthful profile of the agents, again considering all possible profiles $\boldsymbol{J}$. The crucial observation is that in order to compute that lower bound we took into account *all* possible profiles, including those that could be an equilibrium state of an iteration procedure. Hence, we can use the same lower bound for the social welfare obtained by the iteration of the plurality rule. In total, the Dynamic Price of Anarchy of the plurality rule for any profile $\boldsymbol{J}$ can be at most $\frac{knm-k^2+k}{nm-k^2-2k}$, which is $O(k)$.

We conclude that the DPoA of the plurality rule is $\Theta(k)$. □

Proof of Theorem 5.19. To illustrate the idea of the proof, we give an example of an agenda Φ with four complete and consistent subsets and four outcome-focused agents. Consider the profile $\boldsymbol{J} := (J_1, J_2, J_3, J_4)$, where $J_1 := 0000, J_2 := 0001, J_3 := 1110, J_4 := 1100$, and assume that the tie-breaking rule selects J_1, which is also the socially optimal outcome. This means that $PoT(F^{pl}, \boldsymbol{J}) = 1$. However, assume that an iteration procedure takes place, where at first an agent switches from J_4 to J_3 (which she prefers with regard to the current collective decision J_1). Then, the agent who truthfully holds J_1 moves to J_2 that is more desirable for her than J_3, and the procedure terminates with J_2 winning. It is straightforward to measure that $DPoA(F^{pl}, \boldsymbol{J}) \geqslant \frac{5}{4} > PoT(F^{pl}, \boldsymbol{J})$. □

Proof of Theorem 5.20. We sketch the proof using an agenda Φ with five complete and consistent subsets and five outcome-focused agents. Consider the profile $\boldsymbol{J} := (J_1, J_2, J_3, J_4, J_5)$, where $J_1 := 111, J_2 := 001, J_3 := 010, J_4 := 100, J_5 := 000$, and assume that the tie-breaking rule selects J_1. By following the definitions, we compute that $PoT(F^{pl}, \boldsymbol{J}) = \frac{9}{6}$. However, if an iteration procedure takes place, in worst case scenario we have an equilibrium where one of the judgment sets J_2, J_3, J_4 wins. Then, $DPoA(F^{pl}, \boldsymbol{J}) = \frac{9}{8} < PoT(F^{pl}, \boldsymbol{J})$. □

List of Symbols

$\succsim_i$ The preference relation of agent i over all possible collective decisions. It is taken to be reflexive, transitive, and complete 12

2^{Φ} The set of all subsets (the powerset) of the agenda 6

C Closeness-respecting preferences: among two possible opinions, they prioritize the one whose intersection with the agent's truthful judgment is a superset of the other's intersection with the agent's truthful judgment 13

Φ The agenda, i.e., the set of formulas that the agents decide upon 5

H Hamming-distance preferences: the smaller the Hamming-distance of an opinion to the agent's truthful judgment is, the higher that opinion is ranked 14

$H(J, J')$ The Hamming-distance between the judgment sets J and J'. That is, the number of formulas that they disagree on 8

$H(\boldsymbol{J}, J)$ The Hamming-distance between the profile $\boldsymbol{J}$ and the judgment set J'. That is, the sum of the number of formulas that all the individual judgment sets in $\boldsymbol{J}$ and J disagree on 9

H_w Weighted Hamming-distance preferences 31

$\boldsymbol{J}$ The profile of judgments of all the agents 6

J_i The truthful judgment set (opinion) of agent i. It is complete and consistent 5

$J_i^{\heartsuit}$ The desired judgment set of agent i. It is a subset of the truthful judgment J_i . 70

$\boldsymbol{J}_{-i}$ The partial profile of judgments of all the agents except for agent i 6

Bibliography

Airiau, S. and Endriss, U. (2009). Iterated Majority Voting. In *Proceedings of the 1st International Conference on Algorithmic Decision Theory (ADT)*, pages 38–49.

Aleskerov, F., Karabekyan, D., Sanver, M. R., and Yakuba, V. (2011). On the Degree of Manipulability of Multi-valued Social Choice Rules. *Homo Oeconomicus*, 28(1):205–216.

Arad, A. and Rubinstein, A. (2012). The 11–20 Money Request Game: A Level-k Reasoning Study. *The American Economic Review*, 102(7):3561–3573.

Armstrong, D. M. (1973). *Belief, Truth and Knowledge*. Cambridge University Press.

Arrow, K. J. (1951a). Alternative Approaches to the Theory of Choice in Risk-taking Situations. *Econometrica*, 19(4):404–437.

Arrow, K. J. (1951b). *Social Choice and Individual Values*. Yale University Press.

Bacharach, M., Gerard-Varet, L. A., Mongin, P., and Shin, H. S. (2012). *Epistemic Logic and the Theory of Games and Decisions*. Springer.

Bartholdi, J. J., Tovey, C. A., and Trick, M. A. (1989). The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241.

Bassi, A. (2015). Voting Systems and Strategic Manipulation: An Experimental Study. *Journal of Theoretical Politics*, 27(1):58–85.

Baumeister, D., Erdélyi, G., and Rothe, J. (2016). Judgment Aggregation. In Rothe, J., editor, *Economics and Computation: An Introduction to Algorithmic Game Theory, Computational Social Choice, and Fair Division*, pages 361–391. Springer.

Baumeister, D., Rothe, J., and Selker, A.-K. (2017). Strategic Behavior in Judgment Aggregation. In Endriss, U., editor, *Trends in Computational Social Choice*. AI Access. To appear.

Botan, S., Novaro, A., and Endriss, U. (2016). Group Manipulation in Judgment Aggregation. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 411–419.

Brânzei, S., Caragiannis, I., Morgenstern, J., and Procaccia, A. D. (2013). How Bad is Selfish Voting? In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI)*, pages 138–144.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2004). A Cognitive Hierarchy Model of Games. *The Quarterly Journal of Economics*, 119(3):861–898.

Caragiannis, I., Procaccia, A. D., and Shah, N. (2014). Modal Ranking: A uniquely Robust Voting Rule. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI)*, pages 616–622.

Chapman, B. (2002). Rational Aggregation. *Politics, Philosophy and Economics*, 1(3):337–354.

Chopra, S., Pacuit, E., and Parikh, R. (2004). Knowledge-theoretic Properties of Strategic Voting. In *Proceedings of the 8th European Conference on Logics in Artificial Intelligence (JELIA)*, pages 18–30.

Christodoulou, G. and Koutsoupias, E. (2005). The Price of Anarchy of Finite Congestion Games. In *Proceedings of the 37th annual ACM Symposium on Theory of Computing (STOC)*, pages 67–73.

Costa-Gomes, M. and Crawford, V. P. (2006). Cognition and Behavior in two-person Guessing Games: An Experimental Study. *The American Economic Review*, 96(5):1737–1768.

Costa-Gomes, M., Crawford, V. P., and Broseta, B. (2001). Cognition and Behavior in Normal-form Games: An Experimental Study. *Econometrica*, 69(5):1193–1235.

de Haan, R. (2017). Complexity Results for Manipulation, Bribery and Control of the Kemeny Procedure in Judgment Aggregation. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1151–1159.

Dietrich, F. (2007). A Generalised Model of Judgment Aggregation. *Social Choice and Welfare*, 28(4):529–565.

Dietrich, F. and List, C. (2007a). Arrow's Theorem in Judgment Aggregation. *Social Choice and Welfare*, 29(1):19–33.

Dietrich, F. and List, C. (2007b). Judgment Aggregation by Quota Rules: Majority Voting Generalized. *Journal of Theoretical Politics*, 19(4):391–424.

Dietrich, F. and List, C. (2007c). Strategy-Proof Judgment Aggregation. *Economics and Philosophy*, 23(03):269–300.

Dietrich, F. and List, C. (2008). Judgment Aggregation without Full Rationality. *Social Choice and Welfare*, 31(1):15–39.

Dietrich, F. and List, C. (2010). Majority Voting on Restricted Domains. *Journal of Economic Theory*, 145(2):512–543.

Dietrich, F. and List, C. (2013). Where Do Preferences Come From? *International Journal of Game Theory*, 42(3):613–637.

Dokow, E. and Holzman, R. (2009). Aggregation of Binary Evaluations for Truth-functional Agendas. *Social Choice and Welfare*, 32(2):221–241.

Dokow, E. and Holzman, R. (2010). Aggregation of Binary Evaluations. *Journal of Economic Theory*, 145(2):495–511.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669.

Elster, J. (1998). *Deliberative Democracy*. Cambridge University Press.

Endriss, U. (2016). Judgment Aggregation. In Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D., editors, *Handbook of Computational Social Choice*, pages 399–426. Cambridge University Press.

Endriss, U. and de Haan, R. (2015). Complexity of the Winner Determination Problem in Judgment Aggregation: Kemeny, Slater, Tideman, Young. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 117–125.

Endriss, U. and Grandi, U. (2014). Binary Aggregation by Selection of the Most Representative Voters. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI)*, pages 668–674.

Endriss, U., Grandi, U., de Haan, R., and Lang, J. (2016a). Succinctness of Languages for Judgment Aggregation. In *Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 176–186.

Endriss, U., Grandi, U., and Porello, D. (2010). Complexity of Judgment Aggregation: Safety of the Agenda. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 359–366.

Endriss, U., Grandi, U., and Porello, D. (2012). Complexity of Judgment Aggregation. *Journal of Artificial Intelligence Research*, 45(1):481–514.

Endriss, U., Obraztsova, S., Polukarov, M., and Rosenschein, J. S. (2016b). Strategic Voting with Incomplete Information. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 236–242.

Everaere, P., Konieczny, S., and Marquis, P. (2015). Belief Merging versus Judgment Aggregation. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 999–1007.

Farquharson, R. (1969). *Theory of Voting*. Yale University Press.

Gibbard, A. (1973). Manipulation of Voting Schemes: A General Result. *Econometrica*, 41(4):587–601.

Gilboa, I. (2009). *Theory of Decision under Uncertainty*. Cambridge University Press.

Grandi, U. (2012). *Binary Aggregation with Integrity Constraints*. PhD thesis, Institute for Logic, Language and Computation, University of Amsterdam.

Grandi, U. and Endriss, U. (2011). Binary Aggregation with Integrity Constraints. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 204–209.

Grandi, U., Loreggia, A., Rossi, F., Venable, K. B., and Walsh, T. (2013). Restricted Manipulation in Iterative Voting: Condorcet Efficiency and Borda Score. In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory (ADT)*, pages 181–192.

Griffioen, S. (2017). Covertly Controlling Choices: Manipulating Decision Making Under Partial Knowledge. Master's thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam.

Halpern, J. Y. (2005). *Reasoning about Uncertainty*. MIT press.

Hemaspaandra, E., Spakowski, H., and Vogel, J. (2005). The Complexity of Kemeny Elections. *Theoretical Computer Science*, 349(3):382–391.

Hendricks, V. F. (2006). *Mainstream and Formal Epistemology*. Cambridge University Press.

Hintikka, J. (1962). *Knowledge and Belief*. Cornell University Press.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11:1563–1600.

Jeffrey, R. (1983). Bayesianism with a Human Face. *Testing Scientific Theories*, 10:133–156.

Kemeny, J. G. (1959). Mathematics without Numbers. *Daedalus*, 88(4):577–591.

Knight, F. H. (1921). Risk, Uncertainty and Profit. *Houghton Mifflin Company*.

Koolyk, A., Lev, O., and Rosenschein, J. S. (2016). Convergence and Quality of Iterative Voting under non-scoring Rules. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1329–1330.

Kornhauser, L. A. and Sager, L. G. (1986). Unpacking the court. *The Yale Law Journal*, 96(1):82–117.

Koutsoupias, E. and Papadimitriou, C. (1999). Worst-case Equilibria. In *Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science (STACS)*, pages 404–413.

Lehtinen, A. (2007). The Welfare Consequences of Strategic Voting in Two Commonly Used Parliamentary Agendas. *Theory and Decision*, 63(1):1–40.

Lev, O. and Rosenschein, J. S. (2012). Convergence of Iterative Voting. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 611–618.

List, C. (2003). A Possibility Theorem on Aggregation over Multiple Interconnected Propositions. *Mathematical Social Sciences*, 45(1):1–13.

List, C. (2012). The Theory of Judgment Aggregation: An Introductory Review. *Synthese*, 187(1):179–207.

List, C. and Pettit, P. (2002). Aggregating Sets of Judgments: An Impossibility Result. *Economics and Philosophy*, 18(1):89–110.

List, C. and Pettit, P. (2004). Aggregating Sets of Judgments: Two Impossibility Results Compared. *Synthese*, 140(1):207–235.

Meir, R. (2015). Plurality Voting under Uncertainty. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2103–2109.

Meir, R. (2017). Iterative Voting. In Endriss, U., editor, *Trends in Computational Social Choice*. AI Access. To appear.

Meir, R., Lev, O., and Rosenschein, J. S. (2014). A Local-dominance Theory of Voting Equilibria. In *Proceedings of the 15th ACM Conference on Economics and Computation (EC)*, pages 313–330.

Meir, R., Polukarov, M., Rosenschein, J. S., and Jennings, N. R. (2010). Convergence to Equilibria in Plurality Voting. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI)*, pages 823–828.

Miller, D. (1992). Deliberative Democracy and Social Choice. *Political Studies*, 40(1):54–67.

Miller, M. K. and Osherson, D. (2009). Methods for Distance-based Judgment Aggregation. *Social Choice and Welfare*, 32(4):575–601.

Mongin, P. (2012). The Doctrinal Paradox, the Discursive Dilemma, and Logical Aggregation Theory. *Theory and Decision*, 73(3):1–41.

Nagel, R. (1995). Unraveling in Guessing Games: An Experimental Study. *The American Economic Review*, 85(5):1313–1326.

Nehring, K. and Puppe, C. (2007). The Structure of Strategy-proof Social Choice—Part I: General Characterization and Possibility Results on Median Spaces. *Journal of Economic Theory*, 135(1):269–305.

Nowé, A., Vrancx, P., and De Hauwere, Y.-M. (2012). Game Theory and Multi-agent Reinforcement Learning. In Wiering, M. and van Otterlo, M., editors, *Reinforcement Learning*, pages 441–470. Springer.

Obraztsova, S., Markakis, E., Polukarov, M., Rabinovich, Z., and Jennings, N. R. (2015). On the Convergence of Iterative Voting: How Restrictive Should Restricted Dynamics Be? In *Proceedings of the 29th AAAI Conference on Artificial Intelligence(AAAI)*, pages 993–999.

Obraztsova, S., Markakis, E., and Thompson, D. R. (2013). Plurality Voting with Truth-biased Agents. In *Proceedings of the 6th International Symposium on Algorithmic Game Theory (SAGT)*, pages 26–37.

Pauly, M. and Van Hees, M. (2006). Logical Constraints on Judgement Aggregation. *Journal of Philosophical Logic*, 35(6):569–585.

Penczynski, S. P. (2016). Strategic Thinking: The Influence of the Game. *Journal of Economic Behavior and Organization*, 128:72–84.

Perea, A. (2012). *Epistemic Game Theory: Reasoning and Choice*. Cambridge University Press.

Pettit, P. (2001). Deliberative Democracy and the Discursive Dilemma. *Philosophical Issues*, 11(1):268–299.

Pigozzi, G. (2006). Belief Merging and the Discursive Dilemma: An Argument-based Account to Paradoxes of Judgment Aggregation. *Synthese*, 152(2):285–298.

Ray, P. (1973). Independence of Irrelevant Alternatives. *Econometrica*, 41(5):987–991.

Reijngoud, A. (2011). Voter Response to Iterated Poll Information. Master's thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam.

Reijngoud, A. and Endriss, U. (2012). Voter Response to Iterated Poll Information. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 635–644.

Reyhani, R. and Wilson, M. C. (2012). Best Reply Dynamics for Scoring Rules. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI)*, pages 672–677.

Satterthwaite, M. A. (1975). Strategy-proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions. *Journal of Economic Theory*, 10(2):187–217.

Sen, A. (1969). Quasi-transitivity, Rational Choice and Collective Decisions. *The Review of Economic Studies*, 36(3):381–393.

Sen, A. (1970). *Collective Choice and Social Welfare*. San Francisco, CA: Holden–Day.

Sen, A. (1973). Behaviour and the Concept of Preference. *Economica*, 40(159):241–259.

Stahl, D. O. and Wilson, P. W. (1995). On Players' Models of other Players: Theory and Experimental Evidence. *Games and Economic Behavior*, 10(1):218–254.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Van Ditmarsch, H., Lang, J., and Saffidine, A. (2012). Strategic Voting and the Logic of Knowledge. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1247–1248.

Van Hees, M. (2007). The Limits of Epistemic Democracy. *Social Choice and Welfare*, 28(4):649–666.

Von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

Zwicker, W. (2016). Introduction to the Theory of Voting. In Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D., editors, *Handbook of Computational Social Choice*, pages 23–56. Cambridge University Press.