# Strategic manipulation in voting under higher-order reasoning

**MSc Thesis** *(Afstudeerscriptie)*

written by

**Kyah Elisabeth Mercedes Smaal**
(born July 15th, 1992 in Nijmegen, the Netherlands)

under the supervision of **dr. Ronald de Haan** and **dr. Fernando R. Velázquez Quesada**, 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:**
*February 7, 2019*

**Members of the Thesis Committee:**
dr. Alexandru Baltag
dr. Ulle Endriss
dr. Ronald de Haan
Zoi Terzopoulou MSc
dr. Fernando R. Velázquez Quesada
prof. dr. Yde Venema (chair)

Institute for Logic, Language and Computation

# Abstract

The Gibbard-Satterthwaite Theorem states that any non-dictatorial and surjective social choice function is susceptible to manipulation if there are at least three alternatives. This classical result assumes that manipulators are naive: they think that every other voter will cast a sincere ballot. Furthermore, it is assumed that voters have full information regarding the preferences of other voters. These assumptions make it unrealistically easy to manipulate an election. We argue that voters are likely to realise that other voters may act strategically too, and choose the best strategy accordingly. This thesis investigates the strategic incentives of higher-order reasoning voters, that is, voters who reflect on the uncertainty about the uncertainty of other voters, and so on. We develop a dynamic epistemic model for strategic voting and use this model to analyse strategic behaviour of higher-order reasoning voters. In the traditional 'one-shot' voting setting, voters use their cognitive capacities to predict the votes of fellow voters, in order to determine their own optimal (sincere or insincere) ballot. We show that in general, sophisticated agents who apply higher-order reasoning will not refrain from manipulation. We also consider higher-order reasoning in iterative voting procedures. We investigate whether voters that are able to predict (possibly harmful) future manipulations by fellow voters will avoid a strategic vote. For positional scoring rules and Condorcet extensions, we prove that this is not the case. Finally, we investigate how strategic incentives are affected if we allow voters to communicate with each other. It is shown that in many cases, voters cannot improve the outcome of the election by sharing personal information with their peers.

# Acknowledgements

First of all, I wish to thank my supervisors Ronald de Haan and Fernando Velázquez Quesada for their confidence in me and their enthusiasm during the whole process of writing this thesis. I would like to express my gratitude for your constant support, guidance and for all the discussions that have shaped this thesis. You were always open to discuss my work. Ronald, your optimistic attitude and helpful thoughts were extremely valuable. Our meetings gave me a lot of energy and I always left with more confidence about what I was doing. Fernando, thank you for all your interesting ideas, prompt responses to my questions and queries, and your countless helpful suggestions. Your thoughtful remarks and instant feedback have been of utmost help to me.

I would also like to thank Sonja Smets for the fruitful meetings in the beginning of the process. Your ideas and accurate comments have helped me to get a clearer view of existing research and to form a better understanding of the many possible directions for this thesis.

Nick Bezhanishvili, who introduced me to so many interesting topics in logic when I was still a bachelor student, has been a great academic mentor and guide through the Master of Logic. I thank you above all for trusting and believing in me.

Thanks to the members of the thesis committee for taking the time to read this thesis, but also for teaching many inspiring courses that I was lucky enough to attend during my time as a Master of Logic student.

I would also like to express my gratitude to my family and friends for all the love and support they have so generously given me. Spending a week in Rivières-le-Bois with my friends last summer while writing my thesis was one of the best experiences. I would like to thank my parents for their unconditional support in all its forms. Finally, Bram, thank you for your encouragement, patience and your belief in me, but most of all thank you for being there for me, always.

# Contents

# Chapter 1

# Introduction

Social choice theory is the study of mechanisms for aggregating individual preferences, opinions or judgments into a collective decision. In this thesis, we focus on a specific method for collective decision making: voting. Voting is a commonly used procedure to make a collective choice, for example when a group of individuals wishes to resolve a disagreement, determine a common opinion, elect a representative, choose a public policy, find a winner in a contest or to solve any other problem of aggregating individual preferences over a set of candidates into a group decision. One of the main questions is: how can a collective make a democratic and reasonable decision between the alternatives, on the basis of its members' individual preferences?

An important topic in social choice theory and in voting theory in particular is strategic manipulation. Often, we want the voting procedure to be strategyproof: we do not want individuals to have an incentive to lie about their preferences. This incentive exists whenever it is beneficial for an agent to lie about her preference[1]. Strategic manipulation is undesirable, because a fair and democratic method to make a collective decision should elect the best candidate given everyone's preference. If the voting procedure creates an incentive for voters to misrepresent their preferences, the outcome of the election may not be a good reflection of the voters' opinions.

A central result in Computational Social Choice is the Gibbard-Satterthwaite Theorem, which states that every reasonable voting rule is sensitive to manipulation. However, this does not imply that any democratic election has to deal with manipulative voters. There are some implicit assumptions in the Gibbard-Satterthwaite theorem that are disputable. First of all, the Gibbard-Satterthwaite Theorem assumes that any preference order is possible (the universal domain assumption), while in reality, single-peaked preferences are more natural in many cases. Moulin (1980) showed that if we restrict to single-peaked preferences, there exist strategyproof voting rules. Another escape from the Gibbard-Satterthwaite theorem that is widely studied is the complexity of manipulation: even though it is theoretically possible to manipulate, it is computationally too hard to find a successful manipulative ballot (see Conitzer and Walsh (2016) for an overview of work on computational complexity as barrier to manipulation).

Finally, it is assumed that there is just a single manipulator with full information about the preferences of the other voters. This means that only situations in which a single voter might cast a strategic vote are considered, while every other voter votes truthfully, no matter what

[1] In this thesis, we use feminine pronouns to refer to a voter. When a second voter is considered, we refer to that voter by masculine pronouns.

the outcome of the election will be. Many recent papers have studied situations where a single voter or a small coalition tries to manipulate an election, assuming that the other voters always cast a sincere vote. Since there is only one manipulator, the knowledge of the other agents does not have to be taken into consideration: no matter how much information they have, they will always vote truthfully. Therefore, the manipulative voter does not have to worry about the strategic behaviour of the other voters: she only cares about how her own (possibly untruthful) ballot can affect the outcome of the election.

These circumstances make it unrealistically easy for the manipulator to strategise, because a manipulation is completely risk-free. When voters have less information about the preferences of other voters, strategising can be risky: when a manipulative voter is uncertain about the ballots of the other voters, her manipulation might turn out badly and may result in an outcome that is less favourable than if she had not strategised. Situations where voters have some uncertainty about the preferences of the others haven been considered in recent work (see for example Conitzer et al. (2011); Reijngoud and Endriss (2012)). In these analyses of manipulation under partial information, it is generally assumed that all available information is publicly announced. This could be interpreted as poll information provided by a reliable authority and sharing information results in common knowledge of all participating voters. This is a very basic framework: information is static, and there is no form of communication between individuals.

To get an understanding of how voting rules function under more realistic assumptions, we have to study a more sophisticated model of strategic manipulation in which there can be multiple voters who consider a strategic vote. A situation where multiple voters might strategise is more complex than the situation with a single manipulator: a voter should not only reason about the effect of her own strategic ballot on the outcome of the election, but she should also take into account the possible strategic ballots of other manipulators. When reflecting on her information about the preferences of other voters, a voter might realise that some of her fellow voters have an incentive to vote strategically. In that case, it makes sense that she will determine her best (sincere or insincere) ballot given that some other voters report an untruthful ballot. So, she starts reasoning about the strategic reasoning of other voters, which we call higher-order reasoning.

When we study the manipulability of voting procedures, communication between voters is important, because new information might affect the manipulability of the election. Information about other voters' ballots will be useful for a voter to determine whether she has a strategic vote. From new information, a voter might learn that she has the possibility to cast a strategic vote and change the outcome of the election in her favour, or she might learn that an untruthful ballot will not result in a better outcome. To analyse situations in which multiple manipulators operate, for example cases where manipulators try to form coalitions, a richer framework of information exchange can be useful.

A very interesting question will be what happens in a situation where voters are higher-order reasoners with the ability to communicate. For example, voter 1 knows that voter 2 has a strategic manipulation that is beneficial for both of them, but she also knows that voter 2 does not know this. Then she has an incentive to inform voter 2 about the manipulation. In other situations, communication between voters might enable them to form coordinated coalitions that try to manipulate the election as a group.

We investigate this issue by developing a model of strategic voting, in which agents are able to apply higher-order reasoning. We want to develop a framework in which we can model

information exchange, model-changing actions like changing a ballot and higher-order knowledge (knowledge about the knowledge of other agents). This is exactly what Dynamic Epistemic Logic (DEL) allows us to do: DEL, broadly conceived, is the study of logics of information change and makes it possible to analyse epistemic and doxastic consequences of new information and factual change (Van Benthem, 2011; Van Ditmarsch et al., 2007). Before elaborating on the content of this thesis in more detail, we will first motivate the general idea with an example.

**Example 1.1.** Four friends, Alice, Bob, Carol and Dave, have decided to go on a vacation, but they have not agreed on their holiday destination yet. They have a shortlist of three countries: Austria ($a$), the Bahamas ($b$) and China ($c$). Alice and Bob really like skiing and hiking, so Austria is their top choice. Their second choice is the Bahamas. Carol and Dave are not really into a very active vacation, so Austria is their least favourite option. Carol really likes beaches with white sand and clear blue water, so she prefers the Bahamas, and her second choice is China. Dave likes visiting big cities and cultural attractions, so he prefers China over the Bahamas. Table 1.1 summarises their preferences. Alice is voter 1, Bob is voter 2, Carol is voter 3 and Dave is voter 4. We abbreviate $a \succ b \succ c$ by $abc$, $b \succ c \succ a$ by $bca$ and so on.

| | |
|---|---|
| 1 | $abc$ |
| 2 | $abc$ |
| 3 | $bca$ |
| 4 | $cba$ |

Table 1.1: The preferences of the four friends

They decide to vote about the alternatives on the shortlist. They use the Borda rule with lexicographic tie-breaking[2]. If all friends vote truthfully, they will go to the Bahamas. Carol and Dave know exactly what everyone's preferences are, but are not considering a manipulative vote. Alice and Bob both know the preferences of Carol and Dave. Moreover, they know from each other that Austria is their top choice. However, they are uncertain about each others' second and third choice.

In Figure 1.1, the voting situation and possible worlds are shown[3]. In the actual world, indicated with grey, alternative $b$ (the Bahamas) is elected. Alice is uncertain about the ballot of Bob, and Bob is uncertain about the ballot of Alice: they both think that the other voter will vote either $abc$ or $acb$. Alice (voter 1) cannot distinguish between the two upper worlds. If she changes her ballot to $acb$, there are two options: if the left upper world is the actual world, changing her ballot to $acb$ makes alternative $a$ win, so then she has a successful manipulation, and they will go to Austria. However, if the right upper world is the actual world, changing her ballot makes alternative $c$ win, which is worse than $b$ for her, because she prefers the Bahamas to China. Since this is risky, she is not able to strategise. In the same way, Bob (voter 2) is not able to cast a safe strategic vote. Even if Alice and Bob would know that they both have preference $abc$, it is still impossible to manipulate: if Alice considers it possible that Bob manipulates the election by voting $acb$, reporting $acb$ as well would result in electing $c$. The same holds for Bob.

[2]Under the Borda rule, every voter assigns a number of points to the alternatives. Every voter submits a full ballot order. In the case of four alternatives, for every submitted ballot, the alternative ranked first gets 3 points, the alternative ranked second gets 2 point, the alternative ranked third gets 1 point and the alternative ranked last gets no points. The score of an alternative is the sum of all its points, and the alternative with the highest score wins. Ties are broken according to the order $a \succ b \succ c$.

[3]See Section 2.2 for a basic introduction to dynamic epistemic models.

So, if Alice and Bob are uncertain about each other's votes, it is too risky to manipulate. They cannot safely manipulate individually, and they need some form of communication to coordinate a manipulation.

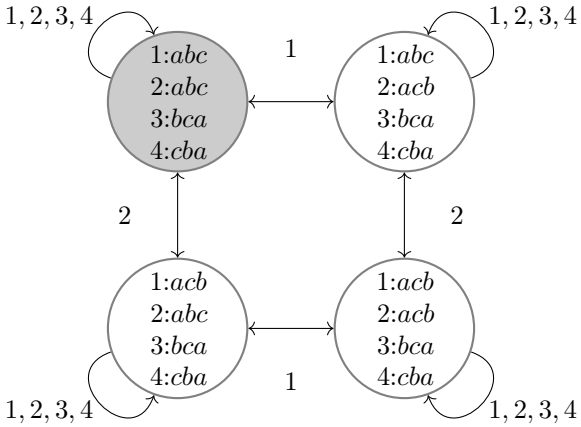



Figure 1.1: Alice (voter 1) and Bob (voter 2) are not able to vote strategically

If Bob in front of the group says: 'I will vote *abc*' (and it is common knowledge that no voter considers it possible that this is false information), then we can update the model by simply removing the worlds in which this statement is not true. The new model is shown in Figure 1.2. Now, Alice knows that she can cast a safe strategic vote: by voting *acb* instead of *abc*, Austria will become the holiday destination of the group.

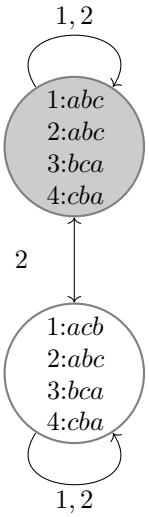



Figure 1.2: Alice (voter 1) learns that she has a successful strategic vote

The main goal of this thesis will be to investigate what kind of knowledge, reasoning and

communication is required to be able to manipulate, and under which conditions the Gibbard-Satterthwaite Theorem ceases to hold. In this thesis we start with developing a framework that models higher-order reasoning agents who reflect on the reasoning of their peers, different types of partial information and communication between agents in the setting of voting and strategising.

We will demonstrate how this framework can be used to model classical voting settings, where manipulative voters are naive and think that every other voter reports a sincere ballot. First, we will analyse strategic behaviour of higher-order reasoning in the setting where voting is a 'one-shot' event: you have one chance to cast a ballot and you are not able to change your vote afterwards. This setting fits many traditional political elections. When voting is a one-shot event, higher-order reasoning voters will use their cognitive capacities to predict the votes of fellow voters, in order to be able to determine their own optimal (sincere or insincere) ballot.

Another way to understand an election is to treat it as a process, and see if we can reach some point of equilibrium, where all voters are satisfied with their votes, no longer wishing to change them. This setting is called iterative voting and fits many non-traditional settings of voting, for example the process of making a social decision via websites as Doodle and Facebook. In an iterative voting process, higher-order reasoning voters use their cognitive capacities to obtain a more farsighted view of the consequences of their manipulative ballots. Before they decide to manipulate, they predict how their manipulation could trigger future manipulations in the iterative voting procedure, and whether the long-term consequences of a manipulation are beneficial or not. In this thesis, we will investigate how higher-order reasoning affects the manipulability of voting rules.

Finally, we will analyse whenever it is beneficial for a voter to share information with other voters. In some cases, a voter is not able to manipulate the election with her own ballot, but she might be able to stimulate another voter to strategise by sharing certain information. We will make a first attempt to explore the relation between communication and strategic manipulation in voting.

## 1.1 Outline of the thesis

**Chapter 2**: In this chapter, we will provide the technical background that forms the baseline of this thesis.

**Chapter 3**: Our starting point is the standard DEL framework. In Chapter 3, we extend this standard model to a framework that can be used to reason about strategic manipulation, and for which there is a sound and complete axiom system.

**Chapter 4**: In Chapter 4, we will discuss how standard approaches in strategic voting can be modelled in our framework. We will discuss the characteristics of models that satisfy the assumptions of the Gibbard-Satterthwaite theorem, but we will also consider some examples with weaker assumptions, such as partial information about the preferences of other voters.

**Chapter 5**: In Chapter 5, we will introduce higher-order reasoning agents. We will first show how the notion of safe manipulation (Slinko and White, 2014) can be seen as a setting in which some of the agents apply higher-order reasoning. Then, we generalise this to a setting in which all agents may apply some form of higher-order reasoning. We follow the cognitive hierarchy theory and define level-$k$ reasoners as agents who believe that every other agent is a level-$(k-1)$

reasoner. We will show that in general, reasoning on a higher level will not make voters refrain from manipulation.

**Chapter 6**: We propose a new perspective on higher-order reasoning about manipulation in iterative voting. If voters apply higher-order reasoning in iterative voting and are able to predict future manipulations, there might be iterative voting procedures that do not converge in the classical setting, but do converge in this higher-order reasoning setting. We show that positional scoring rules and Condorcet extensions are not strategyproof for voters that are able to look one step ahead. However, there are iterative voting procedures that always converge if voters apply second-order reasoning instead of first-order reasoning.

**Chapter 7**: In Chapter 7, we will explore how communication between agents affects strategic manipulation in voting. We will introduce the idea of strategic communication: a voter has an incentive to strategically communicate some information, if a voter thinks that the outcome of the election can be improved by sharing that information with like-minded voters. We discuss under which conditions a voter may benefit from sharing information.

**Chapter 8**: In Chapter 8, we give a conclusion of this thesis and we discuss some directions for future research.

## 1.2 Related work

A link between epistemic logic and voting has been given in Chopra et al. (2004). They use knowledge graphs to indicate the uncertainty of voters regarding other voters' preferences and develop a bimodal logic.

Van Ditmarsch et al. (2013) propose a model for strategic voting in which voters have partial information about other voters' preferences or about other voters' knowledge about their own vote. They define notions of manipulation and equilibrium, and they model information updates about preferences as public announcements. They show that some forms of manipulation are preserved under such updates and others not. A local preference of an agent in a state induces a global preference of that agent on the model. This means that some states are more preferable for an agent than others. Technically, these models are similar to epistemic plausibility models.

Bakhtiari et al. (2018) follow-up on this paper and use a similar framework for strategic voting and higher-order knowledge via knowledge profiles, which are standard S5 epistemic models. Information updates are modelled as truthful public announcements. An election is modelled as an (imperfect information) Bayesian game. Bakhtiari et al. explore how dominant manipulation and knowledge of manipulation are related and which forms of manipulation and (conditional) equilibria are preserved under such information updates and present a logic for strategic voting. In this thesis, we will work with a dynamic epistemic model that also captures insincere announcements and false belief, but also more complex forms of communication such as private announcements.

Modal logics of social choice and of voting have been proposed by Troquard et al. (2011) and, building on that, by Ciná and Endriss (2016) and Perkov (2016). In these logics the semantic primitives are ballots of individual voters and ballot profiles. There are two levels of preferences: reported preferences, given by the valuation on the model, and true preferences given by a general parameter of the model. The modalities only encode manipulability and true preferences,

and not uncertainty, as we do.

Van Eijck (2013) discusses how Propositional Dynamic Logic (PDL) can be used as a multi-agent strategic logic. This logic for strategic reasoning has group strategies as first class citizens, and brings game logic closer to standard modal logic. It is demonstrated that Multi-Agent Strategy Logic (MASL) can express key notions of game theory, social choice theory and voting theory in a natural way. Van Eijck gives a sound and complete proof system for MASL and also discusses an extension of this language to epistemic multi-agent strategic logic.

Terzopoulou (2017) extends the basic framework of judgment aggregation, which is another branch of Computational Social Choice, to a framework that deals with partial knowledge and higher-order reasoning about strategic behaviour. Terzopoulou follows the idea of the cognitive hierarchy theory, that defines level-$k$ reasoners as agents who believe that every other voter reasons at level-$(k-1)$. We translate this framework to the setting of voting theory, and show how level-$k$ reasoners can be modelled in our dynamic epistemic framework.

# Chapter 2

# Background

This chapter is meant to provide a baseline for the formal notions of this thesis.

## 2.1 Voting theory

Let $N = \{1, \ldots, n\}$ be a finite set of *voters* and $X = \{x_1, \ldots, x_m\}$ a finite set of *alternatives*. Let $\mathcal{L}(X)$ denote the set of all strict linear orders on $X$, which means that the order is irreflexive, transitive, and for any two alternatives $x, y \in X$, exactly one of the following is true: $x \succ y$, $y \succ x$ or $x = y$. We use elements of $\mathcal{L}(X)$ to model *true preferences* and *declared ballots*. Each voter has a preference over the alternatives in $X$, where $p_i$ denotes the preference of voter $i$. The preferences of all voters together are called a *preference profile* $\boldsymbol{p}_i = (p_1, \ldots, p_n) \in \mathcal{L}(X)^n$. Every voter submits a ballot $b_i$, which is again a strict linear order over the alternatives in $X$, giving rise to a *ballot profile* $\boldsymbol{b} = (b_1, \ldots, b_n) \in \mathcal{L}(X)^n$. Given a preference profile $\boldsymbol{p}$, $\boldsymbol{p}(i)$ refers to the $i^{\text{th}}$ component of the vector, so the preference order of voter $i$ in $\boldsymbol{p}$. We define $\boldsymbol{b}(i)$ in the same way. For a set of voters $G \subseteq N$, $\boldsymbol{p}(G)$ and $\boldsymbol{b}(G)$ respectively denote the (partial) preference profile and (partial) ballot profile. Let $-i := N \backslash \{i\}$. We use $x \succ_{p_i} y$ to denote that $x$ is ranked higher in the preference $p_i$ of voter $i$, and $x \succ_{b_i} y$ to denote that $x$ is ranked higher in the ballot $b_i$ of voter $i$. A *social choice function* or *voting rule* $F$ for $N$ and $X$ selects one or more winners for every ballot profile:

$$F : \mathcal{L}(X)^n \to \wp(X) \setminus \{\emptyset\}.$$

The following examples are common social choice functions:

- Positional scoring rules (PSRs): a positional scoring rule is defined by a so-called scoring vector $\mathsf{score} = (\mathsf{score}_1, \ldots, \mathsf{score}_m) \in \mathbb{R}^m$ with $\mathsf{score}_1 \geq \mathsf{score}_2 \geq \cdots \geq \mathsf{score}_m$ and $\mathsf{score}_1 > \mathsf{score}_m$. A PSR with scoring vector $(m-1, m-2, \ldots, 0)$ is called *Borda*, *plurality* is a PSR with scoring vector $(1, 0, \ldots, 0)$, *anti-plurality* or *veto* is a PSR with scoring vector $(1, \ldots, 1, 0)$ and for any $k < m$, *k-approval* is a PSR with scoring vector $(\underbrace{1, \ldots, 1}_{\text{k}}, \underbrace{0, \ldots, 0}_{\text{m-k}})$.

  An alternative receives $\mathsf{score}_j$ points for each voter who ranks him at the $j$th position. The alternative(s) with the most points win(s) the election.

- Copeland: an alternative's score is the number of pairwise majority contests he wins minus the number he loses. The alternative with the highest score wins. A pairwise majority contest between candidates $x$ and $y$ is won by $x$ if a majority of voters rank $x$ above $y$.

- Single transferable vote (STV): An STV election proceeds in rounds. In each round the alternative ranked first by the fewest voters gets eliminated. This process is repeated until only one candidate remains (or until all remaining candidates are ranked first equally often).

Sometimes, it is useful to refer directly to the position of an alternative in the ballot of an agent $i$. We define it as follows:

**Definition 2.1.** Let $i \in N$ be a voter with ballot $b_i$. Let $x \in X$ be an alternative. Then

$$\mathsf{rank}_{b_i}(x) := |\{y \mid y \succ_{b_i} x\}| + 1.$$

If $|F(\boldsymbol{b})| = 1$ for all ballot profiles $\boldsymbol{b}$, then $F$ is called *resolute*. Most natural voting rules are irresolute and have to be paired with a tie-breaking rule to always select a unique election winner. A tie-breaking rule picks a unique winner from the set of initial winners. We assume that tie-breaking rules are choice functions: $T : 2^X \setminus \{\emptyset\} \to X$, where $T(Y) \in Y$ for every $Y \subseteq X$. An example of a tie-breaking rule that is *not* a choice function is the random tie-breaking rule which breaks ties randomly. Another example is a tie-breaking rule that depends on the ballot profile, for example the tie-breaking rule that picks the alternative that is ranked highest by voter 1. Sometimes we further restrict attention to *rationalisable* tie-breaking rules, i.e., tie-breaking rules under which ties are broken according to some fixed but arbitrary order $\triangleright$ over the candidates. If $F$ is resolute, instead of $F(\boldsymbol{b}) = \{x\}$, we will write $F(\boldsymbol{b}) = x$.

A resolute voting rule is surjective if each candidate wins under at least one ballot profile. If voters are treated equally, we say that a voting rule is anonymous. When all alternatives are treated equally, we call the voting rule neutral. Constant voting procedures always elect the same, unique winner. If a voting rule is a dictatorship, the alternative that is ranked first by the dictator always wins. A voting procedure is unanimous if it elects alternative $x$ whenever $x$ is ranked first by all voters. A voting procedure satisfies the Pareto condition if it does not return a alternative that is ranked below some other alternative by all voters. A social choice function is Condorcet-consistent if an alternative wins whenever it is ranked higher than every other alternative by a majority of the voters. It is strongly Condorcet-consistent if all alternatives are elected that are ranked higher than every other alternative by at least half of the voters. These properties are formally defined as follows:

**Definition 2.2** (Properties of social choice functions)**.** Some important properties of social choice functions are:

- A resolute social choice function $F$ is *surjective* if for any alternative $x \in X$ there is a ballot profile $\boldsymbol{b} \in \mathcal{L}(X)^n$ such that $F(\boldsymbol{b}) = x$.
- A social choice function $F$ is *anonymous* if $F(b_1, \ldots, b_n) = F(b_{\tau(1)}, \ldots, b_{\tau(n)})$ for any ballot profile $\boldsymbol{b} \in \mathcal{L}(X)$ and any permutation $\tau : N \to N$.
- A social choice function $F$ is *neutral* if alternatives are treated symmetrically: $F(\tau(\boldsymbol{b})) = \tau(F(\boldsymbol{b}))$ for any profile $\boldsymbol{b}$ and any permutation $\tau : X \to X$ (with $\tau$ extended to preferences and profiles in the natural manner)
- A social choice function $F$ is *constant* if there is a candidate $x \in X$ such that $F(\boldsymbol{b}) = x$ for any ballot profile $\boldsymbol{b} \in \mathcal{L}(X)^n$.
- A social choice function is a *dictatorship* if there exists $i \in N$ such that for all $\boldsymbol{b} \in \mathcal{L}(X)^n$, $F(\boldsymbol{b}) = x$ for $x \in X$ with $\mathsf{rank}_{\boldsymbol{b}(i)}(x) = 1$. That is, $F$ always selects the alternative that is ranked first by voter $i$.

- A social choice function $F$ is *unanimous* if it always selects an alternative that is top-ranked by everyone. Thus, if $\mathsf{rank}_{\boldsymbol{b}(i)}(x) = 1$ for all $i \in N$, then $F(\boldsymbol{b}) = x$.
- A social choice function $F$ is *Pareto-efficient* if for every ballot profile $\boldsymbol{b} \in \mathcal{L}(X)^n$: $y \notin F(\boldsymbol{b})$ if there exists $x \in X$ such that $y \succ_{\boldsymbol{b}(i)} x$ for every voter $i \in N$.
- A social choice function $F$ is *Condorcet-consistent* if for any ballot profile $\boldsymbol{b} \in \mathcal{L}(X)^n$:

  $$F(\boldsymbol{b}) = \{x\} \text{ whenever } |b(x \succ y)| > |b(y \succ x)| \text{ for all } y \in X \backslash \{x\}.$$

  Social choice functions that are Condorcet-consistent are called *Condorcet extensions*.
- A social choice function $F$ is *strongly Condorcet-consistent* if for any ballot profile $\boldsymbol{b} \in \mathcal{L}(X)^n$:

  $$F(\boldsymbol{b}) = W^{\boldsymbol{b}} \text{ whenever } W^{\boldsymbol{b}} \neq \emptyset,$$

  where $W^{\boldsymbol{b}}$ is the set of all weak Condorcet winners of $\boldsymbol{b}$, i.e.:

  $$W^{\boldsymbol{b}} = \{x \in X \mid |b(x \succ y)| \geq |b(y \succ x)| \text{ for all } y \in X \backslash \{x\}\}$$

### 2.1.1 Manipulation

A voter is pivotal at a ballot profile if she can change the outcome of an election by just changing her own ballot. A voter is pivotal with respect to an alternative $x$, if she can make $x$ a winner of the election by just changing her own ballot.

**Definition 2.3.** A voter $i$ is *pivotal* at ballot profile $\boldsymbol{b}$ under a social choice function $F$ if there exists a ballot $b_i' \neq \boldsymbol{b}(i)$ such that

$$F(b_i', \boldsymbol{b}(-i)) \neq F(\boldsymbol{b}).$$

If $F(b_i', \boldsymbol{b}(-i)) = x$, we say that voter $i$ is *pivotal with respect to* $x$.

A voter $i$ votes *truthfully* if she reports her true preference $p_i$, and *untruthfully* otherwise. In classical voting theory, a voter $i$ is said to have an incentive to manipulate if she can improve the election outcome with respect to $p_i$ by voting untruthfully. A resolute voting rule is susceptible to manipulation if there is a profile in which some voter has an incentive to manipulate. If a resolute voting rule is not susceptible to manipulation, then it is immune to manipulation.

**Definition 2.4** (Incentive to manipulate - classical voting theory)**.** Given a resolute voting rule $F$, a voter $i$ and a ballot profile $\boldsymbol{b}$ with $b_i = p_i$, $i$ has an *incentive to manipulate* if there exists a ballot $b_i' \neq b_i$ such that

$$F(b_i', \boldsymbol{b}(-i)) \succ_{p_i} F(\boldsymbol{b})$$

**Definition 2.5.** A social choice function $F$ is *strategyproof* if for any ballot profile, no individual voter has an incentive to manipulate.

From the perspective of computational social choice, strategyproof voting procedures are desirable: a fair and democratic voting rule should elect the best candidate given everyone's preference, and the voting rule should not create an incentive for voters to misrepresent their preferences. In particular when many voters try to manipulate, the resulting ballot profile may turn out to be very far from the electorate's true preferences and thus not representative. In addition, voters should not have to waste resources pondering over what other voters will do and trying to figure out how best to respond. Unfortunately, it turns out that for any 'democratic' voting rule, there exist situations in which there is a voter with an incentive to manipulate. Gibbard (1973) and Satterthwaite (1975) independently proved the following theorem:

**Theorem 2.6** (Gibbard-Satterthwaite)**.** *When $m \geq 3$, any resolute social choice function that is surjective and strategyproof, is a dictatorship.*

For a simple proof of the Gibbard-Satterthwaite Theorem, see Benoît (2000). This central result has far-reaching implications: it states that if you have three or more alternatives, any 'reasonable' voting rule is susceptible to manipulation. As discussed in the introduction, the assumption of full information is very strong and makes it unrealistically easy for a voter to manipulate.

### 2.1.2 Manipulation under partial information

In this section, we will recall the framework of Conitzer et al. (2011) and Reijngoud and Endriss (2012) that is used to analyse manipulation of a single manipulator under partial information. In order to study strategic behaviour of voters with incomplete information, we have to refine the definition of strategic manipulation. Since a voter might have incomplete information about the ballots of the other voters, she might be uncertain about the winner of the election under the current profile and about the election outcome when she changes her ballot. We define the information set of a voter $i$ as the set of partial ballot profiles (with a ballot order for every voter except $i$) that $i$ considers possible, given the information $\pi(\boldsymbol{b})$.

A *poll information function* (PIF) is a function $\pi$ that maps ballot profiles to 'pieces of information'. This piece of information allows an agent to define a set of ballot profiles, exactly those satisfying the provided information. The formal definition of that pieces of information and the corresponding poll information function depends on the type of information that is communicated to the voters. The following examples are choices for PIFs that were introduced by Reijngoud and Endriss (2012):

- The profile-PIF $\pi : \mathcal{L}(X) \to \mathcal{L}(X)$ simply outputs the full ballot profile: $\pi(\boldsymbol{b}) = \boldsymbol{b}$.
- Given a social choice function $F$, the corresponding winner-PIF maps ballot profiles to the winning alternative under the ballot profile with respect to $F$: $\pi : \mathcal{L}(X) \to X$, where $\pi(\boldsymbol{b}) = F(\boldsymbol{b})$.
- The MG-PIF returns the majority graph of the ballot profile. A majority graph is a directed graph in which each node represents an alternative. There is an edge $(x, y)$ from $x$ to $y$ if and only if $x$ wins the pairwise majority contest between $x$ and $y$. Let $G$ denote the set of all finite graphs. Then $\pi : \mathcal{L}(X) \to G$, and if $mg(\boldsymbol{b})$ is the majority graph of ballot profile $\boldsymbol{b}$, we have $\pi(\boldsymbol{b}) = mg(\boldsymbol{b})$.
- The WMG-PIF returns the weighted majority graph of the ballot profile. A weighted majority graph is a majority graph in which every edge is assigned a label. Each edge $(x, y)$ is labelled with the difference between the number of voters ranking $x$ above $y$, and the number of voters ranking $y$ above $x$. Let $G_{\text{weighted}}$ be the set of all finite weighted graphs, then $\pi : \mathcal{L}(X) \to G_{\text{weighted}}$. Let $wmg(\boldsymbol{b})$ be the weighted majority graph of ballot profile $\boldsymbol{b}$. Then $\pi(\boldsymbol{b}) = wmg(\boldsymbol{b})$.
- Given a social choice function $F$, the corresponding score-PIF returns for each candidate its score under the input profile according to $F$. $F$ should assign points to each candidate for this PIF to be well-defined. Formally, $\pi(\boldsymbol{b}) = (\mathsf{score}_F(x_1, \boldsymbol{b}), \ldots, \mathsf{score}_F(x_m, \boldsymbol{b}))$, where $\mathsf{score}_F : X \times \mathcal{L}(X)^n \to \mathbb{N}$ computes the score of an alternative $x \in X$ under ballot profile $\boldsymbol{b} \in \mathcal{L}(X)^n$ according to $F$.

Then, given the information $\pi(\boldsymbol{b})$, every agent has an information set consisting of partial ballot profiles that she considers possible.

**Definition 2.7** (Information set)**.**

$$\mathcal{W}_i^{\pi(\boldsymbol{b})} = \{\boldsymbol{b}'(-i) \in \mathcal{L}(X)^{n-1} \mid \pi(b_i, \boldsymbol{b}(-i)') = \pi(\boldsymbol{b})\}$$

We say that a voter has an incentive to manipulate if there is a scenario in her information set for which some untruthful ballot would result in a better outcome for her, and there is no scenario in which changing her ballot to that new ballot would result in a worse outcome than when she would vote truthfully.

**Definition 2.8** (Incentive to manipulate with partial information)**.** Given an info set $\mathcal{W}_i^{\pi(\boldsymbol{b})}$ of voter $i$, we say that voter $i$ has an *incentive to manipulate* by voting $b_i' \neq \boldsymbol{b}(i)$ if $b_i' \neq \boldsymbol{p}(i)$, there exists a partial ballot profile $\boldsymbol{b}'(-i) \in \mathcal{W}_i^{\pi(\boldsymbol{b})}$ such that $F(b_i', \boldsymbol{b}'(-i)) \succ_{\boldsymbol{p}(i)} F(\boldsymbol{b}(i), \boldsymbol{b}'(-i))$ and for every ballot profile $\boldsymbol{b}''(-i) \in \mathcal{W}_i^{\pi(\boldsymbol{b})}$, $F(b_i', \boldsymbol{b}''(-i)) \succeq_{\boldsymbol{p}(i)} F(\boldsymbol{b}'')$.

Thus, manipulation must be 'safe': by manipulating, the outcome under the untruthful ballot will never be worse than the outcome under the truthful ballot. We say that voters are *risk-averse*: if there exists a situation in which an untruthful ballot would result in a worse outcome, they refrain from strategising. This amounts to *de re knowledge of manipulation* (Van Ditmarsch et al., 2013): there is a vote that is strategic for any ballot profile she considers possible. A weaker form of knowledge is *de dicto knowledge of manipulation*: in this case, the voter knows that for every ballot profile she considers possible, she has a strategic vote, but this vote is not the same strategic vote for every ballot profile. So, she knows that she has a strategic vote, but she does not know what the manipulation is.

Another important aspect of the definition is that the manipulative vote only has to (weakly) improve the outcome given a certain ballot profile: the outcome under a partial ballot profile and a strategic ballot of voter $i$, should be (weakly) better than the outcome under that same partial profile and a truthful ballot of voter $i$.

An alternative way to specify the incentives of an agent to manipulate an election is by looking at her best strategies under the information she holds. Consider a social choice function $F$, a ballot profile $\boldsymbol{b}$, preference profile $\boldsymbol{p}$ and an agent $i \in N$ with information set $\mathcal{W}$. We say that a ballot order $b_i$ is *undominated* if there is no other ballot $b_i'$ such that

- $F(b_i', \boldsymbol{b}(-i)) \succeq_{\boldsymbol{p}(i)} F(b_i, \boldsymbol{b}(-i))$ for all $\boldsymbol{b}(-i) \in \mathcal{W}$, and
- $F(b_i', \boldsymbol{b}(-i)) \succ_{\boldsymbol{p}(i)} F(b_i, \boldsymbol{b}(-i))$ for some $\boldsymbol{b}(-i) \in \mathcal{W}$.

If agent $i$'s true preference $p_i$ is undominated, then this will be her unique best strategy. Otherwise, all the undominated ballots form her set of best strategies. We assume that she will pick a strategy from this set by some decision mechanism $D : \wp(\mathcal{L}(X)) \to \mathcal{L}(X)$. For example, a voter may choose the strategy that is closest to her truthful ballot. In this thesis, we will assume that every voter uses the same decision mechanism $D$ to pick a strategy from the set of best strategies, and that $D$ is common knowledge in the group of voters. We define the best strategy of a voter $i$ as follows:

**Definition 2.9** (Best strategy)**.** Let $i \in N$ be a voter with information set $\mathcal{W}$ and true preference $p_i$. Then her best strategy is defined as

$$\mathcal{S}_i(\mathcal{W}, p_i) := \begin{cases} p_i & \text{if } p_i \text{ is undominated} \\ D(\{b_i \mid b_i \in \mathcal{L}(X) \text{ is undominated}\}) & \text{otherwise} \end{cases}$$

So, $\mathcal{S}_i$ always returns a single ballot order.

### 2.1.3 Iterative voting

So far we have modelled voting as a one-shot event: voters declare their preferences and the voting rule computes a definitive outcome. There is no option to change or revise a decision. While this assumption fits some political voting settings, the reality is more complex. Committees often follow an informal voting process where members are free to revise their votes or hold straw polls. Online voting tools such as Facebook and Doodle allow voters to see previous votes and to change their vote later on, and even in traditional political voting, polls broadcast in the media may trigger voters to change their vote.

In iterative voting, voting proceeds in rounds. We assume that in the first round, every voter reports her truthful preference order. Every round, the reported ballot profile and the outcome are announced, but voters may change their votes after observing the current ballot profile and outcome. The game proceeds in turns, where a single voter changes her vote at each turn, until no voter has objections and the final outcome is announced. The common assumption in iterative voting is that voters do not reason about the other voters' preferences or who might change their vote, and thus act in a *myopic* way. That is, the voters vote in every round as if it is the last one, since they are not able to make a future prediction. In game-theoretic terms, each voter will play a best response to the current ballot profile of the other voters.

We use the following notation: $\boldsymbol{b}^t = (b_1^t, b_2^t, \ldots, b_n^t)$ is the ballot profile declared in round $t \geq 0$. In general, we assume that $\boldsymbol{b}^0 = \boldsymbol{p}$. Let $F$ be a resolute voting rule. After round $t$, voter $i \in N$ has a *better response* $b_i' \in \mathcal{L}(X)$ with $b_i' \neq \boldsymbol{b}^t(i)$ if $F(b_i', \boldsymbol{b}(-i)^t) \succ_i F(\boldsymbol{b}^t)$.

A profile without better responses, is a Nash equilibrium. After each round, one voter with better responses implements one of them. The process stops when there are no more better responses. We speak of *convergence* for the voting rule $F$, if the process always stops eventually. Often, we restrict to best responses. A *best response* for voter $i$ is a better response $b_i^*$ that cannot be improved:

$$F(b_i^*, \boldsymbol{b}(-i)^t) \succ_i F(\boldsymbol{b}^t) \text{ and } F(b_i^*, \boldsymbol{b}(-i)^t) \succeq_i F(b_i', \boldsymbol{b}(-i)^t) \text{ for all } b_i' \in \mathcal{L}(X)$$

A central question for iterative voting is under which conditions a process of iterative voting converges. The following result was proved by Meir et al. (2010) and later strengthened by Reijngoud (2011) and Brânzei et al. (2013).

**Theorem 2.10.** *Iterative voting restricted to arbitrary best responses converges for the plurality rule paired with lexicographic tie-breaking.*

Lev and Rosenschein (2012) and Reyhani and Wilson (2012) proved (independently from each other) a very similar result for anti-plurality.

**Theorem 2.11.** *An iterative election with anti-plurality as voting rule, a rationalisable linear-order tie-breaking and voters that use a best-response strategy, converge even when not starting from a truthful profile.*

Unfortunately, plurality and anti-plurality are the only positional scoring rules for which such a result is attainable. It was shown by Lev and Rosenschein (2016) that for any other scoring rule, the iterative voting process will not converge:

**Theorem 2.12.** *Under the iterative procedure, using a best response strategy and when voters are myopic, no scoring rule apart from plurality and veto converges.*

For other types of voting rules, it is an open problem whether iterative voting converges or not.

In the next section, we will introduce the basic epistemic framework to model knowledge, belief and information change.

## 2.2 Dynamic Epistemic Logic

Dynamic Epistemic Logic (DEL) is an extension of the basic epistemic logic with event models and product updates. It is a powerful framework that can be used to study model-changing actions (Baltag et al., 1999; Van Ditmarsch et al., 2007; Van Benthem et al., 2006; Van Benthem, 2011). The basic language of DEL is the same as standard epistemic logic.

**Definition 2.13** (Language)**.** The language of multi-agent epistemic logic $\mathscr{L}_{\text{DEL}}$ is generated by:

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

where $p \in P$ and $P$ is a countable set of atomic sentences. Other Boolean connectives are defined in the standard way. "$K_a\varphi$" reads as "The agent $a$ knows $\varphi$".

**Definition 2.14** (Epistemic model)**.** For a set $P$ of propositions, an *epistemic model* is a triple $M = (S, \{R_a\}_{a\in A}, V)$, where $S$ is a non-empty set of possible worlds, $\{\sim_a\}_{a\in A} \subseteq S \times S$ a family of binary equivalence relations over $S$ indexed by agents $a \in A$, and $V$ a valuation $V : P \to \wp(S)$, assigning a set of states to each proposition $p \in P$. By $(M, s^*)$ we denote the epistemic model $M$ with an actual world $s^* \in S$.

**Definition 2.15** (Semantics)**.**

$$\begin{array}{lcl}
M, s \vDash \top & \iff & \text{always} \\
M, s \vDash p & \iff & s \in V(p) \\
M, s \vDash \neg\varphi & \iff & M, s \nvDash \varphi \\
M, s \vDash \varphi_1 \wedge \varphi_2 & \iff & M, s \vDash \varphi_1 \text{ and } M, s \vDash \varphi_2 \\
M, s \vDash K_a\varphi & \iff & \text{for all } t \in S \text{ with } s \sim_a t, M, t \vDash \varphi
\end{array}$$

Event models are relational structures that can be used to describe a variety of informational actions, from public announcements to more subtle communications that may be private or (semi-)public, deception, and suspicion, and many other more complex forms of communication (Baltag et al., 1999).

**Definition 2.16** (Event model)**.** An *event model* is a tuple $U = (\Sigma, \{R_a\}_{a\in A}, pre)$ where $\Sigma$ is a non-empty set of events, $\{R_a\}_{a\in A} \subseteq \Sigma \times \Sigma$ is a set of equivalence relations over $\Sigma$ associated to the agents $A$ and $pre$ is a precondition map associating a formula $pre_e \in \mathscr{L}_{\text{DEL}}$ to each event $e \in \Sigma$. By $(U, e^*)$, we denote the event model $U$ with actual event $e^*$.

Given an initial epistemic model, we determine the model-transforming effect of an event model by constructing a new epistemic model. This is called the product update and it is used to model changes of information or belief[1].

[1] In this thesis, we will not deal with the philosophical meaning of the terms *knowledge* and *belief*. We assume that voters directly use whatever they think is true to reason about strategic behaviour, and hence we do not make a distinction between knowledge and belief

**Definition 2.17** (Product update)**.** Given an epistemic model $M = (S, \{R_a\}_{a\in A}, V)$ with actual state $s^*$ and an event model $U = (\Sigma, \{R_a\}_{a\in A}, pre)$ with actual event $e^*$, we define their product update $M \otimes U = (S \otimes \Sigma, \{R_a^{M\otimes U}\}_{a\in A}, V^{M\otimes U})$ to be a new state model, given by

- $S \otimes \Sigma = \{(s, e) \in S \times \Sigma \mid M, s \vDash pre_e\}$
- $(s, e)R_a^{M\otimes U}(s', e')$ if and only if $sR_a^M s'$ and $eR_a^U e'$
- $V^{M\otimes U}(p) = \{(s, e) \in S \otimes \Sigma \mid s \in V^S(p)\}$.

and with actual state $(s^*, e^*)$.

**Proposition 2.18.** Let $M$ be an epistemic model with actual world $s^*$ and let $U$ be an event model $U$ with actual event $e^*$ such that $M, s^* \vDash pre(e^*)$. Then, $M \otimes U$ is again an epistemic model.

*Proof.* See Van Ditmarsch et al. (2007). □

In the next chapter, we will extend this basic DEL framework to a model and a logic that can be used to analyse strategic manipulation in voting.

# Chapter 3

# The model

## 3.1 A dynamic epistemic model for strategic voting

In this section, we will introduce the framework that will be used in this thesis to model strategic voting.

Let $N = \{1, \ldots, n\}$ be the set of voters, and $X = \{x_1, \ldots, x_m\}$ be the set of alternatives they have to elect a winner from. Let $\mathcal{B}$ denote all possible ballot profiles and let $\mathcal{P}$ denote all possible preference profiles. Recall that ballot orders and preference orders are strict linear orders over the set of alternatives. Hence, $\mathcal{B} = \mathcal{P} = \mathcal{L}(X)^n$. The framework presented here is based on the DEL framework that was first introduced in Baltag et al. (1999) and extended by Van Benthem et al. (2006), but here we follow Van Ditmarsch and Kooi (2006).

Since we want to consider both truthful and untruthful votes, we have to make a distinction between true preferences and reported ballots. Therefore, we will use two types of propositional variables in the model: preference profiles and ballot profiles. We assume that every voter knows her own preference and ballot, so she will never be doubtful between two states in which her preference orders or ballot orders are distinct. This technical condition leads to a dependence of the accessibility relation for an agent $i$ on the valuation on the model. Hence, we first define a valuation on the states of the model, and then we define the accessibility relations. This is all formalised in the following definition.

**Definition 3.1** (Epistemic model for strategic voting)**.** Let $N$ be a set of voters, $X$ a set of alternatives and $F : \mathcal{B} \to X$ a resolute social choice function. An *epistemic model for strategic voting* for $F$ is a triple $M = (S, V, R)$, where

(i) $S$ is a non-empty set of states

(ii) $V : S \to \mathcal{P} \times \mathcal{B}$ is a valuation on $S$ that assigns a preference profile and a ballot profile to each world

(iii) $R : N \to \wp(S \times S)$ assigns an accessibility relation[1] to every agent $i \in N$, such that $sR_it$ implies that $\mathsf{proj}_1(V(s))(i) = \mathsf{proj}_1(V(t))(i)$ and $\mathsf{proj}_2(V(s))(i) = \mathsf{proj}_2(V(t))(i)$.

[1] We do not impose further restrictions on the accessibility relations of the voters in $N$. However, all models that will be used in this thesis are KD45 models, which means that the accessibility relations are serial, transitive and Euclidean.

By $(M, s^*)$, we denote an epistemic model for strategic voting with actual world $s^*$.

Given an epistemic model for strategic voting $M = (S, V, R)$, if $(s, t) \in R(i)$, we also denote this by $sR_it$. If it is clear from the context which valuation is used, we denote $\mathsf{proj}_1(V(s))$ by $\boldsymbol{p}_s$ (the preference profile in state $s$), and $\mathsf{proj}_2(V(s))$ by $\boldsymbol{b}_s$ (the ballot profile in state $s$). We refer to the preference and ballot order of voter $i$ in state $s$ by $\boldsymbol{p}_s(i)$ and $\boldsymbol{b}_s(i)$. We can extend this notation to sets of voters: if $G \subseteq N$, then $\boldsymbol{p}_s(G)$ is the (partial) preference profile of the voters in $G$ and $\boldsymbol{b}_s(G)$ is their (partial) ballot profile. To denote an arbitrary ballot profile or a partial ballot profile of $G \subseteq N$, we use $\boldsymbol{b}$ and $\boldsymbol{b}(G)$ respectively. Finally, we write $-i := N \backslash \{i\}$ and $-G := N \backslash G$.

We use a language with three types of atoms: atoms that express the winner, atoms that express the order of two alternatives in a ballot and atoms that express the order of two alternatives in a preference. The modalities are PDL-style. Like Van Eijck (2013), we use two types of actions: one to express the accessibility relations of the voters (so, doxastic relations), the other to express updates.

**Definition 3.2** (Language)**.** Let $x, y \in X$ be alternatives, and let $i \in N$ be a voter. The language $\mathscr{L}$ is defined as

$$\begin{array}{lcl} \varphi & ::= & \top \mid x \mid x \succ_i^p y \mid x \succ_i^b y \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid [\gamma]\varphi \mid [\alpha]\varphi \\ \gamma & ::= & (U, e) \mid \gamma_1 \cup \gamma_2 \\ \alpha & ::= & i \mid G^* \end{array}$$

The modality $[G^*]$ is associated with 'common knowledge among the agents in $G$', and $(U, e)$ is an update model as defined below. We work with the usual abbreviations: $\bot$ is shorthand for $\neg\top$, $\varphi_1 \vee \varphi_2$ is shorthand for $\neg(\neg\varphi_1 \wedge \neg\varphi_2)$, $\varphi_1 \to \varphi_2$ is shorthand for $\neg\varphi_1 \vee \varphi_2$ and $\langle\alpha\rangle\varphi$ is shorthand for $\neg[\alpha]\neg\varphi$. The weak order $\succeq_i^p$ is defined as:

$$x \succeq_i^p y := \begin{cases} \top & \text{if } x = y \\ x \succ_i^p y & \text{otherwise} \end{cases}$$

$x \succeq_i^b$ is defined analogously. Furthermore, let $b_i = x_1 \succ x_2 \succ \cdots \succ x_m$, then in the language $b_i$ is shorthand for the conjunction of all atoms that express the relative orders of alternatives with respect to $b_i$:

$$b_i = (x_1 \succ_i^b x_2) \wedge (x_2 \succ_i^b x_3) \wedge \cdots \wedge (x_{m-1} \succ_i^b x_m)$$

Since the ballot must be a linear order, the conjunction of atoms uniquely defines a ballot $b_i$. In the same way, if $p_i = x_1 \succ \cdots \succ x_m$, we define $p_i$ in the language as follows:

$$p_i = (x_1 \succ_i^p x_2) \wedge (x_2 \succ_i^p x_3) \wedge \cdots \wedge (x_{m-1} \succ_i^p x_m).$$

Although we overload notation for $b_i$ and $p_i$, because $b_i$ and $p_i$ both may refer to a linear order of alternatives (a semantic object) and to a sentence in the logic (a syntactic object), this will not cause confusion. Since we are only working with finite number of voters and alternatives, there is a canonical way to translate formulas into ballots and preferences or vice versa, provided that the formula has the appropriate shape. It will be clear from the context whether we talk about the linear orders or the logic.

The update model should capture two types of dynamics: not only the model should be able to update the agent's knowledge, the model also has to capture that voters might want to change their ballots when they have an incentive to manipulate. Let $B$ denote the set of all ballot atoms

(i.e., atoms of the form $x \succ_i^b y$). Let $B_i \subseteq B$ denote the set of all ballot atoms of voter $i$. If a voter changes her ballot, we model this by defining postconditions for every ballot atom. The postconditions turn some ballot atoms 'on' (namely, exactly the ballot atoms that define the new ballot), and some ballot atoms 'off' (namely, the ballot atoms that are the reverse of the new ballot). Of course, the new ballot should also be a strict linear order. This is all formalised in the following definition.

**Definition 3.3** (Update model for strategic voting)**.** An *update model for strategic voting* is a tuple $U = (\Sigma, R, pre, post)$, where

(i) $\Sigma$ is a non-empty set of events

(ii) $pre : \Sigma \to \mathscr{L}$ assigns a precondition to each event

(iii) $post : \Sigma \to (B \to B \cup \{\top, \bot\})$ assigns a postcondition to each event. A postcondition is a function that maps a ballot atom to either $\top$, $\bot$ or itself[2]. Postconditions must satisfy the following conditions: let $e \in \Sigma$, take $i \in N$ and any alternatives $x, y, z \in X$. Then

(a) if $post(e)(x \succ_i^b y) \neq \top$ and $post(e)(x \succ_i^b y) \neq \bot$, then $post(e)(x \succ_i^b y) = x \succ_i^b y$
(b) if $post(e)(x \succ_i^b y) = x \succ_i^b y$, then for all $x', y' \in X$, $post(e)(x' \succ_i^b y') = x' \succ_i^b y'$
(c) if $x = y$, then $post(e)(x \succ_i y) \neq \top$
(d) if $post(e)(x \succ_i^b y) = \top$ and $post(e)(y \succ_i^b z) = \top$, then $post(e)(x \succ_i^b z) = \top$
(e) if $x \neq y$, then $post(e)(x \succ_i^b y) = \top \iff post(e)(y \succ_i^b x) = \bot$

(iv) $R : N \to \wp(\Sigma \times \Sigma)$ assigns an accessibility relation[3] to every agent $i \in N$, such that $eR_i f$ implies that the following holds: for all $x \succ_i^b y \in B_i$, $post(e)(x \succ_i^b y) = post(f)(x \succ_i^b y)$

An update model for strategic voting $(U, e^*)$ with an actual event $e^* \in \Sigma$ is called an *update*. An update model with a singleton set of events, accessible to all agents, and precondition $\top$, is a *public assignment*. An update model with a singleton set of events, accessible to all agents, and the identity function as postcondition, is a *public announcement*.

The semantics for our language $\mathscr{L}$ is given in the following definition:

**Definition 3.4** (Semantics)**.** Let $F$ be a social choice function and let $M = (S, V, R)$ be an epistemic model for strategic voting for $F$. Then we define

$$
\begin{array}{lcl}
M, s \vDash \top & \iff & \text{always} \\
M, s \vDash x \succ_i^p y & \iff & x \succ y \in \mathsf{proj}_1(V(s))(i) \\
M, s \vDash x \succ_i^b y & \iff & x \succ y \in \mathsf{proj}_2(V(s))(i) \\
M, s \vDash x & \iff & F(\mathsf{proj}_2(V(s))) = x \\
M, s \vDash \neg\varphi & \iff & M, s \nvDash \varphi \\
M, s \vDash \varphi_1 \wedge \varphi_2 & \iff & M, s \vDash \varphi_1 \text{ and } M, s \vDash \varphi_2 \\
M, s \vDash [U, e]\varphi & \iff & \text{if } M, s \vDash pre(e), \text{ then } M \otimes U, (s, e) \vDash \varphi \\
M, s \vDash [\gamma_1 \cup \gamma_2]\varphi & \iff & M, s \vDash [\gamma_1]\varphi \text{ and } M, s \vDash [\gamma_2]\varphi \\
M, s \vDash [i]\varphi & \iff & \text{for all } s' \text{ with } (s, s') \in R(i) : M, s' \vDash \varphi \\
M, s \vDash [G^*]\varphi & \iff & \text{for all } s' \text{ with } (s, s') \in R^*(G) : M, s' \vDash \varphi.
\end{array}
$$

Here, $R^*$ denotes the reflexive and transitive closure of $R$.

[2] each postcondition is required to be only finitely different from the identity $\mathsf{id}$; this condition is met since there are always finitely many ballot atoms

[3] We do not impose further restrictions on the accessibility relations of the voters in $N$. However, the accessibility relations of all update models that will be used in this thesis are serial, transitive and Euclidean.

We now define the effect of the execution of an update model on an epistemic model for strategic voting.

**Definition 3.5** (Product update)**.** Given an epistemic model for strategic voting $M = (S, V, R^M)$ with actual state $s^*$ and an update $(U, e^*)$ with $U = (\Sigma, R^U, pre, post)$ such that $M, s^* \vDash pre(e^*)$, the result of updating $(M, s^*)$ with $(U, e^*)$ is the epistemic model for strategic voting $M \otimes U = (S', R', V')$ with actual state $s'^*$, where

(i) $S' = \{(s, e) \mid M, s \vDash pre(e)\}$

(ii) $V' : S' \to \mathcal{P} \times \mathcal{B}$ with $V'((s, e)) = (\mathsf{proj}_1(V(s)), \{x \succ_i^b y \mid M, s \vDash post(e)(x \succ_i^b y)\})$

(iii) $R'(i) = \{((s, e), (t, f)) \mid (s, e), (t, f) \in S' \text{ and } (s, t) \in R^M(i) \text{ and } (e, f) \in R^U(i)\}$

(iv) $s'^* = (s^*, e^*)$

The atomic valuation $V$ of an updated model $M \otimes U$ is constructed as follows: for each new state $(s, e)$, it simply takes the preference profile of the world $s$, since preferences never change. For the ballot profile, it takes all those ballot atoms whose postcondition at $e$ holds in the initial model at $s$. Given the restrictions on the postcondition function of an update model, such set does define a ballot profile, and thus the valuation function in the resulting model is of the appropriate form.

**Proposition 3.6.** The result of updating an epistemic model for strategic voting with an update model, is again an epistemic model for strategic voting.

*Proof.* Let $M = (S, V, R^M)$ be an arbitrary epistemic model and $U = (\Sigma, R^U, pre, post)$ an arbitrary update model. We have to show that $M \otimes U = (S', V', R')$ is an epistemic model for strategic voting. The non-trivial part is proving that $V'$ is a well-defined valuation on $M \otimes U$ and that the relation $R'$ satisfies the requirements.

To show that $V'$ is well-defined, we have to check that linearity of ballot orders still holds. Let $(s, e) \in M \otimes U$. Let $x, y \in X$ two alternatives, and let $i \in N$. Now suppose that $post(e)(x \succ_i y) = x \succ_i y$. Then it must hold that for every $x', y' \in X$, $post(e)(x \succ_i^b y) = x \succ_i y$. Then $V'(s, e) = (\mathsf{proj}_1(V(s)), \{x \succ_i^b y \mid M, s \vDash post(e)(x \succ_i^b y)\})$. Note that $M, s \vDash post(e)(x \succ_i^b y$ if and only if $M, s \vDash x \succ_i^b y$, and hence $V'(s, e) = (\boldsymbol{p}_s, \boldsymbol{b}_s)$, so in this case $V$ is well-defined.

If $post(e)(x \succ_i^b y) \neq x \succ_i y$, we have that for every $x', y' \in X$, either $post(e)(x' \succ_i^b y') = \top$ or $post(e)(x' \succ_i^b y') = \bot$. If we have $x'$ and $y'$ such that $x = y$, then $post(e)(x' \succ_i^b y') = \bot$, which implies that $M \otimes U, (s, e) \nvDash x \succ_i^b y$, so irreflexivity is guaranteed. Furthermore, for any $x', y', z' \in X$ we have that $post(e)(x' \succ_i^b y') = \top$ and $post(e)(y' \succ_i^b z') = \top$, then $post(e)(x' \succ_i^b z') = \top$. This implies that since $M, s \vDash \top$, in particular $M, s \vDash post(e)(x \succ_i^b z$. Hence, if $M \otimes U, (s, e) \vDash x \succ_i^b y$ and $M \otimes U, (s, e) \vDash y \succ_i^b z$, then $M \otimes U, (s, e) \vDash x \succ_i^b z$, so this guarantees transitivity. Finally, since $x \neq y$, then $post(e)(x \succ_i^b y) = \top \iff post(e)(y \succ_i^b x) = \bot$, we have that the ballot order of $i$ in $(s, e)$ must be total. It follows that $V'$ must assign a linear ballot order for $i$ to state $(s, e)$. Since $i$ was arbitrary, this holds for every voter.

It is left to show that $R'$ is well-defined. Let $(s, e), (t, f) \in S'$ such that $((s, e), (t, f)) \in R'(i)$ for some $i \in N$. We have to show that $\boldsymbol{b}_{(s,e)}(i) = \boldsymbol{b}_{(t,f)}(i)$ and $\boldsymbol{p}_{(s,e)}(i) = \boldsymbol{p}_{(t,f)}(i)$. Since $(s, e) R'_i (t, f)$, it must hold that $s R_i^M t$ and $e R_i^U f$. Thus, $\boldsymbol{b}_s(i) = \boldsymbol{b}_t(i)$ and $\boldsymbol{p}_s(i) = \boldsymbol{p}_t(i)$. The valuations are given by $V'((s, e)) = (\mathsf{proj}_1(V(s)), \{x \succ_i^b y \mid M, s \vDash post(e)(x \succ_i^b y)\})$ and $V'((t, f)) = (\mathsf{proj}_1(V(t)), \{x \succ_i^b y \mid M, t \vDash post(f)(x \succ_i^b y)\})$. Since $\boldsymbol{p}_s(i) = \boldsymbol{p}_t(i)$, it follows that

$\boldsymbol{p}_{(s,e)}(i) = \boldsymbol{p}_{(t,f)}(i)$.

Since $eR_i^U f$, it holds that for every $x \succ_i^b y \in B_i$, $post(e)(x \succ_i^b y) = post(f)(x \succ_i^b y)$. Furthermore, $M, s \vDash x \succ_i^b y$ if and only if $M, t \vDash x \succ_i^b y$. Hence, $\mathsf{proj}_2(V'(s,e))(i) = \{x \succ_i^b y \mid M, s \vDash post(e)(x \succ_i^b y)\} = \{x \succ_i^b y \mid M, s \vDash post(f)(x \succ_i^b y)\} = \{x \succ_i^b y \mid M, t \vDash post(f)(x \succ_i^b y)\} = \mathsf{proj}_2(V'(t,f))(i)$, as was required. □

Composition of two update models is not part of the language. We define composition of two update models semantically:

**Definition 3.7** (Composition of update models)**.** Let $U = (\Sigma, R, pre, post)$ and $U' = (\Sigma', R', pre', post')$ be update models with actual events $e^*$ and $e'^*$ respectively. The composition $(U, e^*); (U', e'^*)$ of these update models is $(U'', e''^*)$, where $U'' = (\Sigma'', R'', pre'', post'')$ is defined by

(i) $\Sigma'' = \Sigma \times \Sigma'$

(ii) $R''(i) = \{((f, f'), (g, g')) \mid (f, g) \in R(i) \text{ and } (f', g') \in R'(i)\}$

(iii) $pre''(f, f') = pre(f) \wedge [U, f]pre'(f')$

(iv)
$$post''(f, f')(x \succ_i^b y) = \begin{cases} post(f)(x \succ_i^b y) & \text{if } post'(f')(x \succ_i^b y) = x \succ_i^b y \\ post'(f')(x \succ_i^b y) & \text{otherwise} \end{cases}$$

and the actual event in $U''$ is $e''^* = (e^*, e'^*)$.

Next, we prove that composition of update models is well-defined.

**Proposition 3.8.** $M, s \vDash [(U, e); (U', e')]\varphi \iff M, s \vDash [U, e][U', e']\varphi$

*Proof.* Let $M = (S, V, R)$ be an arbitrary model with actual state $s$. It suffices to show that $(M \otimes U) \otimes U'$ is isomorphic to $M \otimes (U; U')$. A detailed proof for purely epistemic update can be found in Van Ditmarsch et al. (2007). The postconditions only play a part in the proof that the valuations correspond. Let $V'$ be the valuation of $M \otimes U$, $V_1$ be the valuation of $(M \otimes U) \otimes U'$ and $V_2$ the valuation of $M \otimes (U; U')$. We show that $V_1$ and $V_2$ correspond.

Let $((s, e), e')$ be a state in $(M \otimes U) \otimes U'$. We have that

$$V'(s, e) = (\mathsf{proj}_1(V(s)), \{x \succ_i^b y \mid M, s \vDash post(e)(x \succ_i^b y)\}).$$

If $post'(e')(x \succ_i^b y) = x \succ_i^b y$, it holds that

$$\begin{aligned} V_1((s, e), e') &= (\mathsf{proj}_1(V'(s, e)), \{x \succ_i^b y \mid M \otimes U, (s, e) \vDash post'(e')(x \succ_i^b y)\}) \\ &= (\mathsf{proj}_1(V(s)), \{x \succ_i^b y \mid M \otimes U, (s, e) \vDash x \succ_i^b y\}) \\ &= (\mathsf{proj}_1(V(s)), \{x \succ_i^b y \mid M, s \vDash post(e)(x \succ_i^b y)\}). \end{aligned}$$

In that case, we have

$$\begin{aligned} V_2(s, (e, e')) &= (\mathsf{proj}_1(V(s)), \{x \succ_i^b y \mid M, s \vDash post''(e, e')(x \succ_i^b y\}) \\ &= (\mathsf{proj}_1(V(s)), \{x \succ_i^b y \mid M, s \vDash post(e)(x \succ_i^b y\}) \\ &= V_1((s, e), e'). \end{aligned}$$

If $post'(e')(x \succ_i^b y) \neq (x \succ_i^b y)$, it holds that for all $x', y' \in X$, either $post'(e')(x' \succ_i^b y') = \top$ or $post'(e')(x' \succ_i^b y') = \bot$. Hence, we have

$$\begin{aligned} V_1((s,e),e') &= (\mathsf{proj}_1(V'(s,e)), \{x \succ_i^b y \mid M \otimes U, (s,e) \vDash post'(e')(x \succ_i^b y)\}) \\ &= (\mathsf{proj}_1(V(s)), \{x \succ_i^b y \mid post'(e')(x \succ_i^b y) = \top\}) \\ &= (\mathsf{proj}_1(V(s)), \{x \succ_i^b y \mid M, s \vDash post'(e')(x \succ_i^b y)\}). \end{aligned}$$

Furthermore, we have

$$\begin{aligned} V_2(s,(e,e')) &= (\mathsf{proj}_1(V(s)), \{x \succ_i^b y \mid M, s \vDash post''(e,e')(x \succ_i^b y)\}) \\ &= (\mathsf{proj}_1(V(s)), \{x \succ_i^b y \mid M, s \vDash post'(e')(x \succ_i^b y)\}) \\ &= V_1((s,e),e'). \end{aligned}$$

This shows that $V_1 = V_2$. □

In modal logic, the notion of bisimulation is central. We define a bisimulation for epistemic models for strategic voting in the standard way. The only difference is that our bisimulation is working over a different kind of valuation.

**Definition 3.9** (Bisimulation)**.** Let two epistemic models for strategic voting $M = (S, V, R)$ and $M' = (S', V', R')$ be given. A non-empty relation $\mathcal{R} \subseteq S \times S'$ is a *bisimulation* if for all $s \in S$ and all $s' \in S'$ with $(s, s') \in \mathcal{R}$:

(i) $V(s) = V'(s')$

(ii) for all $i \in N$ and all $t \in S$: if $sR_it$, then there is $t' \in S'$ such that $s'R_i't'$ and $(t, t') \in \mathcal{R}$

(iii) for all $i \in N$ and all $t' \in S'$: if $s'R_i't'$, then there exists $t \in S$ such that $sR_it$ and $(t, t') \in \mathcal{R}$.

If there exists a bisimulation between two models $M$ and $M'$, we call $M$ and $M'$ *bisimilar*.

The notion of bisimulation captures the expressivity of the language.

**Theorem 3.10.** *Let $M$ with actual state $s^*$ and $M'$ with actual state $s'^*$ be epistemic models for strategic voting. If $M, s^*$ and $M', s'^*$ are bisimilar, then $M, s^* \vDash \varphi \iff M', s'^* \vDash \varphi$ for every modal formula $\varphi$.*

*Proof.* The proof is standard, by induction on $\varphi$. There are three base cases. For atoms of the form $x \succ_i^p y$ and $x \succ_i^b y$, it follows directly by part (i) of Definition 3.9. For atoms of the form $x$, we have that when two worlds have the same ballot profile, the voting rule assigns the same winner; thus, bisimilar worlds coincide on formulas of the form $x$. The final non-standard part of the proof is to show that formulas of the form $[U, e]\varphi$ are respected by bisimulations. It is well-known that the product update operation preserves bisimulation: if you take two bisimilar models and apply a product update with the same event model on both, the resulting models are also bisimilar (Van Ditmarsch et al., 2007). □

This shows that bisimular models $(M, s^*)$ and $(M', s'^*)$ represent essentially the same epistemic situation.

## 3.2 Manipulation in the dynamic epistemic model for strategic voting

Now we can define an incentive to manipulate in an epistemic model for strategic voting. A voter $i$ has an incentive to manipulate if there is a strategic ballot $b_i$ that (weakly) improves the outcome from the perspective of that voter and voter $i$ *believes* that this is the case. So, *de re* belief of her manipulation is required. For now, we stick to the classical notion of an incentive to manipulate and we assume that the manipulator is naive and does not take in consideration manipulations of other voters. This definition follows the train of thought the notion of a *dominant manipulation of an information set* introduced by Conitzer et al. (2011) and used in Reijngoud and Endriss (2012) and Bakhtiari et al. (2018).

**Definition 3.11** (Incentive to manipulate)**.** In an epistemic model for strategic voting $M = (S, V, R)$ with actual state $s^*$, we say that voter $i$ with preference $p_i$ has a *dominant manipulation* $b'_i \neq p_i$ if:

- There exists a state $s$ with $s^* R_i s$ such that $F(b'_i, \boldsymbol{b}_s(-i)) \succ_{p_i} F(\boldsymbol{b}_s)$
- For every $t$ with $s^* R_i t$, it holds that $F(b'_i, \boldsymbol{b}_t(-i)) \succ_{p_i} F(\boldsymbol{b}_t)$

Voter $i$ has an *incentive to manipulate* if there exists a ballot $b_i$ such that $b_i$ is a dominant manipulation.

As discussed in Subsection 2.1.2, another way to specify the incentives of a voter to manipulate is by looking at her best strategy under the information she has. In Definition 2.9, a voter's best strategy is defined given her information set. We translate this notion and define the best strategy of a voter in an epistemic model for strategic voting:

**Definition 3.12** (Best strategy in an epistemic model for strategic voting)**.** Let $M = (S, V, R)$ be an epistemic model for strategic voting with actual state $s^*$.

$$\mathcal{S}_i(M, s^*) := \mathcal{S}_i(\{\boldsymbol{b}(-i) \mid \text{there exists } s \in S \text{ with } s^* R_i s \text{ and } M, s \vDash \boldsymbol{b}\}, \boldsymbol{p}_{s^*}(i))$$

**Proposition 3.13.** A voter $i$ has an incentive to manipulate in a model $M$ with actual world $s^*$ if and only if $\mathcal{S}_i(M, s^*) \neq \boldsymbol{p}_{s^*}(i)$.

*Proof.* This follows directly from the definitions. □

If a voter $i$ has an incentive to manipulate, we assume that the voter will actually commit to her strategic ballot. This means that she will factually change her ballot. We model this as a single manipulation by $i$ and fully private announcement to a group of agents $G \subseteq N$ with $i \in G$. This means that the commitment of $i$ to her strategic ballot $b_i$ is noticed by every voter in $G$ and this is common knowledge within the group $G$, while the other agents do not suspect anything. This is modelled via the execution of an update model.

**Definition 3.14.** Let $M$ be an epistemic model for strategic voting with actual state $s^*$. Take $i \in N$ and a insincere ballot $b_i \neq p_i$ (here, $p_i$ is the truthful preference of $i$ in $M$, so $M, s^* \vDash p_i$) and a group of voters $G \subseteq N$, with $i \in G$. Recall that we can write $b_i$ in the language as a conjunction of ballot atoms. We say that $U_{b_i^G}$ is an $b_i^G$*-manipulation update model for* $M$ if $U_{b_i^G} = (\Sigma, R, pre, post)$ with

- $\Sigma = \{e_1, e_2\}$

- $R(i) = \{(e_1, e_1), (e_2, e_2)\}$ for every $i \in G$ and $R(j) = \{(e_1, e_2), (e_2, e_2)\}$ for every $j \notin G$
- $pre(e_1) = pre(e_2) = \top$
- For any $j \neq i$, for any ballot atom $x \succ_j^b y \in B_j$, $post(e_1)(x \succ_j^b y)$ is the identity function. For any ballot atom $x \succ_i^b y \in B_i$, $post(e_1)(x \succ_i^b y) = \top \iff x \succ_i^b y \in b_i$.

The actual event in $U$ is $e_1$.

In the language, instead of using $[U_{b_i^G}]$, we will write $[b_i^G]$ for short.

**Example 3.15.** Suppose we have voters $N = \{1, 2, 3\}$, three alternatives and voter 1 changes her ballot to $b_1 = a \succ b \succ c$, which is only noticed by herself and voter 2. Then $U_{b_1^{\{1,2\}}}$ is the following update model illustrated in Figure 3.1. Here, we only specify the postconditions that are not the identity function.

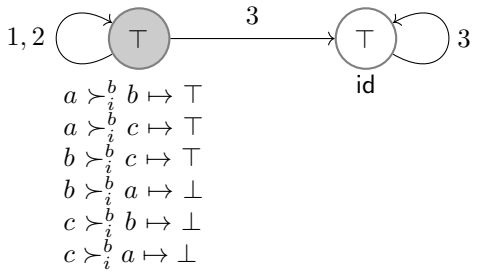



Figure 3.1: The update model $U_{b_1^{\{1,2\}}}$

Intuitively, if two agents $i$ and $j$ ($i \neq j$) change their ballots and a group $G \subseteq N$ with $i \in G$ notices the first change, and $H \subseteq N$ with $j \in H$ the second, then the order in which the agents change their ballots should not matter. The following Proposition shows that this is indeed correct.

**Proposition 3.16.** $M, s \vDash [b_i^G][b_j^H]\varphi \iff M, s \vDash [b_j^H][b_i^G]\varphi$.

*Proof.* See Appendix. □

## 3.3 Expressing manipulation in the language

We give some examples to demonstrate that our language expresses important concepts of strategic voting in a dynamic epistemic setting, introduced in Section 3.1. In the examples discussed below we do not specify $G$, because in the definition of a dominant manipulation of an information set and an incentive to manipulate, we assume that voters only reason about the effect of changing their own vote to the outcome of the election, and the knowledge of other voters is not taken into account.

Recall Definition 3.11: an agent $i$ has a dominant manipulation $b_i$ for her information set if, in at least one accessible world, after changing her ballot to $b_i$, the new winning alternative $x'$ in that world is strictly preferred over the original winner $x$. In every accessible world, changing her ballot to $b_i$ should result in a winner that is at least as good as the original winner $x$. Recall

that $\mathcal{L}(X)$ denotes the finite set of all possible linear orders over the finite set of alternatives, and that $b_i$ (in the language) denotes a conjunction of atoms that indicate the relative order of alternatives in the ballot $b_i$ of $i$. Then the statement that voter $i$ has an incentive to vote $b_i$ is expressed in the language as follows:

$$\boldsymbol{\alpha}(b_i) := \langle i \rangle \Big( \bigwedge_{x \in X} \bigwedge_{x' \in X, x' \neq x} ((x \wedge [b_i^G] x') \to x' \succ_i^p x) \Big) \wedge [i] \Big( \bigwedge_{x \in X} \bigwedge_{x' \in X, x' \neq x} ((x \wedge [b_i^G] x') \to x' \succeq_i^p x) \Big)$$

Voter $i$ has an incentive to manipulate when there exists a dominant manipulation for her information set:

$$\bigvee_{b_i \in \mathcal{B}} \Big( \boldsymbol{\alpha}(b_i) \Big)$$

There is a voter with an incentive to manipulate is expressed as follows:

$$\bigvee_{i \in N} \bigvee_{b_i \in \mathcal{B}} \Big( \boldsymbol{\alpha}(b_i) \Big)$$

If no voter has an incentive to manipulate, the voting situation is strategyproof. In the logic, we express this as

$$\bigwedge_{i \in N} \bigwedge_{b_i \in \mathcal{B}} [i] \Big( \bigwedge_{x \in X} \bigwedge_{x' \in X, x' \neq x} ((x \wedge [b_i^G] x') \to x \succeq_i x') \Big),$$

which is logically equivalent to $\neg \bigvee_{i \in N} \bigvee_{b_i \in \mathcal{L}(X)} \Big( \boldsymbol{\alpha}(b_i) \Big)$.

## 3.4 Axiomatisation

In Table 3.1, a proof system for the dynamic epistemic logic of strategic voting is given.

| **axiom/rule** | **description** |
|---|---|
| all instantiations of propositional tautologies | |
| from $\varphi$ and $\varphi \to \psi$, infer $\psi$ | modus ponens |
| from $\varphi$, infer $[\alpha]\varphi$ | necessitation for belief |
| $[\alpha](\varphi \to \psi) \to ([\alpha]\varphi \to [\alpha]\psi)$ | distribution for knowledge |
| $[G^*]\varphi \to (\varphi \wedge [G][G^*]\varphi)$ | mix |
| $[G^*](\varphi \to [G]\varphi) \to (\varphi \to [G^*]\varphi)$ | induction axiom |
| $x \succ_i^b y \to [i]x \succ_i^b y$ | knowledge of own ballot |
| $x \succ_i^p y \to [i]x \succ_i^p y$ | knowledge of own preference |
| $\neg(x \succ_i^b x)$ | irreflexivity of ballot orders |
| $(x \succ_i^b y \wedge y \succ_i^b z) \to x \succ_i^b z$ | transitivity of ballot orders |
| $x \succ_i^b y \vee y \succ_i^b x \quad$ if $x \neq y$ | totality of ballot orders |
| $\neg(x \succ_i^p x)$ | irreflexivity of preference orders |
| $(x \succ_i^p y \wedge y \succ_i^p z) \to x \succ_i^p z$ | transitivity of preference orders |
| $x \succ_i^p y \vee y \succ_i^p x \quad$ if $x \neq y$ | totality of preference orders |
| $x \leftrightarrow \bigvee_{\boldsymbol{b} \in \mathcal{B}, F(\boldsymbol{b}) = x} \boldsymbol{b}$ | axioms that characterise $F$ |
| from $\varphi$, infer $[\gamma]\varphi$ | necessitation for updates |
| $[\gamma](\varphi \to \psi) \to ([\gamma]\varphi \to [\gamma]\psi)$ | distribution for updates |
| $[U,e]x \succ_i^p y \leftrightarrow (pre(e) \to x \succ_i^p y)$ | update and preferences |
| $[U,e]x \succ_i^b y \leftrightarrow (pre(e) \to post(e)(x \succ_i^b y)$ | update and ballots |
| $[U,e]x \leftrightarrow \big(pre(e) \to \bigvee_{\boldsymbol{b} \in \mathcal{B}, F(\boldsymbol{b}) = x} \bigwedge_{x \succ_i^b y \in \boldsymbol{b}} [U,e]x \succ_i^b y\big)$ | update and winners |
| $[U,e]\neg\varphi \leftrightarrow (pre(e) \to \neg[U,e]\varphi)$ | update and negation |
| $[U,e]\varphi \wedge \psi \leftrightarrow ([U,e]\varphi \wedge [U,e]\psi)$ | update and conjunction |
| $[U,e][i]\varphi \leftrightarrow (pre(e) \to \bigwedge_{(e,f) \in R(i)} [i][U,f]\varphi)$ | update and knowledge |
| $[U,e][U',e']\varphi \leftrightarrow [(U,e);(U',e')]\varphi$ | update composition |
| Let $(U,e)$ be an update model and let a set of formulas $\chi_f$ be given for every $f$ such that $(e,f) \in R^*(G)$. From $\chi_f \to [U,f]\varphi$ and $(\chi_f \wedge pre(f)) \to [i]\chi_g$ for every $f \in \Sigma$ such that $(e,f) \in R^*(G)$, $i \in G$ and $(f,g) \in R(i)$, infer $\chi_e \to [U,e][G^*]\varphi$. | updates and common knowledge |

Table 3.1: Axiomatisation for the dynamic epistemic logic of strategic voting

Since there are finitely many ballot profiles, one can indeed characterise the winner by listing the ballots in which it wins, according to the chosen social choice function. However, in this

thesis, we will not focus on a particular social choice function, so we will not give an explicit list of axioms.

**Lemma 3.17.** The axioms for knowledge of own ballot and preference, the axioms for linearity of ballot and preferences and the axioms for updates and preferences, ballot and winners are sound with respect to the class of epistemic models for strategic voting.

*Proof.* The axioms for knowledge of own ballot and knowledge of own preference formalise the constraint on epistemic models for strategic voting that $sR_it$ implies that $\mathsf{proj}_1(V(s))(i) = \mathsf{proj}_1(V(t))(i)$ and $\mathsf{proj}_2(V(s))(i) = \mathsf{proj}_2(V(t))(i)$. The axioms for linearity of ballot and preferences are valid since a valuation always assigns a single preference profile and ballot profile to each state. The axiom for update and preferences are valid since preferences never change: if $M$ is a model with valuation $V$, and $M \otimes U$ is a product update of $M$ with any update model $U$ such that $V'$ is the valuation of $M \otimes U$, then $\mathsf{proj}_1(V'(s)) = \mathsf{proj}_1(V(s))$ for every $s \in S$. Hence, preferences never change. The axiom for update and ballots corresponds to the axiom for update and atoms in Van Ditmarsch and Kooi (2006). The axiom for update and winners is similar to this axiom, but extended to winners: $x$ is the winner after an update $U$, if and only if there is a ballot profile $\boldsymbol{b}$ such that $F(\boldsymbol{b}) = x$, and after the update, exactly that ballot profile $\boldsymbol{b}$ holds. □

**Theorem 3.18.** *The axiomatisation for the dynamic epistemic logic of strategic voting is sound and complete with respect to the class of epistemic models for strategic voting.*

*Proof (sketch).* We only show the soundness of the updates and common knowledge rule, as soundness of the rest of the axioms is evident or already shown in 3.17. Let $(U, e)$ be an update model and let a set of formulas $\chi_f$ be given for every $f$ such that $(e, f) \in R^*(G)$. From $\chi_f \to [U, f]\varphi$ and $(\chi_f \wedge pre(f)) \to [i]\chi_g$ for every $f \in \Sigma$ such that $(e, f) \in R^*(G)$, $i \in G$ and $(f, g) \in R(i)$. Now we have to show that $\chi_e \to [U, e][G^*]\varphi$ is valid. Let $M, s$ be an epistemic model and suppose $M, s \vDash \chi_e$. We show that $M, s \vDash [U, e][G^*]\varphi$. This is the case if and only if $M, s \vDash pre(e)$ implies that $M \otimes U, (s, e) \vDash [G^*]\varphi$, so suppose $M, s \vDash pre(e)$. Let $(t, f) \in S \otimes \Sigma$ such that $((s, e), (t, f)) \in (R^{M\otimes U})^*(G)$. This means that there is a path of arbitrary but finite length between $(s, e)$ and $(t, f)$ for agents in $G$. We have to prove that $M \otimes U, (t, f) \vDash \varphi$. We use induction on the length of the path.

If the length of the path is 0, we have $(s, e) = (t, f)$. We have $M, s \vDash \chi_e$ and since $(e, e) \in (R^{M\otimes U})^*(G)$, we also have $M, s\chi_e \to [U, e]\varphi$. Hence, $M, s \vDash [U, e]\varphi$ and therefore $M \otimes U, (s, e) \vDash \varphi$. It follows that $M \otimes U, (t, f) \vDash \varphi$.

Now consider a path of length $n + 1$. Let $i \in G$ and $(u, g) \in M \otimes U$ such that $((s, e), (u, g)) \in R^{M\otimes U}(i)$ and $((u, g), (t, f)) \in (R^{M\otimes U})^*(G)$. Since $(e, g) \in R^{M\otimes U}(i)$, it must hold that $(e, g) \in R^U(i)$. Hence, $\chi_g$ is defined and from the validity $(\chi_e \wedge pre(e)) \to [i]\chi_g$, $M, s \vDash \chi_e$ and $M, s \vDash pre(e)$ we infer that $M, s \vDash [i]\chi_g$. Since $(e, g) \in R^{M\otimes U}(i)$, it must hold that $(s, u) \in R(i)$ and this implies that $M, u \vDash \chi_g$. From $\chi_g \to [U, g]\varphi$, we can induce that $M, u \vDash [U, g]\varphi$ and hence $M \otimes U, (u, g) \vDash \varphi$. We now apply the induction hypothesis of the length $n$ path between $(u, g)$ and $(t, f)$, and therefore $M \otimes U, (t, f) \vDash \varphi$, as required.

This axiomatisation is an extension of the logic of ontic and epistemic change (Van Ditmarsch and Kooi, 2006): we add axioms for knowledge of own ballot and preference, axioms for linearity of ballot and preferences, and we distinguish axioms for updating ballots, preferences and winners. For completeness, we observe that the canonical model to determine the completeness of the logic without dynamics is an epistemic model for strategic voting (Lemma 3.17), and that the completeness of the logic without dynamics is as usual (see Van Ditmarsch and Kooi (2006);

Van Ditmarsch et al. (2007); Baltag et al. (1999)) because every formula is equivalent to one without updates (the axioms are rewriting rules, pushing all logical connectives beyond the modal operators). □

## 3.5 Concluding remarks

The dynamic epistemic logic presented here is able to deal with very complex forms of communication. However, there is one disadvantage of this logic: it is essentially monotonic, because uncertainty can only be decreased, and not increased by an update. This means that no belief revision is allowed. The product update works very well when dealing with 'knowledge', or even with (possibly false) beliefs, as long as these false beliefs are never contradicted by new information. If new information is contradictory with the current belief of an agent, the product update gives non-intuitive results: if an agent $i$ is confronted with a contradiction between previous beliefs and new information she starts to believe the contradiction, and so she gets 'crazy' and starts to believe everything. In this thesis, we only work with classes of epistemic models for strategic voting and update models that are KD45, that is, all accessibility relations are serial, transitive and Euclidean. Furthermore, in this thesis we assume that only true information is communicated, and only true commitment to a new ballot is considered. Hence, agents will not have to deal with false information and inconsistent beliefs. Therefore, the absence of belief revision will not cause any problems.

# Chapter 4

# Naive manipulation

In this section, we will discuss some important results from voting theory and show how they fit into our framework. We will illustrate that the presented framework is general enough to model different voting settings. This chapter is focused on the classical setting with a naive manipulators, that is, with manipulators who think that every other voter will report a sincere ballot. We translate the classical setting with full information as well as more recent research on manipulation under partial information to our framework.

## 4.1 Gibbard-Satterthwaite manipulation

In the Gibbard-Satterthwaite theorem, it is assumed that every voter has full information about every other voter's preference. Furthermore, it is assumed that every voter (initially) casts a sincere vote. The class of models that meet these conditions is defined as follows:

**Definition 4.1.** The class of epistemic models for strategic voting $\mathcal{M}_{\text{full info,sincere}}$ consist of all models $M = (S, V, R)$ that satisfy the following conditions:

- $S = \{s^*\}$
- $M, s^* \vDash x \succ_i^b y \iff M, s^* \vDash x \succ_i^p y$
- for all $i \in N$, $s^* R_i s^*$

We rephrase the Gibbard Satterthwaite theorem as follows (Gibbard, 1973; Satterthwaite, 1975):

**Theorem 4.2** (Gibbard-Satterthwaite)**.** *Let $N$ be the set of voters that have to make a choice between $m$ alternatives. Let $F$ be a resolute, non-dictatorial and surjective social choice function. Then there exists a model $M \in \mathcal{M}_{full\ info,sincere}$ for $F$ such that there exists a voter $j \in N$ with an incentive to manipulate.*

## 4.2 Manipulation under partial information

On the other side, there is a situation in which the voters have no information at all about their fellow voters. If every voter votes truthfully, only knows her own preference and has absolutely no information about the preferences of other voters, and this is common knowledge, we obtain a model with $m!^n$ worlds (since there are $m!$ possible preference orders, and hence $m!^n$ possible preference profiles) where two states $s$ and $t$ are $i$-related if and only if the ballots and preferences of voter $i$ in $s$ and $t$ are identical.

**Definition 4.3.** Let $\mathcal{M}_{\text{ign,sincere}}$ denote the class of 'ignorant' models, namely, models $M = (S, V, R)$ that satisfy the following conditions:

- $|S| = n!^m$
- for all $\boldsymbol{p} \in \mathcal{P}$, there exists $s \in S$ with $\mathsf{proj}_1(V(s)) = \boldsymbol{p}$
- for all $i \in N$, it holds that $sR_it \iff \boldsymbol{p}_s(i) = \boldsymbol{p}_t(i)$ and $\boldsymbol{b}_s(i) = \boldsymbol{b}_t(i)$
- for all $s \in S$, $M, s \vDash x \succ_i^b y \iff M, s \vDash x \succ_i^p y$.

For many important voting rules, it is not possible to manipulate in this situation, because voters have too little information. Conitzer et al. (2011) showed this for Condorcet extensions and positional scoring rules. Here, we rephrase their results in our framework.

**Theorem 4.4.** *Let $F$ be a Condorcet-consistent social choice function. Then in every $M \in \mathcal{M}_{ign,sincere}$ for $F$ it holds that there is no voter with an incentive to manipulate.*

**Theorem 4.5.** *If $F$ is the Borda rule, in every $M \in \mathcal{M}_{ign,sincere}$ for $F$ it holds that there is no voter with an incentive to manipulate.*

**Theorem 4.6.** *If $F$ is a positional scoring rule and in every model $M \in \mathcal{M}_{ign,sincere}$ for $F$ and $n \geq 6(m-2)$, it holds that there is no voter with an incentive to manipulate..*

Albeit it is not realistic to assume full information, nor is it realistic to assume no information at all. This is why partial information is more interesting. Reijngoud and Endriss (2012) analyse the strategic behaviour of voters under different types of partial information. If agents have partial information, the model grows from a single-state model to a multi-state model. The framework of Reijngoud and Endriss (2012) can be interpreted as a voting situation in which a single agent responds to an opinion poll. The opinion poll reveals a certain type of information, for example, it reveals the winner under the truthful profile, or it reveals the weighted majority graph. We discuss one susceptibility and one immunity result. Reijngoud and Endriss show that any unanimous positional scoring rule is susceptible to winner-manipulation. We first define the class of models in which voters have winner information. This means that they know the winner under the truthful profile, but nothing more then that (except for their own preference and ballot). Suppose that the winner under the truthful profile is $x$. Then the possible ballot profiles are all ballot profiles under which $x$ would be the winner.

**Definition 4.7.** The class of models $\mathcal{M}_{\text{winner,sincere}}$ is defined as the set of models $M = (S, V, R)$ that satisfy the following conditions: there exists $x \in X$ such that

- $|S| = |\{\boldsymbol{b} \mid F(\boldsymbol{b}) = x\}|$
- for all $\boldsymbol{b} \in \mathcal{B}$ with $F(\boldsymbol{b}) = x$, there exists $s \in S$ with $\mathsf{proj}_2(V(s)) = \boldsymbol{b}$
- for all $i \in N$, it holds that $sR_it \iff \boldsymbol{p}_s(i) = \boldsymbol{p}_t(i)$ and $\boldsymbol{b}_s(i) = \boldsymbol{b}_t(i)$
- for all $s \in S$, $M, s \vDash x \succ_i^b y \iff M, s \vDash x \succ_i^p y$.

The result translates to our framework in the following way:

**Theorem 4.8.** *When $m > 3$ and $n > 4$, and $F$ is a unanimous positional scoring rule (paired with the lexicographic tie-breaking rule) for $N$ and $X$, then there exists a model $M \in \mathcal{M}_{winner,sincere}$ for $F$ such that there is a voter $i$ with an incentive to manipulate.*

An important immunity result from Reijngoud & Endriss is that anti-plurality is immune to manipulation under winner information (when the number of voters is large enough). We rephrase it as follows:

**Theorem 4.9.** *When $n > 2m - 2$, and $F$ is the anti-plurality rule paired with the lexicographic tie-breaking rule for $N$ and $X$, then there is no model $M \in \mathcal{M}_{winner,sincere}$ for $F$ in which a voter has an incentive to manipulate.*

The restatement of some well-known results in voting theory demonstrates that in the classical framework, there are many implicit assumptions: it is generally assumed that there is just a single manipulator reasoning from an information set of possible preference profiles. Since there is only one manipulator, we do not have to consider the epistemic states of the other agents: no matter how much information they have, they will always vote truthfully. Therefore, the manipulative voter does not have to worry about the behaviour of the other voters: she only cares about how her (possibly untruthful) ballot can affect the outcome of the election. These assumptions are very strong. It is more realistic to consider situations in which multiple (groups of) voters try to manipulate the election. In such situations, voters have to reason about the manipulative behaviour of other voters, so we have to model higher-order reasoning, as we will do in the following chapters.

# Chapter 5

# Manipulation under higher-order reasoning

In the previous chapter, we discussed naive manipulation, i.e., cases where just a single voter votes strategically and the remaining voters are sincere. However, it is likely that other voters may consider a strategic vote as well. Moreover, voters could realise that other voters may reason strategically too, and therefore choose the best strategy accordingly. A natural question that arises here is that what will happen if all voters behave strategically and all of them know that too. In this case, the strategy of each agent depends on the strategy of other agents. In this chapter, we study the reasoning and voting behaviour of sophisticated voters. We do this by looking at a voter's higher-order reasoning, which arises when a voter recognizes that the other voters reflect on her uncertainty about their uncertainty, and so on.

To illustrate the impact of higher-order reasoning on an agent's strategy, we first introduce a simple example, which is a variant of the well-known Battle of the sexes game. A couple, Alice and Bob, has agreed to meet this evening. However, they have not yet decided where to go. Alice would prefer to go to the football game, Bob would rather go to the opera. The third option is staying at home. Both would prefer to go somewhere rather than not going at all, and this is all common knowledge. If they vote for the same activity, they will go there. If they choose different activities, they are not going anywhere. What should they choose? Initially, it seems reasonable for Alice to vote for the football game, because that is her top choice. However, if she thinks that Bob will go for his top choice as well, the opera, they will end up going nowhere, which is her least favourite outcome. Therefore, it is better for Alice to vote for the opera as well. But, what if Bob also realised this and therefore decided to vote for the football game? In that case, Alice should vote for the football game as well, because then they will go to the game. However, if Bob reasons in the same way, he thinks that she will vote for the opera, and therefore he will also vote for the opera. This reasoning process can theoretically be continued indefinitely. Depending on the level of reasoning that Alice *thinks* that Bob applies, her best strategy would either be to vote for the football game (a sincere vote) or the opera (a strategic ballot). The same holds for Bob. So, whether Alice and Bob have an incentive to vote insincerely, really depends on their level of higher-order reasoning.

In this chapter, we will analyse voting procedures in which all voters vote simultaneously: all voters have one chance to report a ballot, so voting is a 'one-shot' event. In this situation, voters have to determine the votes of other voters by applying higher-order reasoning before they cast

a vote, and determine their own best strategy based on what they think that the other agents will do. In the first section, we will discuss the notion of *safe manipulation*, introduced by Slinko and White (2014). We show why voters have to apply some form of higher-order reasoning in order to be able to cast a safe strategic vote. The rest of the chapter will be devoted to a more general notion of higher-order reasoning, based on the principles of cognitive hierarchy theory. The cognitive hierarchy model has recursively defined strategic categories: a level-$k$ reasoning agents think that all other agents reason at level $k-1$. So, each player assumes that her strategy is the most sophisticated. This model was first introduced by Stahl and Wilson (1994, 1995) and Nagel (1995). Using this behavioural model to analyse strategic behaviour from the perspective of computational social choice was first done by Terzopoulou (2017), and we will translate some of her results to our framework.

## 5.1 Safe manipulation

Safe manipulation was introduced by Slinko and White (2014). In many cases, a voter would like to change the outcome of the election, but she is not able to achieve this on her own (she is not *pivotal*). However, she may still cast a strategic vote and hope that other voters with a preference order identical to hers act in the same way. If the right set of like-minded fellow voters follow her, they may be able to change the outcome in their favour as a group. There can be cases where it really matters which set of voters joins the manipulative coalition: if the wrong set of voters decide to cast a strategic vote, the collective manipulation results in an outcome that is *worse* (from their perspective) than if they had not strategised. In such scenarios the strategic vote is called unsafe.

When determining whether a strategic vote is safe, voters also use higher-order reasoning, but restricted to fellow voters with the same preference order. If a group of voters with identical preferences has a shared incentive to manipulate, and there is no coordination in the group, it is possible that they overshoot or undershoot. In that case, the outcome of the election will be worse than the outcome under the truthful profile.

We assume that the strategic voters all have the same sincere preferences and all contemplate casting the same strategic vote, while all other voters are not strategic. So, every voter with a preference different from the preference order of the manipulative group, will just vote truthfully. The voters in the manipulative group are higher-order reasoners: each voter in this group knows that the other voters in the group have an incentive to manipulate, and each voter also knows that the other voters know that she has an incentive to manipulate, etcetera. However, in order to manipulate the election in a way that the outcome is beneficial for them, they have to coordinate: the right set of voters in the group should cast a strategic vote.

**Example 5.1.** Recall the example of the introduction. Four friends have to decide if they go on a vacation to Austria, to the Bahamas or to China. Their preferences are given by:

| | |
|---|---|
| 1 | $abc$ |
| 2 | $abc$ |
| 3 | $bca$ |
| 4 | $cba$ |

and the rule to be used is Borda with tie-breaking order $a \triangleright b \triangleright c$. If everybody votes sincerely,

then $b$ is elected. Voters 3 and 4 are non-manipulative, and this is common knowledge. Suppose that voters 1 and 2 are higher-order reasoners and that they know each other's preferences, and this is also common knowledge. Voter 1 can make $a$ win by voting $a \succ c \succ b$ and voter 2 can do the same. However, if they both try to manipulate, their worst alternative $c$ will become the winner. They both know that it is best for both of them if one of the agents casts a strategic vote, while the other votes truthfully. Now suppose that voter 1 realises this: if she thinks that voter 2 will manipulate the election, it is better for her to vote according to her truthful preference. However, if she thinks that voter 2 thinks that voter 1 will manipulate the election, it is better to cast the strategic vote, as voter 2 (by her reasoning) will refrain from strategising. This reasoning process can be continued, because what if in fact voter 1 thinks that voter 2 thinks that voter 1 thinks that voter 2 will cast a strategic vote?

A manipulation is safe when regardless the set of like-minded voters that join in the act of manipulating, the outcome will not be worse than the outcome under the truthful profile. Thus, even if a voter is uncertain about which fellow voters with identical preference will cast the same strategic vote, the outcome must be at least as good as under the truthful profile. Safe manipulation is only defined for scenarios in which all voters have full information and vote sincerely, and this is common knowledge. To define safe manipulation formally, we introduce an update model in which the manipulative voters are uncertain about the set of voters who cast a strategic vote.

**Definition 5.2** (Uncoordinated coalitional manipulation update model)**.** Let $M \in \mathcal{M}_{\text{full info, sincere}}$. Let voter 1 be a voter with preference $p_1$ and an incentive to strategically vote $b'$. Let $G \subseteq N$ be the set of voters with preference $p_1$ (so $1 \in G$). Then every voter in $G$ has an incentive to strategically vote $b'$. Now we define an *uncoordinated coalitional manipulation update model* for $G$ and $b'$ as $U^G_{b'} = (\Sigma, R^U, pre, post)$, with:

(i) Each event represents a subset of $G$: $\Sigma = \wp(G)$. Let $e_{G_1}, \ldots, e_{G_{2^{|G|}}}$ denote the states of $U$.

(ii) For $k \in \{1, \ldots, 2^{|G|}\}$, in event $e_{G_k}$, the subset of voters $G_k$ change their ballots to $b'$. For every $i \in G_k$, and every $x \succ y \in b'$:

$$post(e_{G_k})(x \succ^b_i y) = \top \text{ and } post(e_{G_k})(y \succ^b_i x) = \bot$$

For every $j \notin G_k$, and every $x, y \in X$:

$$post(e_{G_k})(x \succ^b_j y) = x \succ^b_j y$$

(iii) No information is shared: $pre(e_{G_k}) = \top$ for $k \in \{1, \ldots, 2^{|G|}\}$

(iv) Every voter only knows whether she changes her own ballot:
for every $i \in N$: $eR^U_i f \iff$ for all $x \succ^b_i y \in B_i$: $post(e)(x \succ^b_i y) = x \succ^b_i y$

(v) $e_{G_k}$ with $G_k = \emptyset$ is the actual world

We say that a manipulation $b'$ is safe for voter $i$, if the initial model (in which every agent has full information and votes sincerely) updated with the corresponding uncoordinated coalitional manipulation update model entails an incentive to manipulate for voter $i$. This is formally defined as follows:

**Definition 5.3.** Let $M \in \mathcal{M}_{\text{full info, sincere}}$ with $M, s^* \vDash y$, and let $i \in N$ be a voter with an incentive to strategically vote $b'$. Let $U^G_{b'}$ an uncoordinated coalitional manipulation update model. Then voter $i$ has a *safe manipulation* $b'$ if for every $(s, e) \in M \otimes U^G_{b'}$ with $(s^*, e^*)R_i(s, e)$, if $M \otimes U^G_{b'}, (s, e) \vDash x$, then $M, s \vDash x \succeq^p_i y$.

Note that the updated model $M \otimes U_{b'}^G$ is no longer in the class of models $\mathcal{M}_{\text{full info,sincere}}$. We continue with the example from the introduction of this section to show how this definition can be applied.

**Example 5.1** (continued). The initial epistemic model for strategic voting (with full information and every voter casting a sincere ballot) is shown in Figure 5.1. In this model, no voter attempted to manipulate yet.

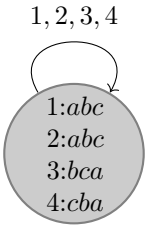



Figure 5.1: The epistemic model $M$ for Borda, with four voters and three alternatives

The uncoordinated coalitional manipulation update model for $G = \{1, 2\}$ and $b' = acb$ is given in Figure 5.2. Here, we abuse notation and use $a \succ_i^b c \succ_i^b b \mapsto \top$ to summarise the postconditions $a \succ_i^b c \mapsto \top$, $a \succ_i^b b \mapsto \top$, $c \succ_i^b b \mapsto \top$, $c \succ_i^b a \mapsto \bot$, $b \succ_i^b a \mapsto \bot$ and $b \succ_i^b c \mapsto \bot$.

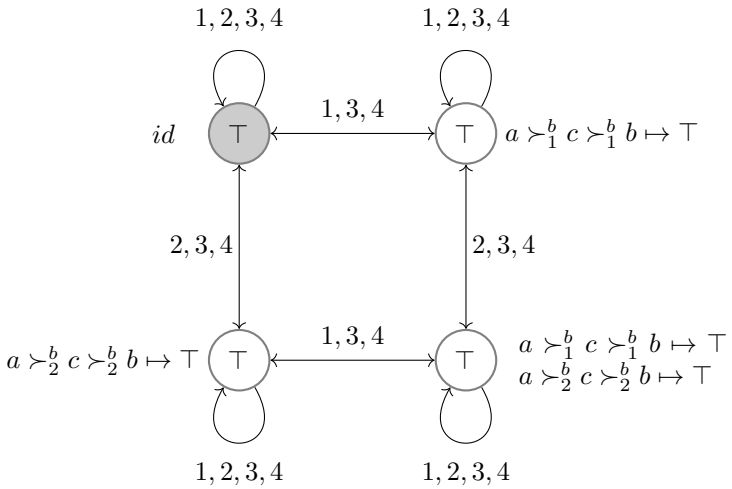



Figure 5.2: The uncoordinated coalitional manipulation update model for $G = \{1, 2\}$ and $b' = acb$

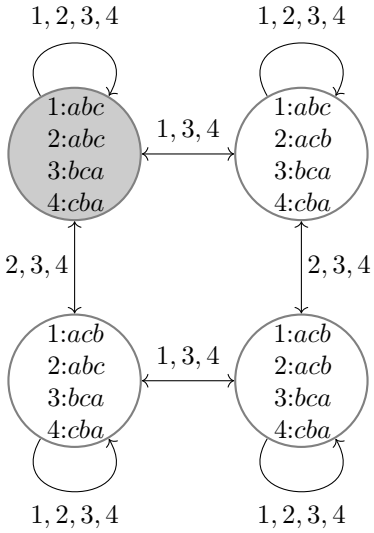



Figure 5.3: The model $M \otimes U_{b'}^{G}$

The result of taking the product update $M \otimes U_{b'}^{G}$ is given in Figure 5.3. In this model, voter 1 and 2 do not have an incentive to manipulate, because they both consider it possible that the other voter strategises, and in that case it is harmful to cast a strategic vote. Thus, this manipulation is unsafe.

Which voting rules are susceptible to safe manipulation? Slinko and White (2014) proved that even if we restrict the possible manipulative ballots to safe manipulations, every non-dictatorial and surjective voting rule is susceptible to manipulation:

**Theorem 5.4.** *Suppose that $F$ is a surjective and non-dictatorial social choice function for $m \geq 3$. Then there exists a model $M \in \mathcal{M}_{\text{full info, sincere}}$ such that there is a voter with a safe manipulation.*

This result can be seen as a strengthening of the Gibbard-Satterthwaite theorem. The natural question that arises is what happens if agents not only consider it possible that like-minded fellow voters will cast the same strategic ballot, but when they start reasoning about the strategic behaviour of all their fellow voters, who might have conflicting interests. In the next section, we will set up the formal framework to answer that question.

## 5.2 Level-$k$ reasoning

The framework introduced in Chapter 3 can be used to analyse very complex voting situations. We could assume that voters are cognitively very capable and that they are able to apply arbitrary levels of higher-order reasoning. They would not only be able to reason about other voters' voting behaviour (like manipulation), but also about the interaction between manipulations of multiple voters at the same time, and they would be able to analyse whenever it would be strategic to share information with other agents. When the number of voters, the number of alternatives and the cognitive levels of voters increase, the reasoning process becomes very complex.

Are real voters capable of such complex forms of higher-order reasoning? Behavourial game theorists study how people in reality actually reason, by conducting experiments in which people interact strategically (Camerer et al., 2004; Costa-Gomes et al., 2001; Camerer, 2003). Theoretical and experimental studies show that human beings are not perfectly rational reasoners, but that on the contrary, our rational behaviour endures serious limitations. Although there is no general consensus on the depth of reasoning by humans in games, experimental studies often show that subjects do not use more than three steps of higher-order reasoning. Many results indicate that it would be very unlikely to observe level-4 reasoning or higher (Arad and Rubinstein, 2012; Camerer et al., 2004; Camerer, 2003). In this thesis, we only consider strategic manipulation under finite levels of reasoning.

Agents with a *theory of mind* have the ability to understand that other agents' perspectives may differ from their own, and they are able to attribute mental states to others. When a voter reasons on a higher level about strategic manipulation in an election, her picture of the situation not only includes information about her own preference and ballot, but also a 'mental model of other minds'. These mental models of other people's minds often have a hierarchical, reflexive structure that is used to reason about strategic situations ('What do you think I think you think...') (Hedden and Zhang, 2002).

Cognitive hierarchy theories capture these mental models of other minds by classifying the voters according to their degree of reasoning in forming expectations of others. In the setting of voting, we define level-0 voters as voters who always vote truthfully. Level-1 voters think that every other voter is a level-0 reasoner. So, a level-1 voter determines her best strategy given that everyone else reports a sincere ballot. We can generalise this to level-$k$ voters: a level-$k$ voter ($k \geq 1$) assumes a homogeneous population consisting of voters reasoning at level $k-1$. Thus, higher-level reasoning voters assume that the other voters do fewer reasoning steps than they do: every higher-order reasoning voter thinks that she is the most sophisticated reasoner. In this way, we can distinguish different strategic types of voters. A strategic type characterises the level of strategic sophistication of a voter and is determined by the number of steps of reasoning that the voter performs in a sequence of iterated best strategies.

Suppose that voter $i$ is a second-order reasoning voter. This means that she assumes that all other voters reason at level 1, i.e., that they assume that every other voter votes truthfully. Let $j$ be some other agent. According to agent $i$'s second-order reasoning, agent $j$ thinks that every other voter reports a sincere ballot. Agent $i$ may realise that following $j$'s reasoning under this assumption, voter $j$ has an incentive to cast a strategic vote. Hence, she will not consider the case in which agent $i$ votes truthfully anymore: in order to determine her own best strategy, agent $i$ will only analyse the cases in which $j$ manipulates. We will construct models in which a $k$-level reasoner has simulated the reasoning process of all other voters, who are (in her mind) $k-1$-reasoners. Since we want to analyse strategic behaviour of higher-order reasoning voters, we have to observe whenever a $k$-level reasoner has an incentive to manipulate.

When voter $i$ simulates the strategic reasoning of the other voters in her own mind, she assumes that the other voters are rational, meaning that she thinks that they will determine their vote in the same way as she does. In Definition 2.9 we defined the procedure that is used by voters to pick a ballot (or in other words, best strategy). A higher-order reasoner thinks that every other voter thinks that every voter chooses her best strategy rationally, that every voter thinks that every voter thinks that every other voter chooses her best strategy rationally, and so on. This assumption is called *common belief in rationality* and is extensively discussed in Perea (2012).

To formalise this, we will use a semantic approach, because our main focus is the semantic interaction of knowledge and voting, not the logic. We think that the semantic approach is more intuitive and comprehensible than expressing notions of higher-order reasoning in the language, as expressing first-order notions already results in complicated and long formulas. We recursively define an epistemic model for strategic voting with a $k$-level reasoner. We will do this by first defining models in which voters apply level-zero reasoning, which means that every voter will report a sincere ballot. In this section, we will make the extra assumption that in the initial model, the voters have no incorrect information or false beliefs about each others' preferences or about other voters' knowledge. This implies that every voters considers the actual state possible, and that we are working with $S5$ models as initial level-zero models. The reason is twofold: first, it is natural to assume that if voters are level-zero reasoners, they have no reason to care about (possibly false) information about the knowledge and preferences of other voters, because in any case, they will just vote truthfully. The second reason is technical: if in the initial model, voters have false beliefs, this will lead to very complicated higher level models. Higher-level reasoning agents 'simulate' the reasoning process of other agents to deduce their decisions. The advantage of starting with an S5 model[1] is that no voter has false beliefs about the knowledge or preferences of other voters, so the reasoning of an agent $j$ about the reasoning of another agent $i$, coincides with the actual reasoning of an agent $i$. This allows us to create level-$k$-models by a relatively simple cut-and-paste operation.

This chapter is mainly based on the Master's Thesis of Terzopoulou (2017). The framework presented here is an extension of the model that is used in Terzopoulou (2017). The main difference is that we do not work with information sets, but dynamic epistemic models. This allows us to consider changes in information (for example as a consequence of a public, semi-private or private announcement) changes in ballots. In Subsection 5.2.3, we will show how some results of Terzopoulou (2017) can be translated to our framework.

### 5.2.1 Level-0 and level-1 models

First, we characterise models in which all voters are level-0 reasoners:

**Definition 5.5** (Level-0 model)**.** An epistemic model for strategic voting $M = (S, V, R)$ is *level-0* if it is an S5 model (this means that for every $i \in N$, $R_i$ is an equivalence relation) and in every state $s \in S$, and every voter $i \in N$, it holds that

$$M, s \vDash x \succ_i^p y \iff M, s \vDash x \succ_i^b y.$$

Let $M = (S, V, R)$ be a level-0 model and let $i$ be a voter. Now suppose that voter $i$ starts reasoning on level 1. This means that she assumes that all other voters are level-0 reasoners, and hence that they will just vote truthfully. Gibbard-Satterthwaite manipulators are first-level reasoners: they assume that they are the only voter who takes a strategic ballot into consideration.

**Definition 5.6** ($i$-level-1 model)**.** Let $M^0 = (S^0, V^0, R^0)$ be a level-0 model with actual state $s^*$. Given this model, we can define a model in which voter $i$ is a first-order reasoner by defining $M^{1,i,s^*} = (S^{1,i,s^*}, V^{1,i,s^*}, R^{1,i,s^*})$ as follows:

- Let $S^0 = \{s_1, \ldots, s_h\}$. To obtain the set of states of the model in which $i$ is a first-level reasoner, we need a copy of $S^0$ for every agent $j \neq i$. The relevant states for agent $i$ are

[1] In an S5 model, all accessibility relations $R$ are equivalence relations.

the states that are accessible for $i$, so we also copy every state that $i$ considers possible in $M^0$. To distinguish the copies of $S^0$, we index the states with the voter concerned: For $j \neq i$, let $S^{0,j} := \{s^j_x \mid 1 \leq x \leq h\}$. For $i$, let $S^{0,i} := \{s \mid s^* R^0_i s\}$. Then we define:

$$S^{1,i,s^*} := \bigcup_{j \in N} S^{0,j}.$$

- In the states that voters $j \neq i$ consider possible, preferences and ballots remain the same as in the level-0 model. So, for all $j \neq i$, $1 \leq x \leq h$ and $s^j_x \in S^{0,j}$:

  $$V^{1,i,s^*}(s^j_x) = V^0(s_x)$$

  However, in the states that voter $i$ considers possible, $i$ uses her best strategy given the truthful votes of the other voters. The preferences and ballots of other voters do not change. Hence, for all $s_x \in S^{0,i}$:

  $$V^{1,i,s^*}(s_x) = (\mathsf{proj}_1(V^0(s_x)), (\bigtimes_{j<i} \mathsf{proj}_2(V^0(s_x))(j) \times \mathcal{S}_i(M^0, s^*) \times \bigtimes_{j>i} \mathsf{proj}_2(V^0(s_x))(j)))$$

- For voter $i$, all states in $S^{0,i}$ are accessible, and we copy the accessibility relations from the original model $M^0$ for every agent $j \neq i$:

  $$R^{1,i,s^*}(i) = \{(s_x, s_y) \mid s_x, s_y \in S^{0,i}\} \cup \{(s^j_x, s^j_y) \mid j \neq i, s^j_x, s^j_y \in S^{0,j} \text{ and } s_x R^0_i s_y\}$$

  For every voter $j \neq i$, all states in her personal copy $S^{0,j}$ are accessible, and we copy the accessibility relations from the original model for every agent $l \neq i$:

  $$\begin{aligned} R^{1,i,s^*}(j) = &\{(s_x, s^j_y) \mid 1 \leq x, y \leq h \text{ and } s_x R^0_j s_y\} \cup \\ &\{(s^l_x, s^l_y) \mid l \in N \text{ such that } l \neq i, 1 \leq x, y \leq h, \text{ and } s_x R^0_j s_y\} \end{aligned}$$

- The actual state of $M^{1,i,s^*}$ is the copy of $s^*$ in this model, so $s^* \in S^{0,i}$.

$M^{1,i,s^*}$ is called an *$i$-level-1 model*.

To get an idea how level-0 and level-1 models look like, we give a simple example of a model in which the truthful preference profile is common knowledge. Suppose that we have a set of voters $N$ with truthful preference profile $\boldsymbol{p}$. The level-0 model is given by the following single-world model:

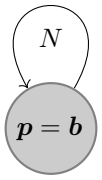



Figure 5.4: A level-0 model $M^0$ in which the preference profile is common knowledge

If voter $i$ starts reasoning at level 1, she assumes that every other voter votes truthfully. Let $\boldsymbol{b}^i$ be the ballot profile in which every voter $j \neq i$ votes truthfully, and voter $i$ votes according to her best strategy $\mathcal{S}_i(M^0, s^*)$. The $i$-level-1 model is presented in Figure 5.5.

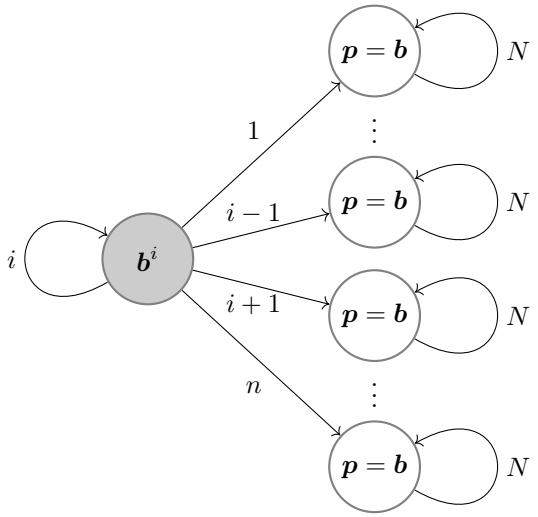



Figure 5.5: The $i$-level-1 model

If the best strategy of first-level reasoning agent $i$ is not her truthful preference, then this notion should be consistent with the concept of a single agent who (secretly) commits to a strategic ballot. The following Proposition shows that these models indeed coincide.

**Proposition 5.7.** Let $M^0$ be a level-0 model with actual state $s^*$. Suppose that $S_i(M^0, s^*) = b_i$. Let $U_{b_i^i}$ be a $b_i^i$-manipulation update model for $M^0$. Then $(M^{1,i,s^*}, s^{1,i,s^*})$ is bisimilar to $(M^0 \otimes U_{b_i^i}, (s^*, e^*))$.

*Proof.* See Appendix. □

### 5.2.2 Higher-level models

What happens if $i$ realises that other voters also might consider a strategic vote? This means that she assumes that all other voters assume that every other voter votes truthfully. So, voter $i$ engages in second-order reasoning. Suppose that according to agent $i$'s second-order reasoning, some other agent $j$ has an incentive to manipulate and casts a strategic vote (following agent $j$'s level-1 reasoning). Then, she will not consider the case in which agent $i$ votes truthfully anymore: in order to determine her own best strategy, agent $i$ will only analyse the cases in which $j$ manipulates. To determine the votes of her fellow voters, voter $i$ first has to figure out which truthful preferences the other voters have, and which preferences profiles every other voter considers possible. As an illustration, the $i$-level-2 model based on the level-0 model in which the preference profile is common knowledge is shown in Figure 5.6. Here, for $j \neq i$, $\boldsymbol{b}^j := (\mathcal{S}_j(M_0, s^*), \boldsymbol{p}(-j))$. The ballot profile $\boldsymbol{b'} := (b_1, \ldots, b_n)$ is the ballot profile in which every voter $j \neq i$ votes according to her best strategy given the level-0 model, so $b_j := \mathcal{S}_j(M^0, s^*)$. Then, voter $i$ determines her best strategy given all the strategies of voters $j \neq i$ and her true preference $\boldsymbol{p}(i)$: $b_i := \mathcal{S}_j(\bigtimes_{j \neq i} \mathcal{S}_j(M^0, s^*), \boldsymbol{p}(i))$.

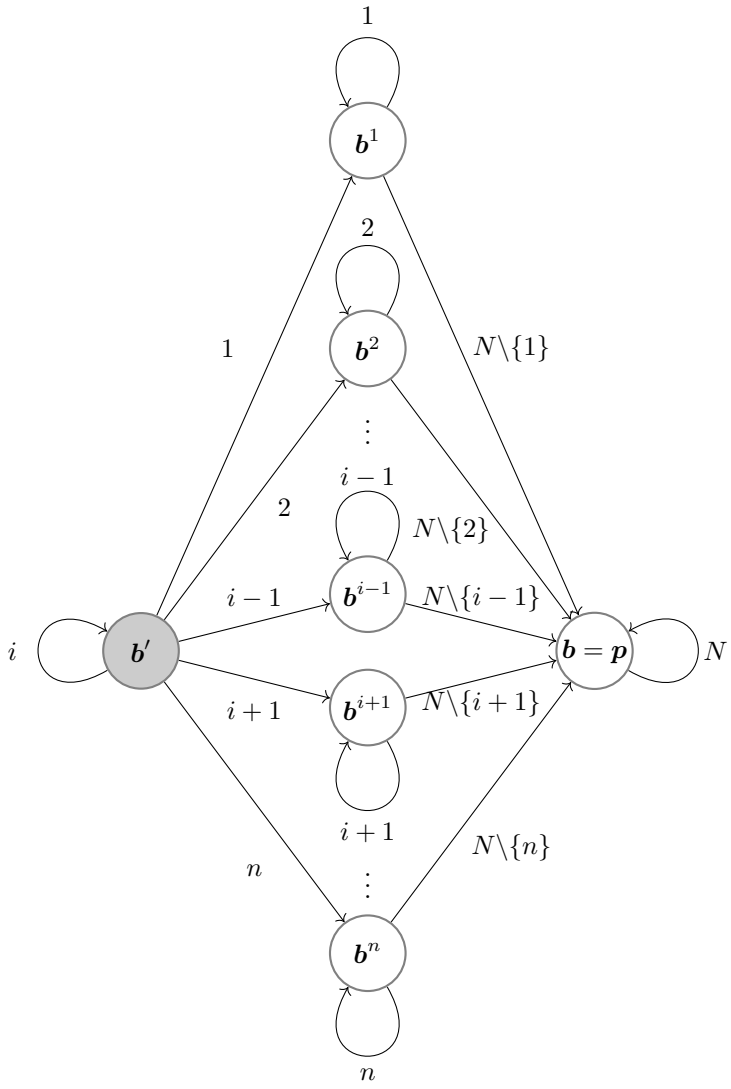



Figure 5.6: An $i$-level-2 model in which the preference profile is common knowledge

We design our level-$k$-framework along the lines of the game-theoretical model of Stahl (1993) and Stahl and Wilson (1995). To formally define models in which a voter reasons at level $k$, we use models in which voters reason at level $k-1$. Suppose that we want to transform voter $i$ to a level-$k$ reasoner. For every voter $j \neq i$, and for every state $s$ that $i$ considers possible in the initial model $M^0$, we take the model in which voter $j$ reasons at level $k-1$, and in which $s$ is the actual state. We combine these models together and 'glue' them together to obtain a model that reflects $i$'s level-$k$ reasoning. So, we an $i$-level-$k$ model is recursively defined and we make it precise in the following way:

**Definition 5.8** ($i$-level-$k$ model)**.** Let $M^0 = (S^0, R^0, V^0)$ be a level-0 model with actual world $s^*$. Suppose that for every voter $i \in N$ and for every $s \in S^0$, we have the model $M^{k-1,i,s} = (S^{k-1,i,s}, V^{k-1,i,s}, R^{k-1,i,s})$ with actual world $s^{k-1,i,s}$. We again define $S^{0,i} := \{s \mid$

$s \in S_0$ and $s^* R_0^i s\}$. Then $M^{k,i,s^*} = (S^{k,i,s^*}, V^{k,i,s^*}, R^{k,i,s^*})$ is defined as follows:

- $S^{k,i,s^*} := S^{0,i} \cup \bigcup_{t \in S^{0,i}} \bigcup_{j \neq i} S^{k-1,j,t}$
- For every $t \in S^{0,i}$, $j \neq i$, $u \in S^{k-1,j,t}$ with $s^{k-1,j,t} R_j^{k-1,j,t} u\}$:

  $V^{k,i,s^*}(u) := V^{k-1,j,t}(u)$

  In the states that represent the level-$k$ reasoning of voter $i$, the preferences are equal to the preferences the level-0 model, the ballots of voters $j \neq i$ are the ballots following their level-$(k-1)$ reasoning, and the ballot of voter $i$ is her best strategy given the $k-1$-ballots of all other voters $j \neq i$. So, for every $t \in S^{0,i}$:

  $$V^{k,i,s^*}(t) := \big(\mathsf{proj}_1(V^0(t)), \big( \bigtimes_{j<i} \underbrace{\mathsf{proj}_2(V^{k-1,j,t}(t))(j)}_{\text{ballot of } j} \times \overbrace{\mathcal{S}_i\big( \bigcup_{t' \in S^{0,i}} \bigtimes_{j \neq i} \mathsf{proj}_2(V^{k-1,j,t'}(t')), \boldsymbol{p}_{s^*}(i)\big)}^{\text{best strategy of } i \text{ given ballots of other voters}} \times \bigtimes_{j>i} \underbrace{\mathsf{proj}_2(V^{k-1,j,t}(t))(j)}_{\text{ballot of } j} \big)\big)$$

  For every $v \in S$ such that $\neg s^{k-1,j^t} R_j^{j,k-1,t} v$ for all $t \in S^{0,i}$ and $j \neq i$:

  $V^{k,i,s^*}(v) := V^{k-1,i,t}(v)$

- For voter $i$:

  $R^{k,i,s^*}(i) := \bigcup_{l \neq i} R^{k-1,l,s}(i) \cup \{(s,t) \mid s,t \in S^{0,i}\}$

  For voters $j \neq i$:

  $R^{k,i,s^*}(j) := \bigcup_{l \neq i} R^{k-1,l,s}(j) \cup \{(t,u) \mid t \in S^{0,i}$ and $u \in S^{k-1,j,t}$ such that $s^{k-1,i,t} R_j^{k-1,j,t} u\}$

- The actual state $s^{k,i,s^*} := s^*$

With every extra step of higher-order reasoning, $i$-level-$k$ models grow exponentially. Therefore, it is hard to visually demonstrate a general $i$-level-$k$ model. In the following section, we will discuss an example in which we recursively construct a higher-level model up to level 2.

We have seen how to construct a model in which a single voter is a level-$k$ reasoner. If a voter is a $k$-reasoner, she thinks that every other voter reasons at level $k-1$. However, this does not have to be the actual situation. For example, if every voter is a level-2 reasoner, every voter (falsely) believes that all other voters in the group are first-order reasoners. A model in which every voter is a level-$k$ reasoner will be called a $N$-level-$k$ model. It is formally defined as follows:

**Definition 5.9** ($N$-level-$k$ model)**.** Let $M^0 = (S^0, V^0, R^0)$ be a level-0 model with actual state $s^*$. Suppose that for every $i \in N$, and every $s \in S^0$ we have the model $M^{k,i,s}$, with actual world $s^{k,i,s}$. Then an *$N$-level-$k$* model $M^{k,N,s^*} = (S^{k,N,s^*}, V^{k,N,s^*}, R^{k,N,s^*})$ is defined as follows:

- $S^{k,N,s^*} = t \cup \bigcup_{i \in N} S^{k,i,s^*}$
- For every $i \in N$:
  $R^{k,N,s^*} = \{(t,u) \mid (s^{k,i,s^*}, u) \in R^{k,i,s^*}\} \cup \bigcup_{j \in N} R^{k,j,s^*}(j)$
- For every $i \in N$, $s \in S^{k,i,s^*}$:
  $V^{k,N,s^*}(s) = V^{k,i,s^*}(s)$,
  and $V^{k,N,s^*}(t) = (\boldsymbol{p}_{s^*}, \bigtimes_{i \in N} \boldsymbol{b}_{k,i,s^*}(i))$.

### 5.2.3 Manipulation under level-$k$ reasoning

The central question now is: are these higher-level models susceptible to manipulation? We first define the notion of a $k^{\text{th}}$-order incentive to manipulate:

**Definition 5.10** ($k^{\text{th}}$-order incentive to manipulate)**.** Let $M^0$ be a level-0 model. Let $M^{k,G,s^*}$ be a model with $G = \{i\}$ or $G = N$ and actual state $s^{k,G,s^*}$. So, $M^{k,G,s^*}$ is either an $i$-level-$k$ model or an $N$-level-$k$ model based on $M^0$. Let $s \in S^{k,G,s^*}$ such that $s^{k,G,s^*} R_i^{k,G,s^*} s$. If

$$\boldsymbol{b}_s(i) \neq \boldsymbol{p}_{s^*}(i),$$

we say that $i$ has a *$k^{th}$-order incentive to manipulate* in $M^0$.

Alternatively, we can define a $k^{\text{th}}$-order incentive to manipulate in terms of the best strategy of a voter $i$ in a model in which $i$ reasons at level $k$. We show that a voter has a $k^{\text{th}}$-order incentive to manipulate if and only if her best strategy in a model in which she reasons at level $k$, is not equal to her truthful preference.

**Proposition 5.11.** Let $M^0$ be a level-0 model. Let $G = \{i\}$ or $G = N$. Then $i \in G$ has an $k^{\text{th}}$-incentive to manipulate if and only if

$$\mathcal{S}_i(M^{k,G,s^*}, s^{k,G,s^*}) \neq \boldsymbol{p}_{s^*}(i)$$

*Proof.* See Appendix. □

**Definition 5.12** (Strategyproofness under level-$k$ reasoning)**.** A level-0 model $M^0$ with actual world $s^*$ is *strategyproof under level-k reasoning* if no voter $i \in G$ has a $k^{\text{th}}$-order incentive to manipulate. A social choice function $F$ is *strategyproof under level-k reasoning* if any level-0 model for $F$ is strategyproof under level-$k$ reasoning.

**Example 5.13.** We consider an example with two voters, four alternatives and social choice function Borda with lexicographic tie-breaking. Voter 1 and 2 are both have some uncertainty about each other's preferences. The initial level-0 model is shown in Figure 5.7. We are going to show that voter 1 does not have a first-order incentive, but does have a second-order incentive to manipulate.

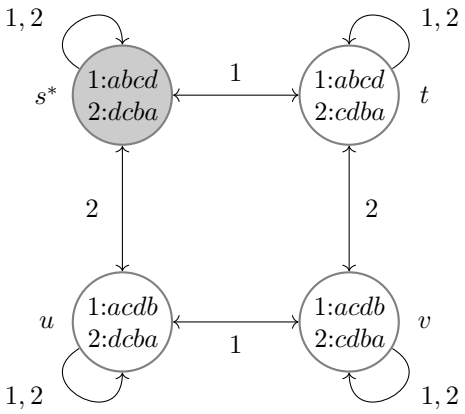



Figure 5.7: The level-0 model $M^0$

Voter 1 does not have an incentive to manipulate: in the upper left world, her favourite alternative $a$ is already winning. In the upper right world, $c$ is winning. Voter 1 could prevent $c$ from winning by voting $abdc$ instead in this case, but if the upper left world is the actual state, this manipulation will result in making $d$ win, which is worse than $b$ for voter 1. Hence, $abdc$ is not a dominant manipulation. Now suppose that voter 1 starts reasoning at level 2. To construct the 1-level-2 model, we first need the models in which voter 2 reasons at level 1. Since voter 1 cannot distinguish between state $s^*$ and state $t$, we need both the 2-level-1 model $M^{1,1,s^*}$ and the 2-level-1 model $M^{1,1,s'}$. These models are presented in Figure 5.8 and 5.9.

If $s^*$ is the actual state, voter 2 cannot distinguish between states $s^*$ and $u$. So, the 2-level-1 model looks as follows[2]. Not all transitive arrows are shown. Voter 2 has a dominant manipulation $cdba$: voter 2 thinks that voter 1 votes truthfully, so either $abcd$ or $acdb$. In case $abcd$ is the ballot of voter 1, then voter 2 can strictly improve the outcome by voting $cdba$. In that case, $c$ will become the winner instead of $a$. If $acdb$ is the ballot of voter 1, then $c$ will be the winner of the outcome. If voter 2 votes $cdba$, $c$ will still be the winner of the election. We indicate the dominant manipulation of voter 2 with red in 5.8.

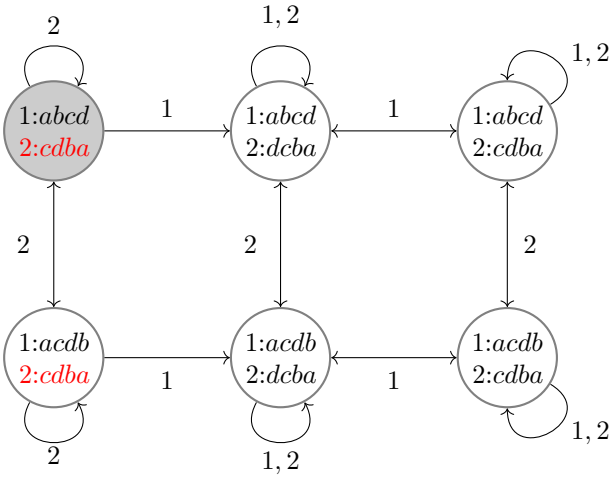



Figure 5.8: The 2-level-1 model $M^{1,2,s^*}$ given that $s^*$ is the actual world

If $t$ would be the actual state, voter 2 cannot distinguish between states $s^*$ and $u$. Here, voter 2 does not have an incentive to manipulate, as she realises that her favourite alternative $c$ is already winning in both scenarios.

[2]If we apply Definition 5.8, we would actually obtain a bigger model, namely the model that contains two copies of the level-0 model. However, that model is bisimilar to the model presented here. For the sake of space and complexity of the model, we show smaller versions of the models here.

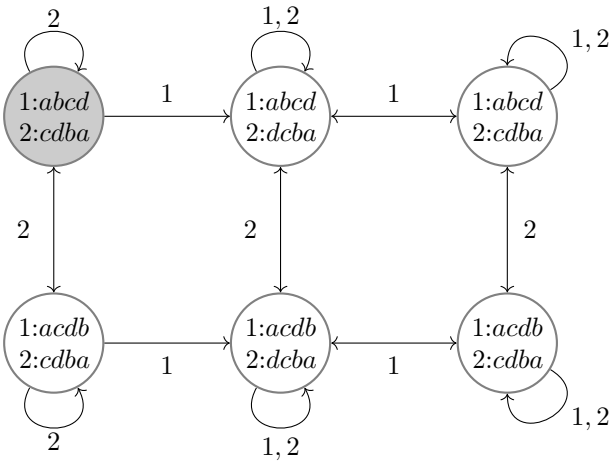



Figure 5.9: The 2-level-1 model $M^{1,2,t}$ given that $t$ is the actual world

Now, we can construct the 1-level-2 model. This will result in the model of Figure 5.10.

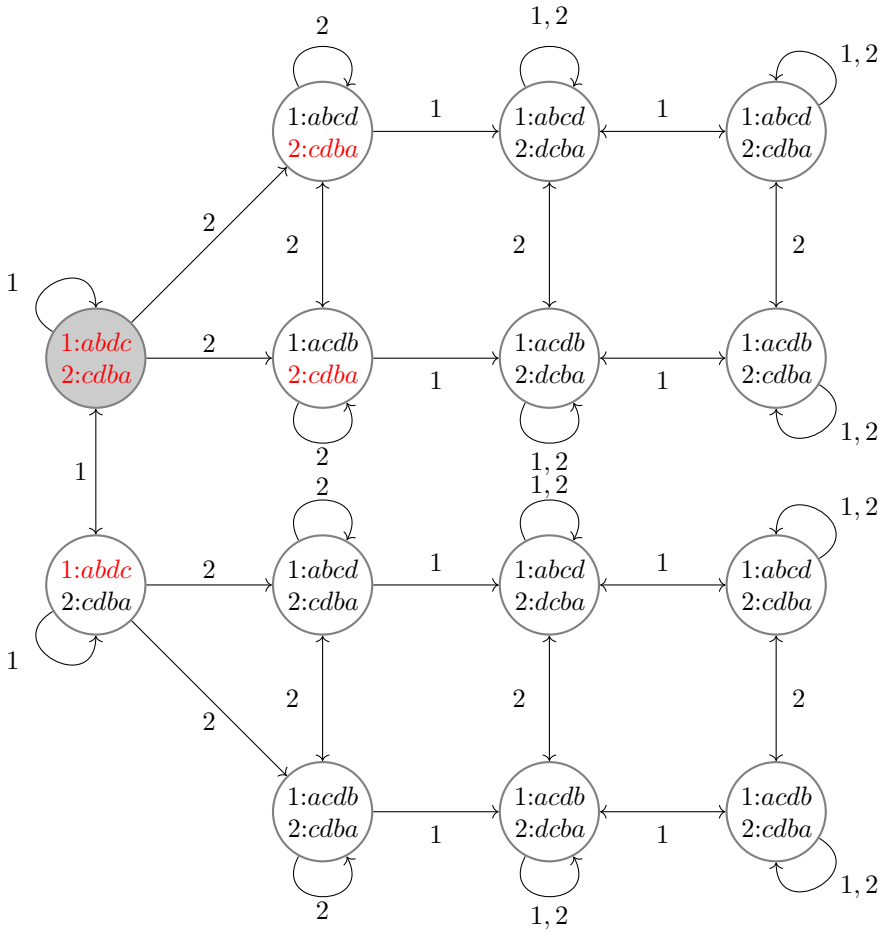



Figure 5.10: The 1-level-2 model of initial model $M^0$. The red ballots are insincere ballots.

In $M^{2,1,s^*}$, we have transformed voter 1 to a level-2 reasoner, so voter 1 thinks that voter 2 is a level-1 reasoner, which means that voter 2 thinks that voter 1 will declare a truthful ballot. After reflecting on this, voter 1 realises that if *dcba* is the truthful preference of voter 2, then voter 2 has an incentive to manipulate by voting *cdba* (indicated in red in Figure 5.10), as we saw in Figure 5.8. If *cdba* is the true preference of voter 2, then voter 2 does not have an incentive to manipulate, as we saw in Figure 5.9. In any case, voter 2 will vote *cdba* and voter 1 knows this. But now, voter 1 has an incentive to manipulate: if voter 2 votes *cdba*, voter 1's best strategy is voting *abdc*. She knows that she can safely strategise by voting *abdc* and make alternative $a$ win.

The following theorems are analogues of the results of Terzopoulou (2017), here applied to our framework in Voting Theory instead of Judgment Aggregation.

**Theorem 5.14.** *For every social choice function $F$, if $F$ is strategyproof under level-1 reasoning, then $F$ is strategyproof under level-$k$ reasoning for every $k \in \mathbb{N}$.*

*Proof.* Let $F$ be a social choice function that is strategyproof under level-1 reasoning. We show this by induction on $k$. Suppose that $F$ is strategyproof under level-$k$ reasoning. Let

$M^{k,N,s^*} = (S^{k,N,s^*}, V^{k,N,s^*}, R^{k,N,s^*})$ be an $N$-level-$k$ model (based on a level-0 model $M^0$) with actual world $s^{k,N,s^*}$. Since $F$ is strategyproof under level-$k$ reasoning, for all $j \in N$ and all $s \in S^0$, it holds that

$$\mathcal{S}_j(M^{k,j,s}, s^{k,j,s}) = \boldsymbol{p}_s(j).$$

Now take arbitrary $i \in N$ consider the model $M^{k+1,i,s^*}$. We apply Proposition 5.11 and show that

$$\mathcal{S}_i(M^{k+1,i,s^*}, s^{k+1,i,s^*}) = \boldsymbol{p}_{s^*}(i).$$

We have

$$\begin{aligned}
\mathcal{S}_i(M^{k+1,i,s^*}) &= \mathcal{S}_i(\{(\mathsf{proj}_2(V^{k+1,i,s^*}(t))(-i) \mid t \in S^{k+1,i,s^*} \text{ with } s^{k+1,i,s^*} R^{k+1,i,s^*} t\} \\
&= \mathcal{S}_i\big(\{\bigtimes_{j\neq i} \mathcal{S}_j(M^{k,j,t}, s^{k,j,t} \mid t \in S^{k+1,i,s^*} \text{ with } s^{k+1,i,s^*} R^{k+1,i,s^*} t\}, \boldsymbol{p}_{s^*}(i)\big) \\
&= \mathcal{S}_i\big(\{\bigtimes_{j\neq i} \boldsymbol{p}_t(j) \mid t \in S^{k+1,i,s^*} \text{ with } s^{k+1,i,s^*} R^{k+1,i,s^*} t\}, \boldsymbol{p}_{s^*}(i)\big) \\
&\quad \text{(since } F \text{ is strategyproof under level-}k \text{ reasoning)} \\
&= \mathcal{S}_i\big(\{\boldsymbol{p}_t(-i) \mid t \in S^{k+1,i,s^*} \text{ with } s^{k+1,i,s^*} R^{k+1,i,s^*} t\}, \boldsymbol{p}_{s^*}(i)\big) \\
&= \boldsymbol{p}_{s^*}(i) \text{ (since } F \text{ is strategyproof under level-1 reasoning)}
\end{aligned}$$

Thus, $i$ does not have a $k+1^{\text{th}}$-order incentive to manipulate. Since $i$ was arbitrary, we conclude that $F$ is strategyproof under $k+1$-level reasoning. □

**Theorem 5.15.** *Let $F$ be the plurality rule along with a lexicographic tie-breaking rule. Then*

*(i) $F$ is susceptible to winner-manipulation under first-level reasoning.*

*(ii) $F$ is immune to winner-manipulation under second-level reasoning.*

*Proof.* (i) Let $X$ be the set of alternatives and $N$ be the set of voters. Consider a model $M \in \mathcal{M}_{\text{winner,sincere}}$ (this is a level-0 model) with truthful preference profile $\boldsymbol{p}$, $F(\boldsymbol{p}) = x_m$, and voter $i$ with preference $p_i = x_1 \succ \cdots \succ x_m$. Then voter $i$ has an incentive to manipulate, because the outcome can never be worse than $x_m$ and in some cases, the outcome of the election will be strictly better.

(ii) Let $M \in \mathcal{M}_{\text{winner,sincere}}$ (which is again a level-0 model), $i$ an arbitrary voter, $\boldsymbol{p}$ the truthful profile and $F(\boldsymbol{p}) = x$. Let voter $i$'s top choice be $y$. Suppose that there is a strategic ballot $b_i \neq p_i$ such that $F(b_i, \boldsymbol{p}(-i)) = z$ and $y \succ_i^p x$ (note that $z \neq y$ and $x \neq y$). Since $i$ only has winner information, there exists a state $s \in S$ such that $M, s \vDash \boldsymbol{p'}$ such that $\{j \in N \mid \mathsf{rank}_{\boldsymbol{p'}(j)}(x) \neq 1 \text{ and } \mathsf{rank}_{\boldsymbol{p'}(j)}(y) = 2\} = \{k \in N \mid \mathsf{rank}_{\boldsymbol{p'}(k)}(x) = 1\}$. Following the first-order reasoning of the voters in $\{j \in N \mid \mathsf{rank}_{\boldsymbol{p'}(j)}(x) \neq 1 \text{ and } \mathsf{rank}_{\boldsymbol{p'}(j)}(y) = 2\}$, they all have an incentive to strategically vote for $y$ (that is, report any ballot with $y$ on the first position). So following the second-order reasoning of voter $i$, there exists a state $(s, e) \in M \otimes U$ with $(s^*, e^*)R_i(s, e)$ such that $M^{1,N,s^*}, s \vDash y$ and $M^{1,N,s^*} \otimes U_{b_i^i}, (s, e) \nvDash y$. In that state $(s, e)$, $i$ is pivotal with respect to $y$. Thus, $i$'s manipulation would obstruct the manipulations of the first-level reasoners, and hence it would prevent alternative $y$ from winning. Therefore, $i$ does not have an incentive to manipulate under second-level reasoning. □

Terzopoulou (2017) showed that if a voting rule is manipulable for level-1 reasoners, it will also be manipulable for some higher level reasoners.

**Theorem 5.16.** *Consider any social choice function $F$. If $F$ is susceptible to manipulation under first-level reasoning, then even if $F$ is immune to manipulation under level-$k$ reasoning for some $k$, it is susceptible to manipulation under level-$k+1$ reasoning.*

*Proof.* See Appendix. □

This theorem shows that taking in consideration the strategic behaviour of other voters will in general not stop the manipulating behaviour of voters.

## 5.3 Concluding remarks

In this chapter, we introduced the cognitive hierarchy theory as model for higher-order reasoning voters. We showed that in general, sophisticated agents who apply higher-order reasoning will also not refrain from voting strategically. Even if a non-dictatorial and surjective voting rule is strategyproof under level-$k$ reasoning for some $k \in \mathbb{N}$, it will be susceptible to manipulation under level-$(k+1)$ reasoning. However, this result does not imply that every voting rule can always be manipulated by real human beings. It would be useful to study the manipulability of specific voting rules under level-2 and level-3 reasoning, as those levels seem to be the most realistic levels of reasoning in practice.

One limitation of the iterated reasoning models described here is that a level-$k$ agent assumes that all other agents are exactly one level of sophistication lower than herself. Every agent assumes that she is the most sophisticated voter. Although there is experimental evidence for this theory, it could be interesting to consider other structures of higher-order reasoning. For example, it is reasonable to assume that a level-$k$ agent thinks that not every agent will reason at level $k-1$ exactly, but that some of them are more sophisticated than others. An even more realistic situation would be when which agents are uncertain about the level of reasoning of their peers: in such scenarios, an agent reasoning at level $k$ cannot distinguish between submodels in which agents reason at level $1, 2, \ldots$ or $k-1$. This would introduce a lot of uncertainty and might even complicate manipulation for that agent. Hence, it would be interesting to analyse to which degree voting rules are susceptible to manipulation under a generalised notion of level-$k$ reasoning.

# Chapter 6

# Higher-order reasoning in iterative voting

In most voting settings, it is assumed that the votes are collected in one step, without allowing the voters to adapt their ballots to the ballots of other voters. However, in many applications, a dynamic approach may represent a more realistic model of the process of collective decision making. Iterative voting is a dynamic procedure where voters are allowed to change their ballots as often as they want. An equilibrium point is reached when no voter wishes to change its preference. Real-life examples include websites as event scheduler Doodle, Facebook, and consensus decision making in Wikipedia.

In many papers that study iterative voting procedures, it is assumed that voters only consider changing their vote if they can improve the current situation, and that they do not take into account future steps by other players. These voters are myopic. If we consider voters with the capacity to reason on a higher level, these voter will take into account future steps by other players when determining a best response. Higher-order reasoners have a more farsighted view of the consequences of their manipulative ballots: they are able to look ahead and predict future steps by other voters.

To illustrate how higher-order reasoning could play a role in iterative voting, we consider an example. Suppose that six colleagues want to schedule a meeting. There are four possible moments: Monday ($a$), Tuesday ($b$), Wednesday ($c$) of Friday ($d$). They use an online internet tool to pick a day of the meeting. In this online internet tool, every participant reports a ballot order and the tool uses 2-approval (and tie-breaking order $b \triangleright c \triangleright a \triangleright d$) to decide on the day the meeting will take place. Every colleague can change her reported ballot order as often as she wants. Suppose that the preferences of the colleagues are as follows:

| | |
|---|---|
| 1 | $abdc$ |
| 2 | $dbca$ |
| 3 | $cadb$ |
| 4 | $cadb$ |
| 5 | $bcda$ |
| 6 | $abdc$ |

Under this truthful profile, alternative $a$ and $b$ are tied, so $b$ will be the winner. Colleague 1

realises that she can improve the outcome by voting *adbc* instead of *abcd*, because then *b* will lose a point and *a* will become the winner:

| | |
|---|---|
| 1 | *adbc* |
| 2 | *dbca* |
| 3 | *cadb* |
| 4 | *cadb* |
| 5 | *bcda* |
| 6 | *abdc* |

However, this is outcome not favourable for colleague 2, as *a* is his least favourite option. He can improve by voting *dcba* instead of *dbca*, and then *c* will become the winner. The new situation is as follows:

| | |
|---|---|
| 1 | *adbc* |
| 2 | *dcba* |
| 3 | *cadb* |
| 4 | *cadb* |
| 5 | *bcda* |
| 6 | *abdc* |

Note that this is harmful for colleague 1: option *c* is her least favourite alternative, and she cannot make *b* win anymore. Moreover, there is no voter who can improve the outcome: alternative *b* and *d* have to little points, so no voter can individually make them win. For alternative *a*, it holds that no voter who prefers *a* to the current winner *c* can improve the outcome: colleague 1 and 6 do not have a strategic ballot that would make *a* win from *c*. After reflecting on this, colleague 1 realises that in retrospect, changing her vote was not strategic, because now she ends up with option *c*, which is worse for her than option *b*. If she had realised this before, by predicting the iterative voting process in advance, she would have avoided this situation by not changing her vote. In that case, *b* would have become the winner.

In the best scenario, iterative voting procedures may even become strategyproof: if a voter knows that a manipulation of the truthful ballot might cause a sequence of manipulations by other voters such that the outcome of the election will be worse for her than the outcome under the truthful profile, she might decide to not manipulate at all in the first round.

In this chapter, we will focus on second-order reasoning in iterative voting. A second-order reasoning voter is able to look one step in the future: she does not just analyse how her vote affects the outcome of the election, but she is also able to analyse how the new ballot profile can be manipulated in the next round. In the next section, we discuss how two subsequent manipulations interact.

## 6.1 Interaction of manipulations

We consider situations in which there is a manipulative voter who casts a strategic vote, such that an incentive to manipulate for a second manipulator is created. We can think of the interaction between the first and the second manipulator as a symbiosis: the second manipulation would not have been possible if the first voter had not manipulated. The second manipulation

can be beneficial for the first voter, in the sense that the winner of the election of the double-manipulated ballot is even better than the winner under the single-manipulated profile, from her perspective. However, the second manipulation can also be harmful for her: the winner of the double-manipulated ballot profile can be worse than the winner under the truthful profile. The interaction between the first and the second manipulator can be characterised in terms of types of symbiosis: the second manipulation can be mutualising, neutralising or parasitising. These types of symbiosis were first introduced by De Bary (1879).

### 6.1.1 Mutualising manipulation

We say that a consecutive manipulation is *mutualising* if both manipulators benefit from the second manipulation. Let $F$ be a social choice function, let $N$ be a set of voters with current ballot profile $\boldsymbol{b}$. Suppose that voter $i$ is the initial manipulator and has an incentive to vote $b_i' \neq \boldsymbol{b}(i)$, so $F(b_i', \boldsymbol{b}(-i)) \succ_{p_i} F(\boldsymbol{b})$. Now suppose that voter $j$ has an incentive to manipulate the profile $((b_i', \boldsymbol{b}(-i)))$ and that $b_j'$ is the dominant manipulation. If $F(b_j', b_i', \boldsymbol{b}(-i,j)) \succ_{p_i} F(\boldsymbol{b})$ and no voter had an incentive to make $F(b_j', b_i', \boldsymbol{b}(-i,j))$ win under the initial profile, then the manipulation of voter $j$ is mutualising. It does not have to hold that the outcome under the double-manipulated profile is better than the outcome when only voter $i$ manipulates: it may hold that $F(b_i', \boldsymbol{b}(-i)) \succ_{p_i} F(b_j', b_i', \boldsymbol{b}(-i,j))$. In this case, the manipulation of $j$ is still mutualising. Although the outcome after two rounds is not as good as the outcome after one round for voter $i$, when reasoning from the initial state, this still stimulates voter $i$ to manipulate.

**Example 6.1.** We have four voters, four alternatives and the voting rule is Borda, paired with the lexicographic tie-breaking rule. Suppose that voter 1 has preference $a \succ b \succ c \succ d$, voter 2 has preference $c \succ a \succ d \succ b$, voter 3 has preference $b \succ d \succ c \succ a$ and voter 4 has preference $c \succ b \succ a \succ d$. If every voter reports a truthful ballot, $c$ will win. This is disadvantageous for voter 1 and she can manipulate by voting *abdc*. If she does so, in the second round $b$ will win. Voter 2 was very happy with the result in the first round, but alternative $c$ is his worst option. He can manipulate by voting *acdb* and make $a$ win. This is very beneficial for voter 1, as this is her top choice. So, the manipulation of voter 2 is mutualising.

| | | | | |
|---|---|---|---|---|
| 1: *abcd* | | 1: *abdc* | | 1: *abdc* |
| 2: *cadb* | | 2: *cadb* | | 2: *acdb* |
| 3: *bdca* | $\xrightarrow{1}$ | 3: *bdca* | $\xrightarrow{2}$ | 3: *bdca* |
| 4: *cbad* | | 4: *cbad* | | 4: *cbad* |
| $a$:6, $b$:7, $c$:8, $d$:3 | | $a$:6, $b$:7, $c$:7, $d$:4 | | $a$:7, $b$:7, $c$:6, $d$:4 |
| $c$ wins | | $b$ wins | | $a$ wins |

### 6.1.2 Neutralising manipulation

When a voter casts a strategic vote and manages to improve the outcome from her perspective, this might be disadvantageous for another voter. The other voter was quite happy with the winner under the truthful ballot profile, but does not like the outcome under the new, untruthful ballot profile. In this case, she might try to *neutralise* the strategic vote: by casting a strategic vote, the elected alternative will be the same alternative as the winner under the truthful ballot profile. This is unfavourable for the first manipulator, but not harmful.

**Example 6.2.** We have five voters, three alternatives and the voting rule is plurality, paired with the lexicographic tie-breaking rule. The preferences of the voters are given in the table

below. If every voter reports a truthful ballot, then $b$ will win. Voter 1 has an incentive to vote $acb$ and make $a$ the winner. However, voter 2 does not like this, since $a$ is her least favourite alternative. She can neutralise this manipulation by voting $bca$, which will make $b$ again the winner.

| | | | | |
|---|---|---|---|---|
| 1: $cab$ | | 1: $acb$ | | 1: $acb$ |
| 2: $cba$ | | 2: $cba$ | | 2: $bca$ |
| 3: $acb$ | $\xrightarrow{1}$ | 3: $acb$ | $\xrightarrow{2}$ | 3: $acb$ |
| 4: $bac$ | | 4: $bac$ | | 4: $bac$ |
| 5: $bac$ | | 4: $bac$ | | 4: $bac$ |
| $a$:1, $b$:2, $c$:2 | | $a$:2, $b$:2, $c$:1 | | $a$:2, $b$:3, $c$:0 |
| $b$ wins | | $a$ wins | | $b$ wins |

### 6.1.3 Parasitising manipulation

Suppose that voter $i$ manipulates and this manipulation creates an incentive for voter $j$ to manipulate as well. If, from the perspective of the initial manipulator, the winner of the double-manipulated profile is less preferred than the alternative that was elected under the truthful profile, this is harmful for voter $i$. So, the second manipulator *parasitises* the manipulation of the first manipulator. Formally, let $\boldsymbol{b}$ be the initial ballot profile and suppose that voter $i$ has an incentive to manipulate by voting $b'_i$. Furthermore, suppose that under the new profile, voter $j$ has an incentive to manipulate by voting $b'_j$ and let $F(b'_i, b'_j, \boldsymbol{b}(-i,j) = x$. If $F(b'_i, \boldsymbol{b}(-i)) \succ_{p_i} F(b'_i, b'_j, \boldsymbol{b}(-i,j)$, while no voter had an incentive to make $x$ win under the initial profile, then the second manipulation is a *parasitising manipulation.*

**Example 6.3.** In this example, the voting rule is Borda, paired with the lexicographic tie-breaking rule.

| | | | | |
|---|---|---|---|---|
| 1: $abcd$ | | 1: $bacd$ | | 1: $bacd$ |
| 2: $dcba$ | | 2: $dcba$ | | 2: $dacb$ |
| 3: $cdab$ | $\xrightarrow{1}$ | 3: $cdab$ | $\xrightarrow{2}$ | 3: $cdab$ |
| 4: $bcda$ | | 4: $bcda$ | | 4: $bcda$ |
| $a$:4, $b$:6, $c$:8, $d$:6 | | $a$:3, $b$:7, $c$:7, $d$:7 | | $a$:5, $b$:6, $c$:6, $d$:7 |
| $c$ wins | | $b$ wins | | $d$ wins |

We assume that it is not the goal of the second manipulator to harm the first manipulation, or to help the first manipulator to improve the outcome. The goal of a manipulator is always to improve the outcome of the election from her own perspective. So, a manipulator can unintentionally cast a parasitising strategic vote, and in some cases the manipulator might not even know whether her manipulation is beneficial or not for the other voters.

## 6.2 Second-order manipulation in iterative voting

As discussed before, the main goal of this thesis is to investigate under which circumstances a voting situation is strategyproof, i.e., under which circumstances no voter has an incentive to manipulate. If a voter is a second-order reasoner in an iterative voting procedure, her reasoning

consists of the following two questions: how does my manipulative ballot affect the outcome of the election, and could my manipulation stimulate a subsequent manipulation by another voter that affects the outcome of the election? We assume that in case of a mutualising or neutralising manipulation, this will not stop the initial manipulator from voting strategically: in the case of a mutualising manipulation, the second manipulation is even beneficial for the first manipulator. In the case of a neutralising manipulation, the first manipulator might be disappointed that her manipulation turned out to be unsuccessful, but the second manipulation was not harmful and therefore there is no reason for the first manipulator to refrain from strategising.

However, in the case of a parasitising manipulation, this might discourage a manipulative voter: if a voter knows that if she manipulates the election, she creates an opportunity for another voter to parasitise that manipulation, she will probably decide not to strategise at all (recall that we assumed voters to be risk-averse). This is because she knows that if she would cast a strategic ballot, another voter would manipulate as well and she would end up with a worse outcome than the initial outcome. Iterative voting procedures that are not convergent if voters are first-level reasoners, might become convergent when voters think a bit more about the consequences of their manipulative behaviour. For voter $i$, a ballot $b_i$ is a second-order risk-free dominant manipulation if $b_i$ is a dominant manipulation and if she knows that there is no voter who could parasitise her strategic vote. The following definitions formalise this idea.

**Definition 6.4** (Second-order risk-free incentive to manipulate)**.** Let $F$ be an iterative social choice function, and let $M$ be an epistemic model for strategic voting for $F$ with actual world $s^*$. Let $i \in N$ with preference $p_i$ and $b_i \neq \boldsymbol{p}_{s^*}(i)$ and $b_i \neq \boldsymbol{b}_{s^*}(i)$. Then $b_i$ is a *second-order risk-free dominant manipulation* if:

- Voter $i$ considers it possible that $b_i$ is a successful manipulation:
  There exists a state $s \in S$ with $s^* R_i s$ such that $F(b_i, \boldsymbol{b}_s(-i)) \succ_{p_i} F(\boldsymbol{b}_s)$
- This manipulation $b_i$ is a dominant manipulation:
  For every $t$ with $s^* R_i t$, $F(b_i, \boldsymbol{b}_t(-i)) \succeq_{p_i} F(\boldsymbol{b}_t)$
- This manipulation $b_i$ is not parasitisable
  For every $t$ with $s^* R_i t$, if there exists a voter $j$ with a dominant manipulation $b_j \neq \boldsymbol{p}_{s^*}(j)$ of the ballot profile $(b_i, \boldsymbol{b}(-i))$, then either:
  - The manipulation $b_j$ is not harmful for voter $i$:
    $F(b_i, b_j, \boldsymbol{b}_t(-i,j)) \succeq_{p_i} F(\boldsymbol{b}_t)$, or
  - Under the initial profile, there was already a voter that had an incentive to make $F(b_i, b_j, \boldsymbol{b}(-i,j))$ win:
    There exists a voter $k \in N$, $u \in S$ with $s^* R_i u$ such that $k$ has an incentive to strategically vote $b_k \neq \boldsymbol{p}_u(k)$ and $F(b_k, \boldsymbol{b}(-k)) = F(b_i, b_j, \boldsymbol{b}(-i,j))$.

If there exists a second-order risk-free dominant manipulation for voter $i$, we say that $i$ has a *second-order risk-free incentive to manipulate*. Instead of second-order risk-free dominant manipulation, we will often say second-order risk-free manipulation for short.

**Definition 6.5.** If no voter has a second-order risk-free incentive to manipulate in a model $M$, then $M$ is *strategyproof under second-order reasoning*. $F$ is strategyproof under second-order reasoning if all models $M$ for $F$ are *strategyproof* under second-order reasoning.

**Example 6.6.** (Second-order risk-free incentive) Suppose we have five alternatives and three voters, the voting rule is Borda paired with tie-breaking order $d \triangleright c \triangleright a \triangleright e \triangleright b$ and the epistemic model $M = (S, V, R)$ is given in Figure 6.1.

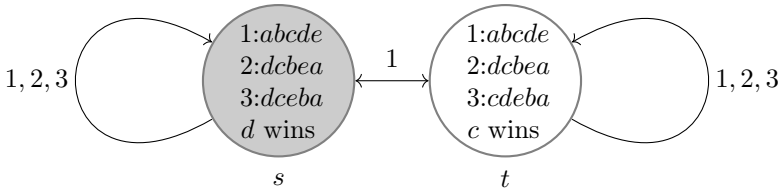



Figure 6.1: Epistemic model $M = (S, V, R)$

Voter 1 has an incentive to vote *cabde*. If voter 1 changes her ballot to *cabde*, the updated model will be $M'$ as illustrated in Figure 6.2.

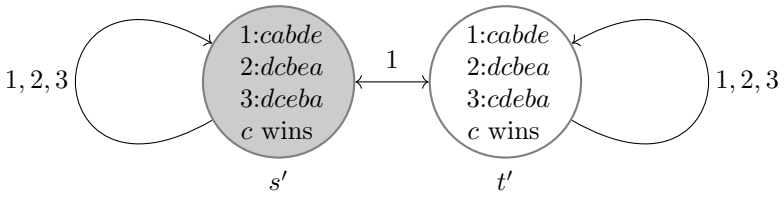



Figure 6.2: Updated epistemic model $M' = (S', V', R')$

This manipulation is also a second order dominant manipulation: this manipulation did not create an opportunity to make an alternative win that is less preferred by 1 than the winner in the corresponding state in $M$. More precisely, in state $s$ the winner is $d$. The only alternative that is worse for 1 than $d$ is alternative $e$. In state $s'$, voter 2 and 3 cannot and do not want to make alternative $e$ win. In state $t$, the winner is $c$, so only alternatives $d$ and $e$ are worse for 1. In state $t'$, if voter 2 changes her ballot to *debac*, then $d$ would win. However, this was already possible in state $s$, so the manipulation *cabde* of voter 1 did not *create* an opportunity for voter 2 to strategise and make $d$ win. Like in state $s$ and $s'$, there is no way voter 2 or 3 could make alternative $e$ win. This means that *cabde* is a second-order risk-free manipulation of voter 1's epistemic state.

**Example 6.7.** (Dominant manipulation that is not second-order risk-free) Suppose we have four alternatives, four voters and the voting rule is Borda paired with lexicographic tie-breaking. Consider the following epistemic model $M = (S, V, R)$:

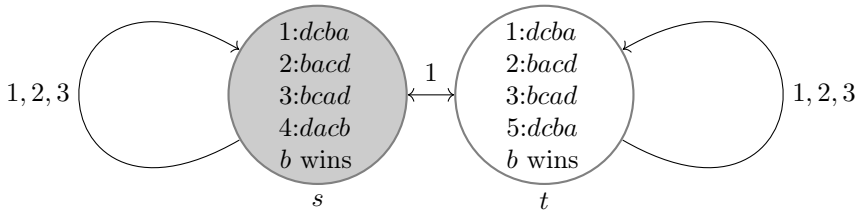



Figure 6.3: Epistemic model $M = (S, V, R)$

Voter 1 is uncertain about the ballot of voter 4, and has a dominant manipulation *cdab*. If voter 1 changes her ballot to *cdab*, the updated model will be $M'$, presented in Figure 6.4.

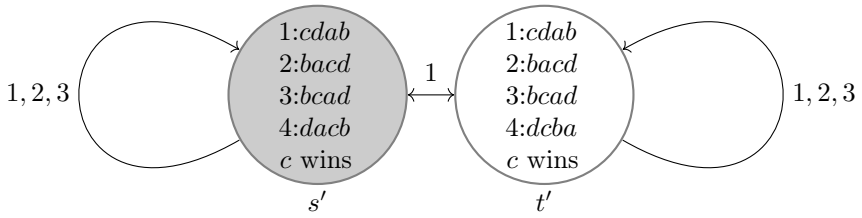



Figure 6.4: Updated epistemic model $M' = (S', V', R')$

In both states that voter 1 considers possible, $c$ will be the new winner, so this manipulation looks successful. However, this manipulation is not risk-free, because in state $s'$, voter 4 can parasitise this manipulation by voting *adcb*, while this was not possible in the initial model $M$.

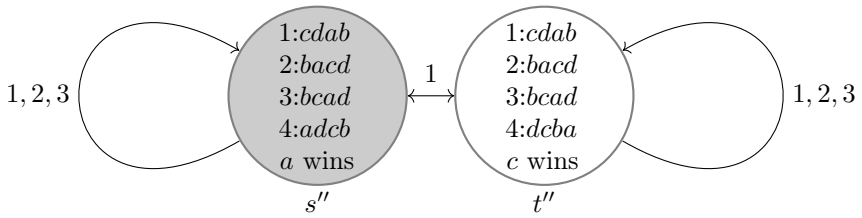



Figure 6.5: Updated epistemic model $M'' = (S'', V'', R'')$

In $M''$, the winner in state $s''$ is alternative $a$, the alternative that is ranked last in voter 1's preference. Hence, the manipulation of voter 4 is parasitising and and thus the ballot *cdab* is a

not second-order risk-free manipulation for voter 1.

Note that every strategyproof social choice function is also second-order strategyproof. A second-order strategyproof social choice function is not necessarily strategyproof: a second-order strategyproof social choice function allows situations where voters have a dominant manipulation, but this manipulation must always be *parasitisable*: there must be another voter that could parasitise this manipulation.

In the situation of full information, we can express the notion that $b_i$ is a second-order risk-free manipulation for voter $i$ in the language as follows:

$$\bigwedge_{x\in X}\bigwedge_{y\in X}\bigwedge_{z\in X}\Big(\langle i\rangle(x\wedge[b_i]y\to y\succ_i^p x)\wedge[i](x\wedge[b_i]y\to y\succeq_i^p x)\Big) \tag{6.1}$$

$$\wedge\Big(\Big(\bigwedge_{j\in N}\bigwedge_{b_j\in\mathcal{B}}\Big(\langle i\rangle\langle j\rangle([b_i][b_j]z\to z\succ_j^p y)\wedge[i][j]([b_i][b_j]z\to z\succeq_j^p y)\Big) \tag{6.2}$$

$$\to\Big(z\succeq_i^p y\vee\bigvee_{k\in N}\bigvee_{b_k\in\mathcal{B}}\langle i\rangle\langle k\rangle([b_k]z\wedge z\succ_k^p x)\wedge[k]([b_k]z\wedge z\succeq_k^p x)\Big)\Big) \tag{6.3}$$

Here, line (6.1) reads as "If voter $i$ has an incentive to vote $b_i$", line (6.2) reads as "And there is a voter $j$ that has an incentive to vote $b_j$ under the profile in which $i$ votes $b_i$", and finally line (6.3) reads as "Then either, this manipulation is risk-free for voter $i$ or there is a voter $k$ that had an incentive to make $z$ win under the initial profile".

The main issue of this thesis is to investigate the manipulability of social choice functions, so the rest of this chapter will focus on that. If voters are second-order reasoners, and refrain from risky manipulations, are there 'reasonable' iterative voting procedures that are immune to manipulation? We show that in the case of three alternatives, the answer is no. The following Theorem can be interpreted as a variant of the Gibbard-Satterthwaite Theorem for second-order strategyproof social choice functions, in the case of three alternatives.

**Theorem 6.8.** *Any surjective and non-dictatorial social choice function for three alternatives is susceptible to second-order risk-free manipulation.*

*Proof.* Let $n$ be the number of voters, and suppose that we have alternatives $a$, $b$ and $c$. Let $F$ be a surjective and non-dictatorial social choice function. We will show that there exists a second-order risk-free manipulation. By the Gibbard-Satterthwaite Theorem, there exists a profile and a voter $i$ such that there is a voter $i\in N$ with an incentive to manipulate. Without loss of generality, say that the preference of $i$ is $a\succ_i^p b\succ_i^p c$. We can assume that the winner under the truthful preference profile must be $b$: if it is $a$, $i$ would not have an incentive to manipulate, and if it is $c$, then the manipulation would certainly be risk-free. Voter $i$ has a manipulative vote $b_i$ of the truthful profile $\boldsymbol{p}$ such that $F(b_i,\boldsymbol{p}(-i))=a$. If this manipulation is second-order risk-free, we are done, so suppose that it is not risk-free. If $b_i=bca$ or $b_i=cba$, then if $(b_i,\boldsymbol{p}(-i))$ would be the truthful preference profile, $i$ would have a risk-free strategic vote $abc$. If $b_i=bac$, then either $abc$ or $bac$ must be a risk-free manipulation for voter $i$: if $bac$ is not risk-free, there exists a parasitising manipulation for some voter $j$ such that $j$ was not able to make $c$ win under the truthful profile, but has an incentive to manipulate and make $c$ win under the manipulated profile $(bac,\boldsymbol{p}(-i))$. But since $bac$ was not risk-free, no voter has an incentive to manipulate and make $c$ win under $\boldsymbol{p}$, so then $abc$ must be a risk-free strategic vote

for $i$ if $(bac, \boldsymbol{p}(-i))$ would have been the truthful profile. It is left to consider $b_i = cab$ and $b_i = acb$.

Suppose that $acb$ is not a manipulative ballot, so only $cab$ is. Then we have the following situation: $F(\boldsymbol{p}) = b$, $F(acb, \boldsymbol{p}(-i)) \neq a$ since $acb$ not a manipulative ballot, and $F(cab, \boldsymbol{p}(-i)) = a$. If $F(acb, \boldsymbol{p}(-i)) = b$, then voter $i$ has a risk-free manipulation $cba$ if $(acb, \boldsymbol{p}(-i))$ would be the truthful profile. If $F(acb, \boldsymbol{p}(-i)) = c$ and $F(cab, \boldsymbol{p}(-i)) = a$, we have that both $\boldsymbol{p}$ and $(acb, \boldsymbol{p}(-i))$ are manipulable by $i$. If one of these manipulations is risk-free, we are done, so suppose they are both not risk-free. This means that there exists a parasitising manipulator that can manipulate $(cab, \boldsymbol{p}(-i))$ and make $c$ win (parasitising the manipulation of the voter with preference $abc$), and that there exist a parasitising manipulator that can make $b$ win (parasitising the manipulation of the voter with preference $acb$). But that means that no matter what the preference of the parasitising manipulators is, the manipulation is risk-free: $a$ is the winner under $(cab, \boldsymbol{p}(-i))$, and there is a voter with an incentive to manipulate and make $a$ win, and there is a voter with an incentive to manipulate and make $b$ win. Thus, their manipulations will never create a new opportunity for a voter to parasitise the manipulation, and hence they are safe. We conclude that if $acb$ is *not* a manipulative ballot for $i$, then there is a risk-free manipulation. So, in the rest of the proof, we only have to consider situations where $acb$ is a manipulative ballot (note that we do not exclude $cab$ to be a manipulative ballot as well).

Since the manipulation of $i$ is not risk-free, there is a parasitising manipulator $j$. Voter $j$ must have preference $cab$, or else $j$ would not have an incentive to manipulate, or the manipulation would be risk-free. In the same way as for voter $i$, we can argue that ballots $b_j = abc$, $b_j = bac$ or $b_j = acb$ entail a risk-free manipulation for $j$. So, we only have to consider $b_j = cba$ and $b_j = bca$. Now suppose that $cba$ is not a manipulative ballot, so $F(b_i, \boldsymbol{p}(-i)) = a$, $F(b_i, cba, \boldsymbol{p}(-i, j)) \neq c$ and $F(b_i, bca, \boldsymbol{p}(-i, j)) = c$. If $F(b_i, cba, \boldsymbol{p}(-i, j)) = a$, voter $j$ would have a risk-free incentive to manipulate under the profile $(b_i, cba, \boldsymbol{p}(-i, j))$. If $F(b_i, cba, \boldsymbol{p}(-i, j)) = b$, then in the same way as for $i$, there is a manipulation of both $(b_i, \boldsymbol{p}(-i))$ and $(b_i, cba, \boldsymbol{p}(-i, j))$. If the manipulations are both not risk-free, then there must exist parasitising manipulations of $(b_i, bca, \boldsymbol{p}(-i, j))$, one that makes $a$, and one that makes $b$ win. Again, there is a risk-free manipulation. So, in the rest of the proof, we only have to consider situations where $cba$ is a manipulative ballot (note that we do not exclude $cab$ to be a manipulative ballot as well).

If the manipulation $cba$ is risk-free, we are done, so suppose that it is not risk-free: there is a parasitising manipulator $k \in N$ who has an incentive to manipulate under the profile $(b_i, b_j, \boldsymbol{p}(-i, j))$ and make $b$ win. The parasitising manipulator $k$ must have preference $bca$, or else $k$ would not have an incentive to manipulate, or the manipulation would be risk-free. Again, if $b_k = acb$, $b_k = cab$ or $b_k = cba$, then there is a risk-free manipulation. So, consider $b_k = bac$ and $b_k = abc$. If $bac$ is not a manipulative ballot, then $F(b_i, b_j, bac, \boldsymbol{p}(-i, j, k)) \neq b$ and $F(b_i, b_j, abc, \boldsymbol{p}(-i, j, k)) = b$ and by the same reasoning as before, there exists a risk-free manipulation. Thus, we only have to consider the situation in which $bac$ is a manipulative ballot. If this manipulation is not risk-free, there must be a parasitising voter with preference $abc$ that manipulates in order to make $a$ win. By the same reasoning as for voter $i$, the manipulative ballot of this voter must be $acb$. So, the manipulation process starts again. Let $n_1$ denote the number voters with preference $abc$, $n_2$ number the voters with preference $acb$, $n_3$ the number voters with preference $cab$, $n_4$ the number voters with preference $cba$, $n_5$ the number voters with preference $bca$ and $n_6$ the number voters with preference $bac$. Now consider the following sequence of manipulations, where the labels of the arrows indicate which type of voter is manipulating in each round, and the red text indicates the strategic ballot of that voter (and hence which group of voters increases by 1):

$$
\begin{array}{c}
n_1\colon abc \\ n_2\colon acb \\ n_3\colon cab \\ n_4\colon cba \\ n_5\colon bca \\ n_6\colon bac \\ \hline \text{b}
\end{array}
\xrightarrow{1}
\begin{array}{c}
n_1-1\colon abc \\ n_2+1\colon acb \\ n_3\colon cab \\ n_4\colon cba \\ n_5\colon bca \\ n_6\colon bac \\ \hline \text{a}
\end{array}
\xrightarrow{3}
\begin{array}{c}
n_1-1\colon abc \\ n_2+1\colon acb \\ n_3-1\colon cab \\ n_4+1\colon cba \\ n_5\colon bca \\ n_6\colon bac \\ \hline \text{c}
\end{array}
\xrightarrow{5}
\begin{array}{c}
n_1-1\colon abc \\ n_2+1\colon acb \\ n_3-1\colon cab \\ n_4+1\colon cba \\ n_5-1\colon bca \\ n_6+1\colon bac \\ \hline \text{b}
\end{array}
\xrightarrow{1}
$$

$$
\begin{array}{c}
n_1-2\colon abc \\ n_2+2\colon acb \\ n_3-1\colon cab \\ n_4+1\colon cba \\ n_5-1\colon bca \\ n_6+1\colon bac \\ \hline \text{a}
\end{array}
\xrightarrow{3}
\begin{array}{c}
n_1-2\colon abc \\ n_2+2\colon acb \\ n_3-2\colon cab \\ n_4+2\colon cba \\ n_5-1\colon bca \\ n_6+1\colon bac \\ \hline \text{c}
\end{array}
\xrightarrow{5}
\begin{array}{c}
n_1-2\colon abc \\ n_2+2\colon acb \\ n_3-2\colon cab \\ n_4+2\colon cba \\ n_5-2\colon bca \\ n_6+2\colon bac \\ \hline \text{b}
\end{array}
\xrightarrow{1}
\begin{array}{c}
\dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \hline
\end{array}
\xrightarrow{\dots}
$$

So, the voters of groups $n_1$, $n_3$ and $n_5$, will one by one move to groups $n_2$, $n_4$ and $n_6$ respectively. This means that eventually, one of the groups will be empty. Suppose for example that $n_3 < n_1$ and $n_3 \leq n_5$. After a finite number of rounds, the iterative voting procedure will continue as follows:

$$
\begin{array}{c}
n_1-n_3\colon abc \\ n_2+n_3\colon acb \\ n_3-(n_3-1)\colon cab \\ n_4+(n_3-1)\colon cba \\ n_5-(n_3-1)\colon bca \\ n_6+(n_3-1)\colon bac \\ \hline \text{a}
\end{array}
\xrightarrow{3}
\begin{array}{c}
n_1-n_3\colon abc \\ n_2+n_3\colon acb \\ n_3-n_3\colon cab \\ n_4+n_3\colon cba \\ n_5-(n_3-1)\colon bca \\ n_6+(n_3-1)\colon bac \\ \hline \text{c}
\end{array}
\xrightarrow{5}
\begin{array}{c}
\boldsymbol{p}^* \\
n_1-n_3\colon abc \\ n_2+n_3\colon acb \\ n_3-n_3\colon cab \\ n_4+n_3\colon cba \\ n_5-n_3\colon bca \\ n_6+n_3\colon bac \\ \hline \text{b}
\end{array}
\xrightarrow{1}
\begin{array}{c}
\boldsymbol{p}^{**} \\
n_1-(n_3+1)\colon abc \\ n_2+n_3+1\colon acb \\ n_3-n_3\colon cab \\ n_4+n_3\colon cba \\ n_5-n_3\colon bca \\ n_6+n_3\colon bac \\ \hline \text{a}
\end{array}
$$

Let $\boldsymbol{p}^*$ be second last ballot profile in this sequence, and $\boldsymbol{p}^{**}$ the last profile. The last manipulation in this sequence is by a voter with preference *abc*, who commits to the strategic ballot *acb*. Now, suppose that profile $\boldsymbol{p}^*$ is the truthful preference profile. Then there is no voter with preference *cab*. Under $\boldsymbol{p}^{**}$, voters with preference *abc* and *acb* do not have an incentive to manipulate. If a voter with preference *bac* has an incentive to manipulate, this is not risky for voters with preference *abc*, because this will be a neutralising manipulation, making *b* win. So, in that case the last manipulation in this sequence is risk-free. If a voter with preference *cba* or *bca* has an incentive to manipulate, this manipulation is risk-free for them, because their least preferred alternative is winning under $\boldsymbol{p}^{**}$. Thus, in either case, there is a second-order risk-free manipulation. If $n_1 \leq n_3, n_5$ or if $n_5 < n_1, n_3$, the sequence of parasitising manipulations will 'terminate' in a similar way, but with a different second-order risk-free manipulation and winner. We conclude that in every case, there is a second-order risk-free manipulation, and hence that every non-dictatorial and surjective social choice function for three alternatives is susceptible to second-order manipulation. □

In the rest of the section, we will work with social choice functions that are not antagonistic:

**Definition 6.9.** A social choice function $F$ is *antagonistic* if there exists a ballot profile $\boldsymbol{b}$ in which every voter ranks a particular alternative $x \in X$ last, but where $F(\boldsymbol{b}) = x$

Non-antagonistic voting rules allow us to define 'subrules' derived from the main rule $F$, which will be used in the remainder of this section. Given any alternative $a$, the subrule $F_{-a}$ is designed to pick alternatives from $X \backslash \{a\}$, and operates as follows. Let $\boldsymbol{p}$ be an arbitrary profile of preferences over the set $X \backslash \{a\}$. Let $\overline{\boldsymbol{p}}$ be the profile of preferences over the original set of alternatives $X$ formed by appending an $a$ to the bottom of every preference order in $\boldsymbol{p}$. Then we define $F_{-a}(\boldsymbol{p}) := F(\overline{\boldsymbol{p}})$. Provided that $F$ is not antagonistic, $F_{-a}$ will not select $a$ at $\overline{\boldsymbol{p}}$ and will therefore be well-defined.

### 6.2.1 Positional scoring rules

**Lemma 6.10.** Let $F$ be a positional scoring rule paired with a rationalisable tie-breaking rule, and let $\boldsymbol{b}$ be a ballot profile such that $F(\boldsymbol{b}) = x$. Suppose that voter $i$ has a strategic ballot $b_i'$ such that $F(b_i', \boldsymbol{b}(-i)) = y$. Then at least one of the following must hold:

(i) The position of $y$ in $b_i'$ is strictly higher than in $b_i$.

(ii) The position of $x$ in $b_i'$ is strictly lower than in $b_i$.

*Proof.* To make $y$ win from $x$, the score of $y$ should be at least as high as the score of $x$. Since $F(\boldsymbol{b}) = x$, the score of $y$ must be (weakly) lower than the score of $x$. To make $y$ win, the score of $y$ must be increased, or the score of $x$ must be decreased. For any positional scoring rule, this means that the rank of alternative $y$ must be increased, or the rank of alternative $x$ must be decreased. □

**Theorem 6.11.** *Anti-plurality is susceptible to second-order risk-free manipulation.*

*Proof.* Let $F$ be the anti-plurality rule and let $x_{m-1} \triangleright x_1 \triangleright x_2 \triangleright \cdots \triangleright x_m$ be the tie-breaking rule. Now consider the preference profile $\boldsymbol{p}$ where $p_i = x_1 \succ x_2 \succ \cdots \succ x_m$ for every $i \in N$. By applying the tie-breaking rule, we obtain $F(\boldsymbol{p}) = x_{m-1}$. However, this outcome is far from ideal for every voter. In this situation, any voter $i \in N$ has an incentive to strategically vote $b_i = x_1 \succ x_2 \succ \cdots \succ x_m \succ x_{m-1}$, since $F(b_i, \boldsymbol{p}(-i)) = x_1$. Obviously, if a single voter $i$ manipulates, this is risk-free, as no voter would have an incentive to manipulate when $i$ casts a strategic vote. □

**Theorem 6.12.** *Any positional scoring rule is susceptible to second-order risk-free manipulation.*

*Proof.* We prove this by induction on the number of alternatives. By Theorem 6.8, we know that this must hold for the case where $m = 3$. Now assume that it holds for $m$ alternatives. Let $F$ be a positional scoring rule for $m+1$ alternatives with scoring vector $\mathsf{score} = (\mathsf{score}_1, \ldots, \mathsf{score}_{m+1})$. Take an alternative $x \in X$ and let $F_{-x}$ be the derived subrule. If $F$ is anti-plurality (or a scoring rule equivalent to anti-plurality), then we can apply Theorem 6.11. So, assume that $F$ is not (equivalent to) anti-plurality. Note that in this case, it must hold that $\mathsf{score}_1 > \mathsf{score}_m$. Hence, $F_{-x}$ is a positional scoring rule. Thus, we can apply the inductive hypothesis to infer that there exists a preference profile $\boldsymbol{p}$ over $m$ alternatives such that there exists a voter with a second-order risk-free manipulation. Without loss of generality, let this be voter 1 with manipulative ballot $b_1$. Let $F_{-x}(\boldsymbol{p}) = a$ and $F_{-x}(b_1, \boldsymbol{p}(-1)) = b$. By Lemma 6.10, in $b_1$ the position of alternative $b$ is increased, or the position of alternative $a$ is decreased (or both).

Now consider profile $\overline{\boldsymbol{p}}$, where $x$ is added to the bottom of every preference order in this profile. By definition of $F_{-x}$, it holds that $F(\overline{\boldsymbol{p}}) = a$ and $F(\overline{b_1}, \overline{\boldsymbol{p}(-1)}) = b$. So, the manipulation of voter

1 is still successful. However, to guarantee that this manipulation is risk-free, we need an extra step. Suppose without loss of generality that voter 2 prefers $c$ to $b$, while $a \succ_1^p c$. Then voter 2 is a potential parasitising manipulator. Since the manipulation of voter 1 of the profile $\boldsymbol{p}$ is risk-free, either some voter could already manipulate the profile $\boldsymbol{p}$ and make $c$ win, or voter 2 is not able to manipulate the profile $(b_1, \boldsymbol{p}(-1))$ to make $c$ win. In the first case, the manipulation that makes $c$ win must also be successful under the extended profile $\overline{\boldsymbol{p}}$. Therefore, in that case voter 1's manipulation is risk-free. In the second case, the extension of the ballot might create the opportunity for voter 2 to cast a successful strategic ballot: by putting alternative $x$ between $c$ and $b$, the gap between the score of $c$ and $b$ can be decreased, and the score of $c$ might even exceed the score of $b$ (depending on the scoring vector). In that case, $c$ would win and the manipulation of voter 1 would not be risk-free. To 'secure' voter 1's manipulative ballot $b_1$, she should move alternative $x$ right below $b$. In this way, she increases the gap between the score of alternative $x_b$ and the score of any less preferred alternative, and hence she prevents any potential parasitising manipulation. This is illustrated in Figure 6.6. We conclude that this manipulation is second-order risk-free.

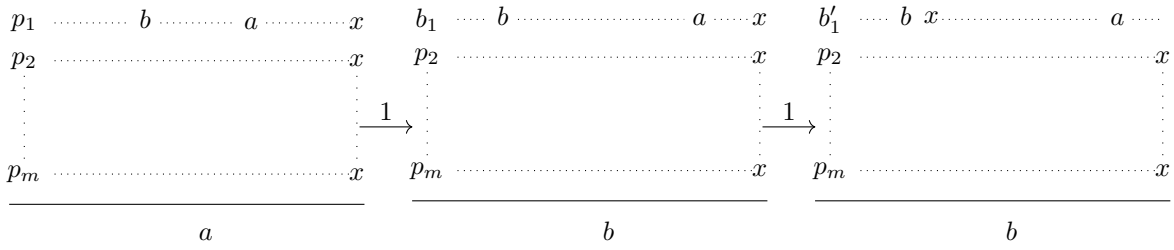



Figure 6.6: Voter 1 has risk-free manipulation $b'_1$

□

Suppose that the voting rule is a positional scoring rule. Let $Y$ be set of alternatives that are at least as bad as $a$ from the perspective of voter $i$. Then if voter $i$ has a strategic ballot in which the positions of the alternatives in $Y$ are not changed, this ballot is a second-order risk-free manipulation.

**Proposition 6.13.** Let $F$ be a positional scoring rule, $\boldsymbol{b}$ some ballot profile with $F(\boldsymbol{b}) = a$ and $b'_i$ a strategic ballot for some voter $i$ such that $F(b'_i, \boldsymbol{b}(-i)) = b$. Let $Y = \{y \mid a \succeq_i^p y\}$. If for all $y \in Y$, $\mathsf{rank}_{b_i}(y) = \mathsf{rank}_{b'_i}(y)$, then $b'_i$ is a second-order risk-free manipulation.

*Proof.* By Lemma 6.10, it must hold that the position of alternative $b$ is strictly higher in $b'_i$ than in $b$, and the score of $b'_i$ must strictly increase. The position of all 'bad' alternatives does not change. Hence, this ballot can not create an opportunity for another voter to parasitise the manipulation and we conclude that $b'_i$ is a second-order risk-free manipulation. □

It is easy to see that by Lemma 6.10, manipulation of plurality under winner-information is always second-order risk-free:

**Corollary 6.14.** If $F$ is the plurality and winner information is given, every manipulation is a second-order risk-free manipulation.

### 6.2.2 Condorcet extensions

In this section, we will show that if the number of voters is even, any non-antagonistic strong Condorcet-extension is susceptible to second-order risk-free manipulation. We will first prove that strong Condorcet extensions are susceptible to second-order risk-free manipulation, and in particular that there exist second-order risk-free manipulations that involve weak Condorcet winners. This implies that the manipulability of strong Condorcet extensions does not (fully) arise from the profiles under which no weak Condorcet winners exist.

**Lemma 6.15.** For $m = 3$ and an even number of voters $n \geq 4$, any strong Condorcet-extension is susceptible to second-order risk-free manipulation, such that the manipulated ballot profile is a Condorcet profile (that is, a profile such that there exists a (weak) Condorcet winner).

*Proof.* Let $n$ be the number of voters and let $X = \{a, b, c\}$. Let $F$ be any strongly Condorcet-consistent social choice function with tie-breaking function $T$, and assume without loss of generality that $T(a, b) = a$. Now consider the following preference profile:

| number of voters | ballot |
|---|---|
| $\frac{1}{2}n - 1$ | $abc$ |
| $\frac{1}{2}n - 1$ | $bca$ |
| 1 | $bac$ |
| 1 | $cab$ |

Under this profile, alternatives $a$ and $b$ are weak Condorcet winners. Hence, $a$ will be the winner. Now, the voter with preference $bac$ has an incentive to manipulate, because if she votes $bca$, then alternative $a$ will no longer win the pairwise majority contest between alternative $c$ and $a$, in which case $a$ would no longer be a weak Condorcet winner. Hence, $b$ would become the unique Condorcet winner and hence the winner of the election. This manipulation is second-order risk-free because no voter has an incentive to manipulate under this ballot profile.

□

We can use this result as the base case for the following theorem:

**Theorem 6.16.** *Any non-antagonistic, strongly Condorcet-consistent social choice function $F$ for at least three alternatives and an even number of voters $n \geq 4$, paired with a tie-breaking choice function $T$ is susceptible to second-order risk-free manipulation.*

*Proof.* We show something stronger, namely that there exists a second-order risk-free manipulation that such that the manipulated profile is a Condorcet profile. We prove this by induction on the number of alternatives. By Proposition 6.15, we know that this must hold for the case where $m = 3$. Now assume that it holds for $m$ alternatives. Let $F$ be a strong Condorcet extension with tie-breaking rule $T$ for $m + 1$ alternatives. Take an alternative $x \in X$ and let $F_{-x}$ be the derived subrule. Now consider a preference profile $\boldsymbol{p}$ over alternatives $X \backslash \{x\}$ and let $\overline{\boldsymbol{p}}$ be the profile $\boldsymbol{p}$ with $x$ attached to the bottom of every preference order. Note that $F_{-x}$ is well defined: in case there are weak Condorcet winners under $\overline{\boldsymbol{p}}$, $T$ chooses an alternative from the set of weak Condorcet winners, but $x$ cannot be a weak Condorcet winner. If there are no weak Condorcet winners, $F(\overline{\boldsymbol{p}}) \neq x$ since $F$ is not antagonistic. Furthermore, $F_{-x}$ is strongly Condorcet-consistent as well: alternative $y \neq x$ is a weak Condorcet winner under $\overline{\boldsymbol{p}}$ if and only if $y$ is a weak Condorcet winner under $\boldsymbol{p}$.

Since $F_{-x}$ is a strong Condorcet extension for $m$ alternatives, we can apply the inductive hypothesis, so we obtain that there exists a voter, say voter 1, with a second-order risk-free manipulation $b_1$, and $(b_1, \boldsymbol{p}(-1))$ is a Condorcet profile. Let $F_{-x}(\boldsymbol{p}) = x_a$ and $F_{-x}(b_1, \boldsymbol{p}(-1)) = x_b$. Then by definition, $F(\overline{\boldsymbol{p}}) = x_a$ and $F(\overline{b_1}, \overline{\boldsymbol{p}(-1)}) = x_b$. So, $\overline{b_1}$ is also a successful manipulation of the profile $\overline{\boldsymbol{p}}$. There are two options: if $x_b$ is a weak Condorcet winner under $\overline{\boldsymbol{p}}$, but loses from $x_a$ in the tie-breaking, the manipulative ballot of voter 1 must result in $x_a$ losing a pair-wise majority contest, in order to eliminate $x_a$. If $x_b$ is not a weak Condorcet winner under $\overline{\boldsymbol{p}}$, then the manipulative ballot of voter 1 must involve making $x_b$ a weak Condorcet winner. Next, we show that voter 1's manipulation must be second-order risk-free.

Since voter 1's manipulation of $\boldsymbol{p}$ was second-order risk-free, voter 2 does not have a parasitising manipulative ballot under $(b_1, \boldsymbol{p}(-1))$. Suppose that there is a voter, say voter 2, that has a parasitising manipulation $b_2$ of the profile $(\overline{\boldsymbol{b_1}}, \overline{\boldsymbol{p}(-1)})$ that makes alternative $x_c$ win. Let $\boldsymbol{b}_X$ denote the ballot profile $(\overline{b_1}, \overline{\boldsymbol{p}(-1)})$. By the inductive hypothesis, we know that under $\boldsymbol{b}_X$, alternative $x_b$ is a weak Condorcet winner. Claim: there exists a manipulation $b_2'$ of voter 2 with $x$ ranked last that makes $x_c$ win. We consider two cases:

**Case 1**: Voter 2 has a manipulation $b_2$ that makes $x_c$ a Condorcet winner under $(b_2, \boldsymbol{b}_X(-2))$. To make $x_c$ an Condorcet winner, voter 2 should report a ballot that results in alternative $x_c$ winning every pairwise majority contest. Under $\boldsymbol{b}_X$, $x_c$ is not a Condorcet winner, so there is at least one pairwise majority contest that $x_c$ loses. This implies that in $b_2$, voter 2 must have swapped alternative $x_c$ with the alternative(s) from which $x_c$ is currently losing. However, $x_c$ does not lose against alternative $x$, because that alternative is ranked last in every ballot in $\boldsymbol{b}_X$. Hence, let $b_2'$ be defined as $b_2$ with $x$ shifted to the last position. Then $b_2'$ should still make $x_c$ a Condorcet winner, and hence the manipulation $b_2'$ makes $x_c$ win.

**Case 2:** Voter 2 has a manipulation $b_2$ that dethrones $x_b$ as Condorcet winner. Claim: there exists a manipulation $b_2'$ of voter 2 with $x$ ranked last that dethrones $x_b$ as Condorcet winner. To dismiss $x_b$ as Condorcet winner, voter 2 should report a ballot that results in alternative $x_b$ losing at least one pairwise majority contest. This implies that in $b_2$, voter 2 must have swapped alternative $x_b$ with at least one alternative, such that $x_b$ now loses the pairwise majority contest with that alternative. This alternative cannot be alternative $x$, because $x$ is ranked last in every other ballot in $\boldsymbol{b}_X$, so $x_b$ will still not lose from $x$. We can define the ballot $b_2'$ as $b_2$ with $x$ shifted to the last position, and $b_2'$ should still dethrone $x_b$, which proves the claim.

Note that since $x$ is ranked last in $b_2$, $b_2' = \overline{b}$ for some ballot $b$ of voter 2, and hence voter 2 has a strategic manipulation of the profile $(b_1, \boldsymbol{p}(-1))$ that elects $x_c$. This contradicts the assumption that voter 2 does not have a parasitising manipulation of that profile. Thus, voter 2 cannot have a manipulation of $\boldsymbol{p}_X$ that makes $x_c$ win.

We conclude that there is no voter with a parasitising manipulation, and hence the manipulation of voter 1 is second-order risk-free. This proves that for any number of alternatives, there is a preference profile that is sensitive to second-order risk-free manipulation. Hence, every non-antagonistic strongly Condorcet-consistent social choice function $F$ is susceptible to second-order risk-free manipulation if the number of voters is even. □

This proof does not work for voting situations with an odd number of voters. With an odd number of voters, Condorcet winner are always unique. Therefore, there is no manipulation that replaces the current Condorcet winner by a new Condorcet winner. In that case, a pivotal voter would have to swap at least two alternatives (the current Condorcet winner and the new

one), but she has never an incentive to do that, because then a less preferred alternative would win. Hence, a successful manipulation of a Condorcet extension should either change the ballot profile from a Condorcet profile to a non-Condorcet profile, or vice versa, or it should change a non-Condorcet profile to another non-Condorcet profile. We were not able to show that in the first case, such a manipulation must be second-order risk-free, because the manipulated profile is non-Condorcet. The Condorcet criterion only requires that Condorcet winners are elected when they exist, so under non-Condorcet profiles, we do not know how the social choice function behaves. Therefore, we cannot prove that there must exist a second-order risk-free manipulation.

### 6.2.3 Other classes of social choice functions

In the previous sections, we showed that any non-dictatorial and surjective iterative voting procedure for three alternatives is susceptible to second-order risk-free manipulation. We also showed that for any number of alternatives, positional scoring rules and non-antagonistic Condorcet extensions are susceptible to second-order risk-free manipulations (the latter if the number of voters is even). It remains an open problem whether for any number of alternatives, any number of voters, and any non-dictatorial and surjective iterative voting procedure, there exists a situation in which a voter has a second-order risk-free incentive to manipulate. We conjecture that this is the case, but we were not able to show it. The inductive argument that is used in the proofs for positional scoring rules and Condorcet extensions cannot be applied here. Suppose that $F$ is any non-dictatorial and surjective social choice function for $m$ alternatives. We consider an alternative $x \in X$ and the derived subrule $F_{-x}$, which is only defined for ballot profiles over $X$ in which every voter ranks $x$ last. We suppose that there is a second-order risk-free manipulation under the subrule $F_{-x}$ of the preference profile $\boldsymbol{p}$ over alternatives $X \backslash \{x\}$. It is easy to show that this manipulation is also a successful manipulation of the extended profile $\overline{\boldsymbol{p}}$, under the original social choice function $F$. However, we cannot guarantee that this manipulation is risk-free: we know (by induction) that there is no parasitising manipulation in which $x$ is ranked last, but the inductive hypothesis says nothing about ballots in which $x$ is *not* ranked last. This means that there could exist a ballot in which $x$ is not ranked last, that is a parasitising manipulation. Therefore, we conclude that this issue remains an open problem.

Even though it seems that most reasonable social choice functions are even manipulable by voters that have a more farsighted view, higher-order reasoning about iterative voting procedures could also shine a new light on another aspect of iterative voting, namely convergence. We will discuss this topic in the next section.

## 6.3 Convergence of iterative voting procedures

An important question for iterative voting procedures is whether an iterative election will always converge. As discussed in Chapter 2, for positional scoring rules it is known that only for iterative elections with the plurality rule or the anti-plurality rule, convergence is guaranteed. The new perspective on manipulation introduced in this Chapter allows us to reconsider these results: if we assume that voters will only cast a strategic ballot when they know that their manipulation cannot be parasitised in the next round, it becomes harder to find a risk-free manipulation. Which implications does this have for the convergence of iterative voting rules? We show that in the case of three alternatives, the Borda rule will always converge.

**Proposition 6.17.** An iterative Borda voting procedure with lexicographic tie-breaking always

converges when $m = 3$.

*Proof.* The idea of the proof is that we show that in every sequence of manipulations, the election tends to become a two-party system: if an alternative does not win in round 1 or round 2, it will never become a winner in a later round. Therefore, the iterative procedure must converge.

Let $X = \{a, b, c\}$. There are (at most) six possible preference orders: *abc*, *acb*, *bac*, *bca*, *cab* or *cba*. First, suppose that under the truthful ballot $\boldsymbol{p}$ profile, $c$ is the winner. This implies that $\mathsf{score}(c)_{\boldsymbol{p}} > \mathsf{score}(a)_{\boldsymbol{p}}, \mathsf{score}(b)_{\boldsymbol{p}}$. In the first round, the ballot profile is $\boldsymbol{b}^1 = \boldsymbol{p}$. Let voter 1 be the first voter to change her ballot. So, voter 1 has a second-order risk-free manipulation $b_1 \neq \boldsymbol{p}(1)$. We consider two cases:

**Case 1**: Voter 1 has a strategic manipulation that makes $b$ the new winner. Then her truthful preference is either *abc* or *bca*, and in both cases her manipulative ballot is *bac*. Claim: for every $t \geq 3$, $F(\boldsymbol{b}^t) \neq a$. We prove this by induction. If voter 1's truthful preference is *abc*, it must hold that $\mathsf{score}_{\boldsymbol{p}}(b) = \mathsf{score}_{\boldsymbol{p}}(c) - 1$. Under the new profile, 1 transfers one point from $a$ to $b$, so $\mathsf{score}_{\boldsymbol{b}^2} a \leq \mathsf{score}_{\boldsymbol{b}^2}(b) - 2$, and $\mathsf{score}_{\boldsymbol{b}^2}(b) = \mathsf{score}_{\boldsymbol{b}^2}(c)$. First, note that since the difference in the score of alternative $a$ and $b$ under profile $\boldsymbol{b}^2$ is at least 2, so an individual voter $j$ could only make $a$ win if $a$ is currently ranked last in $j$'s ballot. However, since under $\boldsymbol{b}^2$, all voters with $a$ ranked last cast a truthful ballot, those voters do not have an incentive to make $a$ win. So, $a$ will not win under ballot profile $\boldsymbol{b}^3$. If voter 1's truthful ballot is *bca*, she will only manipulate after round 2 if she knows that no voter can make alternative $a$ win at $t = 3$, otherwise the manipulation would not be risk-free. This proves the base case.

Now suppose that under all ballot profiles $\boldsymbol{b}^k$ for $k \leq t$, $a$ does not win. This implies that in every round $k \leq t$, the best response of a voter either has the goal to change the winner from $b$ to $c$, or to change the winner from $c$ to $b$. Voters that prefer $b$ to $c$ have true preference *bca*, *bac* or *abc*. If a voter's true preference is *bac*, she cannot manipulate. If a voter's true preference is *bca* or *abc*, her manipulative ballot must be *bac*. In the same way, we can argue that for voters who prefer $c$ to $b$, if they can manipulate, their strategic ballot must be *cab*. So, in every round $k \leq t$, a voter changes her ballot to either *bac* or *cab*.

Furthermore, since $a$ never wins, it holds that $\mathsf{score}_{\boldsymbol{b}^k}(a) < \mathsf{score}_{\boldsymbol{b}^k}(b), \mathsf{score}_{\boldsymbol{b}^k}(c)$ for all $k \leq t$. Now suppose that there is a voter $i$ that can make $a$ win at time $t+1$, so voter $i$ has a strategic manipulation $b_i'$ such that $F(b_i', \boldsymbol{b}^t(-i)) = a$. This implies that (ii) $\mathsf{score}_{\boldsymbol{b}^t}(a) = \mathsf{score}_{\boldsymbol{b}^t}(b) - 1$ and $\mathsf{score}_{\boldsymbol{b}^t}(b) \geq \mathsf{score}_{\boldsymbol{b}^t}(c)$ or (ii) $\mathsf{score}_{\boldsymbol{b}^t}(a) = \mathsf{score}_{\boldsymbol{b}^t}(c) - 1$ and $\mathsf{score}_{\boldsymbol{b}^t}(c) > \mathsf{score}_{\boldsymbol{b}^t}(b)$.

(i) In the first case, $b$ is the winner under $\boldsymbol{b}^t$. That means that after round $t-1$, there was a voter $j$ with an incentive to make $b$ win, and that his strategic ballot must have been *bac*. If his truthful preference is *abc*, then it must hold that $\mathsf{score}_{\boldsymbol{b}^{t-1}}(a) = \mathsf{score}_{\boldsymbol{b}^{t-1}}(b) + 1$. However, then $F(\boldsymbol{b}^{t-1}) = a$, which contradicts our assumption that $a$ is not a winner in round $k \leq t$. Hence, $j$'s truthful preference must be *bca*. But this contradicts our assumption that manipulations are second-order risk-free: if voter $j$ with preference *bca* manipulates after round $t-1$ to make $b$ win in round $t$, he will only do this when alternative $a$ cannot win in round $t+1$. Hence, it must hold that $F(\boldsymbol{b}^{t+1}) \neq a$.

(ii) In the second case, $c$ is the winner under $\boldsymbol{b}^t$. That means that in round $t-1$, there was a voter $j$ with an incentive to make $c$ win, and that his strategic ballot must have been *cab*. If his truthful preference is *acb*, then $\mathsf{score}_{\boldsymbol{b}^{t-1}}(a) = \mathsf{score}_{\boldsymbol{b}^{t-1}}(c) + 1$. However, then $F(\boldsymbol{b}^{t-1}) = a$, which contradicts our assumption that $a$ is not a winner in round $k \leq t$. Hence, $j$'s truthful preference must be *cba*. But this contradicts our assumption that

manipulations are second-order risk-free: if voter $j$ with preference $cba$ manipulates after round $t-1$ to make $c$ win in round $t$, he will only do this when voter $a$ cannot win in round $t+1$. Hence, it must hold that $F(\boldsymbol{b}^{t+1}) \neq a$.

This proves our claim. Since $a$ will never be a winner in any round $t \geq 3$, we can apply the same argumentation to obtain that since $b$ and $c$ will be the only winners, the only manipulative ballots can be $bac$ and $cab$. After every voter with a truthful preference not equal to $bac$ or $cab$ changed her ballot to either $bac$ or $cab$, the voting process converges.

**Case 2**: Voter 1 has a strategic manipulation that makes $a$ the new winner. Following the same argumentation as in Case 1, we can show by induction on the number of rounds that $b$ will never become a winner. It also follows that $abc$ and $cba$ are the only possible manipulative ballots, and hence that the voting process will converge.

We can give a similar argument if the winner under the truthful profile is $a$ or $b$. If $a$ is the winner under the truthful profile, it must hold that $\mathsf{score}(a)_{\boldsymbol{p}} \geq \mathsf{score}_{\boldsymbol{p}}(b), \mathsf{score}_{\boldsymbol{p}}(c)$. If voter 1 has a strategic manipulation that makes $b$ the new winner, her manipulative ballot must be $bca$. Again, we can show by induction that $c$ will never become a winner. In the same way, we can show that if voter 1 has a strategic vote that makes $c$ win, $b$ will never win. Finally, if $b$ is is the winner under the truthful profile, it must hold that $\mathsf{score}_{\boldsymbol{p}}(b) > \mathsf{score}_{\boldsymbol{p}}(a)$ and $\mathsf{score}_{\boldsymbol{p}}(b) \geq \mathsf{score}_{\boldsymbol{p}}(c)$. Again, we distinguish two cases, one in which alternative $a$ wins after the first manipulation, and one in which $c$ wins after the first manipulation. In both cases, we show that the other alternative (respectively $c$ and $b$) will never become a winner. □

This result shows that in the case of three alternatives, iterative voting procedures with Borda as voting rule always converge. It remains an open problem whether similar results are attainable for the general case of $m$ alternatives, or whether similar results hold for other voting rules. As the convergence of iterative voting procedures is not the main focus of this thesis, these question are left for future research.

## 6.4 Concluding remarks

Our investigation concentrated on the hope that agents who are able to look ahead and predict future manipulations of their peers, would realise that their manipulations would be unsuccessful in the end, and that they would therefore choose to remain truthful themselves. Unfortunately, this turned out not to be the case. For two important classes of social choice functions, positional scoring rules and (non-antagonistic) Condorect extensions, we showed that even if voters do not want to risk a worse outcome in the consecutive round of voting, it is still possible to manipulate. It remains an open problem whether this holds for every non-dictatorial and surjective social choice function.

In this chapter, we focused on second-order reasoning agents, that is, agents that are able to look one step ahead. Although they are able to reason about the future, these agents still have a very short-term view. In future work, it would be interesting to consider voters with a more long-term view. One could define a model of rationality in which agents take into account the long-term effects of their choices. Suppose that under the truthful profile $\boldsymbol{p}$, voter $i$ has a manipulative ballot. After she reports this ballot, a sequence of strategic votes begins, eventually terminating in a Nash equilibrium. In this scenario, voter $i$ will be more interested in comparing the winner under the truthful profile $\boldsymbol{p}$ with the outcome after the iterative voting process has converged,

if $i$ casts a strategic vote that triggers a sequence of other manipulations. If the winner under the Nash equilibrium is worse than $F(\boldsymbol{b})$, and $i$ realises that this is a consequence of her own manipulative ballot, it would be better for her to not cast that vote. This is related to the idea of second-order Nash equilibria, introduced by Bilò and Flammini (2011) and based on non-myopic equilibria discussed in Brams and Wittman (1981) and Kilgour (1984). In future work, it would be interesting to analyse the strategic incentives of voters that are able to predict in which Nash equilibrium the election will result.

Here, we only concentrated on parasitising manipulations: the initial manipulator refrains from strategising, because a second manipulation could be harmful. Another direction would be to investigate other intentions of voters for casting a strategic vote. If a voter is able to predict what other voters will do in the future, she might adapt her manipulative behaviour. For example, she may consider a strategic vote that is initially not beneficial for herself, but if that manipulation would provoke a manipulation by another voter that is beneficial for her, she might still cast the strategic vote. So, she attempts to vote strategically in the sense of misleading other voters. Or, following the same kind of reasoning, she might cast a vote that is not resulting in the best outcome from her perspective, but by casting that strategic vote, she blocks a manipulation of another voter that would result in an even worse outcome. For future work, we believe that it would be interesting to investigate these new types of incentives for voters to manipulate an election.

An interesting direction for future research would be to explore whether iterative voting procedures are convergent, or even immune to manipulation if voters have partial information. Meir et al. (2014) and Meir (2015) introduce a behavioural game-theoretic model for iterative voting under uncertainty. Meir et al. (2014) proved that if all voters have the same uncertainty level, start by voting truthfully, then they always converge to a voting equilibrium. It might be interesting to investigate how higher-order reasoning voters will behave in such scenarios.

# Chapter 7

# Strategic communication

Recall the example from the introduction: there are four voters, three alternatives and the voting rule is Borda, and the epistemic model is given in Figure 7.1. Alice and Bob both have preference order *abc*, although they do not know this from each other. Both voters consider it possible that the other voter has preference order *acb*. If both voters report ballot *abc*, then *b* will be the winner. If one of them reports the ballot *abc* and the other reports *acb*, then *a* will be the winner. However, if both voters report *acb*, then *c* will be the winner. Hence, there is no voter with an incentive to manipulate. Alice could improve the situation by voting *acb* if Bob votes *abc*, but she also consider the situation where Bob votes *acb* possible. In that situation, the outcome of the election would be worse. This also holds for Bob. No matter whether Alice and Bob's true preferences are *abc* or *acb*, it is beneficial for both voters if one of them votes *abc* and the other *acb*, because then *a* will win, and in any case *a* is their favourite alternative. Although Bob does not know whether Alice prefers *b* to *c* or vice versa, he knows that if Alice would know that he votes *abc*, then she would definitely report the ballot *acb* (see Figure 7.2). In that case, the outcome of the election would be strictly better for both Alice and Bob than when Bob had not revealed his ballot to Alice. In this situation, Bob does not have an incentive to manipulate, but he rather has an incentive to *communicate*.

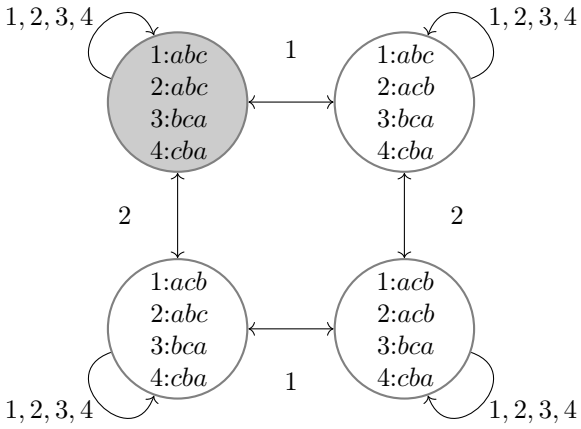



Figure 7.1: Both manipulative voters are not able to vote strategically

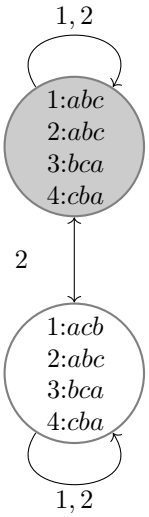



Figure 7.2: Alice learns that she has a successful strategic vote

In this chapter, we will explore the idea of strategic communication in voting. Pre-vote communication can help voters to coordinate their ballots, for example when they try to form a coalition. Assuming that agents are strategic and do anything to obtain the best outcome, which role does communication play in strategic behaviour? Under which circumstances do voters have an incentive to communicate information, and which effect has this information exchange on the outcome of the election?

## 7.1 Sharing factual information

The simplest form of communication is the exchange of factual statements. In this section, we will discuss situations in which a voter believes that sharing personal information with other voters might result in a better outcome of the election.

First, we assume that voters will not cheat by intentionally spreading false information: a voter only shares information that she thinks is true. This is a common assumption in game-theoretic models of reasoning. The most basic type of information exchange treats the source of the information as fully reliable, and this is common knowledge. This means that if voter $i$ shares $\varphi$ with voter $j$, voter $j$ does not consider it possible that $\varphi$ is not true. Furthermore, voter $i$ knows that voter $j$ believes this, and voter $j$ again knows that voter $i$ knows that $j$ believes this, etcetera.

The communicator assumes that the voters who receive her information do not do not care about the reason why she sent that information: they do not question the intentions of the communicator. So, we assume that a potential communicator only reasons about the effect of the information update on the voting behaviour of the other agents: if a group of voters $G$ knows $\varphi$, what is their best strategy? By sharing information, a communicator wants to stimulate the receiver(s) of the message to vote in a certain way.

In this chapter, we do not take into account higher-order reasoning about incentives of other voters to communicate. In theory, if a voter $i$ shares some information with $j$, this could create an incentive to for voter $j$ to share information with voter $k$, which again could create an incentive for a voter $k$ to communicate something, etcetera. We assume that the communicator is not able to predict the consequences of her communication for the incentives of other voters to communicate, but that she will only analyse the consequences of the new information to the *voting* behaviour of other agents.

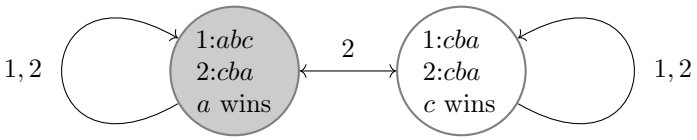
### 7.1.1 When is it beneficial to share information?

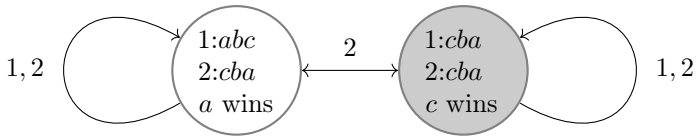
We first have to shine a light on the question whenever communicating a piece of information is beneficial for a voter in the voting framework presented in this thesis.

We have assumed that voters are risk-averse manipulators: even if there is just a single scenario in which a manipulation would result in an unfavourable outcome, the voter will just vote truthfully. Since we have a dynamic epistemic model without belief revision, there are two scenarios in which a voter would change her vote: if she learns that she can safely cast a strategic vote (by ruling out situations in which the strategic vote would result in an unfavourable outcome), or if she learns that her strategic vote in fact will not be successful (by ruling out situations in which the strategic vote would strictly improve the outcome).

Taking these assumptions into account, it follows that a voter only has an incentive to share information with some other voters if she thinks that the new information will stimulate her peers to vote in such a way that it is beneficial for her. Which information should a communicator share in order to stimulate a fellow voter to adopt a certain idea? At first sight, it may seem that providing truthful information can never be harmful. However, it turns out that this is not true: it can sometimes be useful to conceal some information, as the following example will illustrate.

**Example 7.1.** This example is from Bakhtiari et al. (2018). The voting rule is plurality, with tie-breaking according to $b \triangleright a \triangleright c$. In the first figure, voter 1 does not have an incentive to communicate her ballot to voter 2, since voter 2 would then have an incentive to strategically vote $bca$ and $b$ would be the winner. Here, voter 2's uncertainty is beneficial for voter 1, because voter 2 considers it possible that they both have preference order $cba$, in which case a manipulation would be harmful: by voting $bca$, she would make alternative $b$ win.






When a voter $i$ is uncertain about the which state is the actual one, it may be the case that she has a strategic vote that will improve the outcome in some of the situations that she considers possible, but the outcome in the actual state. So in fact, the strategic vote is unsuccessful. However, the strategic vote of $i$ is neither harmful, because the winner under the manipulated ballot profile will be identical to the winner under the truthful profile. This implies that even if a fellow voter $j$ knows all this, he does not have an incentive to inform $i$ that her manipulation will in fact turn out unsuccessful, because this will have no effect on the outcome of the election. Therefore, we assume that a voter only has an incentive to communicate if she believes that the information will stimulate the receiver of the message to vote strategically.

**Definition 7.2** (Communication update model)**.** Let $\varphi$ be a sentence in the logic. We want to define the update model in which $\varphi$ is secretly announced to a group of voters $G \subseteq N$. After the announcement of $\varphi$, $\varphi$ is common knowledge within the group $G$, while none of the other voters suspect anything. Let $U^G_\varphi = (\Sigma, R, pre, post)$ be the following update model:

(i) $\Sigma = (e_1, e_2)$

(ii) For every $j \in G$, $R(j) = \{(e_1, e_1), (e_2, e_2)\}$
For every $l \notin G$, $R(l) = \{(e_1, e_2), (e_2, e_2)\}$

(iii) $pre(e_1) = \varphi$
$pre(e_2) = \tau$

(iv) $post(e_1) = post(e_2) = \mathsf{id}$

(v) The actual event is $e_1$.

We call $U^G_\varphi$ a $\varphi^G$-communication update model. Note that if $G = N$, this update model represents the public announcement of $\varphi$.

In the following definition, we will formalise the idea of having an incentive to communicate $\varphi$. Following the line of this thesis, we assume that voters are risk-averse communicators. This means that a voter will only choose to strategically communicate certain information if in at least one case, the outcome of the election will be strictly better than in the initial situation. In every other situation she considers possible, communicating the information should not make the situation worse. To make the situation strictly better from the perspective of the communicator, at least one of the voters that receive her message has to decide to strategise in such a way that it is beneficial for the communicator. So, her information must create an incentive to manipulate. In all other cases, the information $\varphi$ must not lead to voters manipulating in a way that the outcome is worse from the perspective of the manipulator. The following definition formalises this idea.

**Definition 7.3.** Let $F$ be a voting rule and let $M = (S, V, R)$ be an epistemic model for strategic voting with actual state $s^*$. A voter $i$ has an *incentive to strategically communicate $\varphi$ to a group of voters $G \subseteq N$* with $i \in G$ if:

(i) For all $s \in S$ with $s^* R_i s$, $M, s \vDash \varphi$

(ii) There exists $s \in S$ with $M, s \vDash y$ such that the following hold:

- $s^* R_i s$
- $\bigtimes_{j \in G} \mathcal{S}_j(M \otimes U_\varphi^G, (s, e_1)) \times \bigtimes_{l \notin G} \boldsymbol{b}_s(l))$
- $M, s \vDash x \succ_i^p y$

(iii) For all $s \in S$ with $s^* R_i s$ it holds that:
if $M, s \vDash y$ and $M, s \vDash (\bigtimes_{j \in G} \mathcal{S}_j(M \otimes U_\varphi^G, (s, e_1)) \times \bigtimes_{l \notin G} \boldsymbol{b}_s(l)) \to x$, then $M, s \vDash x \succeq_i^p y$.

Now that we have a formal notion of having an incentive to communicate, we can investigate under which circumstances such an incentive exists.

### 7.1.2 Information-stable models

When studying strategic manipulation in voting under partial information, we should concentrate our search to situations in which no voter has an incentive to communicate some information. Suppose that we have a voting situation that is strategyproof, but that there is a voter with an incentive to communicate some piece of information to a group of fellow voters. So, the communicator thinks that this new information might result in a manipulation of the election by some voters in such a way that the outcome of the manipulated profile is beneficial for the communicator. Although the initial voting situation was strategyproof, the final outcome of the election could still be the result of a manipulation. We call models in which no voter has an incentive to communicate *information-stable*.

**Definition 7.4.** An epistemic model for strategic voting $M$ with actual state $s^*$ is *information-stable* if no voter $i \in N$ has an incentive to strategically communicate some sentence $\varphi$ to some group of voters $G \subseteq N$.

**Proposition 7.5.** Every model $M \in \mathcal{M}_{\text{full info,sincere}}$ is information-stable.

*Proof.* This follows directly by the fact that all voters already have full information. □

**Proposition 7.6.** Every model $M \in \mathcal{M}_{\text{ign,sincere}}$ for plurality is information-stable if $n \geq 3$.

*Proof.* Let $M \in \mathcal{M}_{\text{ign,sincere}}$ be a model for the plurality rule. In $M$, every voter only knows her own preference and ballot. Suppose that voter $i$ considers to communicate her preference to a group $G \subseteq N$ and let $j \in G$, $j \neq i$. We show that voter $j$ does not have an incentive to manipulate. Suppose that voter $j$ has preference $p_j = x_1 \succ x_2 \succ \ldots \succ x_m$. After voter $j$ learns $\varphi$, voter $j$ considers it possible that there are exactly $\lfloor \frac{n}{2} \rfloor$ other voters with alternative $x_1$ as their top choice. In that case, $x_1$ will win if voter $j$ votes truthfully, and $x_1$ will not win if voter $j$ ranks another alternative first. So, no matter the ballot of voter $i$, voter $j$ will not deviate from her truthful ballot. □

**Proposition 7.7.** Every model $M \in \mathcal{M}_{\text{winner,sincere}}$ for plurality is information-stable if $n \geq 3$.

*Proof (sketch).* Let $M \in \mathcal{M}_{\text{winner,sincere}}$ be a model for plurality paired with some tie-breaking rule $T$. Assume, for the sake of contradiction, that voter $i$ has an incentive to communicate $\varphi$ to a group of voters $G$. Recall that she only shares information she knows, so either $\varphi$ contains information about her own preference or ballot, or $\varphi$ is already common knowledge. If $\varphi$ is not common knowledge, it must hold that $M, s^* \vDash (\boldsymbol{b}_{s^*}(i) \wedge \boldsymbol{p}_{s^*}(i)) \to \varphi$. In other words: for any $\varphi$ that voter $i$ is able to send, $\varphi$ cannot contain more information than the ballot and the preference of voter $i$. We show that even if voter $i$ shares all her personal information with $G$, incentives of the voters in $G$ to manipulate will not change. Hence, let $\varphi = \boldsymbol{b}_{s^*} \wedge \boldsymbol{p}_{s^*}$. As discussed before, voter $i$ can only have an incentive to communicate if she believes that her information will stimulate another voter to manipulate. Without loss of generality, we assume that she tries to stimulate another voter to vote for the alternative that she ranks first.

Suppose that $F(\boldsymbol{b}_{s^*}) = x$. Let $j \in G$ and let voter $j$'s preference be $p_j = y \succ z \succ \ldots$, for $y \neq x$ (if $y = x$, $j$ can certainly not be stimulated to manipulate). Since voter $j$'s favourite alternative is not the winner, if $z \neq x$, she has an incentive to vote for her second choice $z$, because she considers it possible that she is pivotal with respect to alternative $z$. Let $b'_j$ a ballot order with $z$ ranked first. Without loss of generality, assume $T(x, z) = x$[1]. Now there are three cases:

Case 1: $\mathsf{rank}_{\boldsymbol{b}_{s^*(i)}}(z) = 1$. Then there exists $s \in S$ with $s^* R_j s$ such that $j$ is pivotal with respect to alternative $z$:

$$|\{k \in N \mid \mathsf{rank}_{\boldsymbol{p}_s(k)}(z) = 1\}| = |\{k \in N \mid \mathsf{rank}_{\boldsymbol{p}_s(k)}(x) = 1\}|,$$

and since $i \in \{k \in N \mid \mathsf{rank}_{\boldsymbol{p}_s(k)}(z) = 1\}$, $M, s \vDash \varphi$. Hence, for $(s, e_1) \in M \otimes U_\varphi^G$ it holds that $(s^*, e_1) R_i (s, e_1)$ and voter $j$ is still pivotal with respect to alternative $z$. Hence, there is no ballot that could dominate $b'_j$, so she will still vote $b'_j$ in $M \otimes U_\varphi^G$.

Case 2: $\mathsf{rank}_{\boldsymbol{p}_{s^*(i)}}(x) = 1$. Then there exists $t \in S$ with $s^* R_j t$ such that $j$ is pivotal with respect to alternative $z$:

$$|\{k \in N \mid \mathsf{rank}_{\boldsymbol{p}_t(k)}(z) = 1\}| = |\{k \in N \mid \mathsf{rank}_{\boldsymbol{p}_t(k)}(x) = 1\}|,$$

and since $i \in \{k \in N \mid \mathsf{rank}_{\boldsymbol{p}_t(k)}(x) = 1\}$, $M, s \vDash \varphi$. Hence, for $(t, e_1) \in M \otimes U_\varphi^G$ it holds that $(s^*, e_1) R_i (t, e_1)$ and voter $j$ is still pivotal with respect to alternative $z$. Hence, there is no ballot that could dominate $b'_j$, so she will still vote $b'_j$ in $M \otimes U_\varphi^G$.

Case 3: There is alternative $a \neq x, z$ with $\mathsf{rank}_{\boldsymbol{b}_{s^*(i)}}(a) = 1$. Then there exists $u \in S$ with $s^* R_j u$ such that $\mathsf{rank}_{\boldsymbol{b}_u(i)}(a) = 1$ and $j$ is pivotal with respect to alternative $z$:

$$|\{k \in N \mid \mathsf{rank}_{\boldsymbol{p}_u(k)}(z) = 1\}| = |\{k \in N \mid \mathsf{rank}_{\boldsymbol{p}_u(k)}(x) = 1\}|,$$

Since $M, s \vDash \varphi$, we have $M, s \vDash pre(e_1)$ and hence for $(s, e_1) \in M \otimes U_\varphi^G$, it holds that $(s^*, e_1) R_i (u, e_1)$ and voter $j$ is still pivotal with respect to alternative $z$. Hence, there is no ballot that could dominate $b'_j$, so she will still vote $b'_j$ in $M \otimes U_\varphi^G$.

So, learning $\varphi$ will not stimulate voter $j$ to change her ballot, because she always considers it possible that she is pivotal with respect to alternative $z$, and hence she will vote in favour of $z$. Since $j$ was arbitrary, it follows that no agent in $G$ will change her vote after learning $\varphi$. Hence, it is not beneficial for voter $i$ to share all her personal information, namely her ballot and her

[1] If $T(x, z) = z$, we can give a similar proof but with one extra voter who ranks $x$ first.

preference. But then, it will also not help to share a information that is weaker. We conclude that $M$ is information-stable. □

These results show that sharing information by a single agent will, in many general cases, not stimulate other voters to manipulate. One reason for this could be that when a group tries coordinate ballots and tries to manipulate the election, there are cases where common knowledge is necessary in order to safely manipulate the election. The group of voters must avoid situations that are similar to the well-known coordinated attack problem (Halpern, 1986) by establishing a way of communication that makes the voting strategy common knowledge among the members of the group, but not known to any voter that is not in the group.

In most cases however, the lack of incentives to communicate is caused by the fact that even though the voters receive personal information from one of their fellow voters, they still consider it possible that their current ballot (truthful or insincere) is pivotal, and hence, the new information will not change their minds. This contradicts our expectations: intuitively, if the voting rule is plurality, we would expect voters with winner information to try to coordinate with other voters whenever they are not happy with the winner under the truthful profile. So, apparently, sharing factual information is not helpful when voters are trying to form coalitions. To get a better understanding of strategic communication between voters, we need a different notion of an incentive to communicate. In the next section, we will discuss some possible directions.

## 7.2 Normative communication

In order to understand coordination, we need a different type of communication. In the previous section, we saw that only sharing factual information will not help voters to coordinate their votes to form a coalition. In order to form a coalition or to coordinate ballots in some way, agents must be able to communicate their ideas about how an agent *should vote* instead of just factual information about how an agent currently votes. Game theorists have started to formalise the role of communication in strategic settings, e.g. Rabin (1990), Parikh (1991) and Farrell (1993). Franke and van Rooij (2015) explore psychological and social aspects of strategic communication in games. Those ideas of rationality and strategic interaction can be used to model communication with a more normative nature.

Farrell (1988) uses the notion of a *suggestion*. In terms of voting theory, if $G \subseteq N$, a suggestion is a (partial) ballot profile $\boldsymbol{b}(G)$ specifying, for each player $i \in G$, a ballot order $\boldsymbol{b}(i)$. We can interpret this as a piece of information, containing a proposal how every voter in $G$ should vote given a model $M$. A suggestion is *consistent* if every ballot suggested to each voter is a best strategy, given that every other voter follows the suggestion. The notion of a suggestion seems appropriate in the case of a group of like-minded voters, and when the fact that they have similar preferences is common knowledge in that group. In order to coordinate their ballots, one member of the group must utter a suggestion, and if it is the case that if every voter votes accordingly, their favourite alternative wins, then every voter in the group will follow the suggestion. A higher-order reasoning voter would check whether following the suggestions is rational for other agents, before determining whether it would be strategic for herself to follow the suggestion.

However, it is not straightforward how we can formally incorporate this idea in our framework. Our dynamic epistemic model can deal with complicated forms of communication regarding facts, but it is not possible to make a distinction between facts and suggestions in the language. In our framework, agents are able to change their ballot, but we assume that they only change

their ballot whenever they are certain that it will (weakly) improve the outcome of the election from their perspective. This means that in every state an agent considers possible, the new ballot must result in an outcome that is at least as good as the outcome when the agent reports her original ballot. In reality however, an agent may have good reasons to be less risk-averse, for example because a credible and like-minded fellow voter does a suggestion how to vote. In order to model this, we need a framework that is more flexible with respect to the ballots a voter commits to. Van Benthem and Liu (2007) provide a model that deals with *preference dynamics*. They argue that statements not only update our current knowledge, but also have other dynamic effects. The idea of this framework is that preferences are not static, but that they can change through commands of moral authorities, suggestions from friends, or just changes in our own evaluation of the world and our possible actions. In voting theory, it is generally assumed that preferences are static, but in future work, ideas from this framework could be applied to design a model that has richer *ballot dynamics*.

Another way to model coalition forming and exchange of (strategic) ideas between voters is to see the set of voters as a *social network*. In Baltag et al. (2018), the idea is that agents are socially connected to each other and that their ideas and behaviour are socially influenced. Adopting a certain opinion or behaviour, or a certain strategic ballot in our case, is contagious: agents adopt new behaviour when the fraction of the people in their network who have already adopted it meets a certain *threshold*. In Baltag et al. (2018) the notion of an *epistemic treshold* is introduced and agents are assumed to only adopt a certain opinion if they have sufficient information about their neighbours. Our notion of changing a ballot could be related to an epistemic treshold in a natural way, because as we have seen in this thesis, voters must have enough information about their fellow voters before they commit to a certain strategic ballot, in particular when they try to form a coalition. In order to understand under which circumstances agents start following each other's strategic suggestions, for example suggestions on how to vote, the process of how social networks within a group of voters are formed is particularly interesting. Smets and Velázquez-Quesada (2018) propose a model for this. An agent is classified by the different features she may have. Following to the *similarity* approach of Smets and Velázquez Quesada, the more similar two agents are, the more likely it is for them to end up in the same social network. In the setting of voting, the preferences of a voter can be modelled as her features. We believe that modelling a group of voters as a social network is a promising approach to get a deeper insight into the ballot coordination of groups of voters.

# Chapter 8

# Conclusion and future research

## 8.1 Conclusion

The main goal of this thesis was to analyse strategic manipulation of higher-order reasoning voters. We presented a dynamic epistemic model for strategic voting and its corresponding sound and complete logic for strategic voting. We showed that this model fits many different types of voting settings. For the classical setting, in which it is often assumed that there is just a single manipulator who assumes that every other voter will report a sincere ballot, we defined classes of models in which voters have partial information about the preferences of their peers.

The focus of this thesis was higher-order reasoning: reasoning about the reasoning of other voters. Voters are likely to realise that other voters may reason strategically too, and therefore choose the best strategy given what they know about the ballots of the other voters. When a voter realises that fellow voters may reflect on her own strategic behaviour as well, she will start reasoning about the reasoning of other voters.

We first discussed the well-known phenomenon of safe manipulation. If a coalition of voters with the same preference tries to manipulate the election, we say that their manipulation is safe if no matter which subgroup of the coalition casts a strategic vote, the outcome will never be worse than under the truthful profile. We argued that in order to determine whether a manipulation is safe, agents have to figure out how they think that the other voters in the coalition will vote. Thus, she reasons on a higher-order level about the voting behaviour of other members of her coalition.

It is likely that sophisticated voters not only reason about the reasoning of like-minded voters, but also about the reasoning of voters that have conflicting interests. We generalised the idea of reasoning about the reasoning of other voters based on the principles of the cognitive hierarchy theory. One can distinguish different strategic types of players. A strategic type captures the level of strategic sophistication of a player and corresponds to the number of steps that the agent will compute in a sequence of iterated best strategies. We developed an epistemic model that represents the $k$ steps of higher-order reasoning of a level-$k$ reasoning voter. We showed that any non-dictatorial and surjective social choice function $F$ for three or more alternatives is susceptible to manipulation under at least some levels higher-order reasoning: even if $F$ is immune to manipulation under level-$k$ reasoning, it will be manipulable by voters who reason at level $k+1$.

In an iterative voting procedure, voters are allowed to change their ballots as often as they want. Higher-order reasoning voters know this and will try to predict the iterative voting process by reasoning about possible future manipulations of fellow voters. We focused on second-order reasoning voters, that is, voters who are able to look one step ahead. Because we assumed that voters are risk-averse, it is reasonable to assume that a voter will refrain from manipulating the election if she knows that a second manipulator could parasitise her manipulation: in that case, the second manipulator casts a strategic vote that results in an outcome that is worse than if she had not strategised at all. We showed that even if a voter is able to predict a future parasitising manipulation, under positional scoring rules and (if the number of voters is even) under non-antagonistic and strongly Condorcet-consistent voting rules, she is still able to cast a second-order risk-free strategic vote.

In this thesis, we regarded higher-order reasoning in one-shot voting and iterative voting as two completely different mental processes. Nonetheless, they may be more closely related than it seems. We will illustrate this with an example. Suppose that voter $i$ has a dominant manipulation of the truthful profile. If a voter is a level-3 reasoner, she thinks that every other voter is a level-2 reasoner, so she thinks that every other agent thinks that she is a level-1 reasoner. In that case, the other agents will realise that she has a strategic vote and apply their best strategy given her strategic vote. Now, suppose that there exists an agent $j$ with a parasitising manipulation: he has a best strategy that results in a winner that is worse than the winner under the truthful profile, from the perspective of $i$. Since $i$ is a level-3 reasoner, she realises that in that case, it is better to not cast the strategic ballot. So, second-order reasoning in iterative voting is in some way similar to level-3 reasoning in one-shot voting. This observation suggests that manipulation under higher-order reasoning in one-shot voting and manipulation under higher-order reasoning in iterative voting are more connected than it seemed, and it might be fruitful to explore which results can be translated from the one-shot voting setting to the iterative voting setting and vice versa.

Finally, we discussed a basic notion of strategic communication in which voters are able to share factual information. We discussed under which conditions a voter has an incentive to share factual information with other voters. We assumed that a voter only wants to reveal certain information is she considers it possible that this will stimulate another voter to vote strategically, in such a way that it is beneficial for the communicator. We showed that in many widely studied voting situations, no voter has an incentive to communicate. However, these results are not very intuitive, because in some cases it can definitely be beneficial for a group of voters to try to form a coalition and coordinate ballots. We realised that the conditions we require in our definition of an incentive to communicate are too strong to explain communication between agents in general. In order to get a better understanding of communication in a voting setting, we need to study more normative models of communication. This requires a model with richer ballot dynamics. Furthermore, we may have to weaken our assumption that voters are completely risk-averse. Models of social networks in which socially connected agents influence each other's opinions and behaviour should also be explored in more detail, in particular their connection with social choice theory.

## 8.2 Directions for future research

This thesis is rich in potential future work. Some possible directions are discussed in the concluding subsections of the chapters, more general ideas are considered here. First, we should mention that we have not used the full potential of this model yet. The model allows us to investigate

more complex forms of partial knowledge, such as scenarios in which not all voters have the same type of information, and where agents are uncertain about each other's uncertainty. Future research could shed light on how higher-order reasoning voter behave in scenarios in which they do not know what the others know. Another direction would be to focus on more concrete social choice functions and analyse under which conditions (classes of) social choice functions are susceptible to manipulation under higher-order reasoning and partial information.

It might also be interesting to delve deeper into manipulation of iterative voting procedures by agents that are able to predict future steps. In particular, if a voter is able to analyse the outcome of a convergent iterative voting process, that voter might adapt her manipulative behaviour in order to affect the final outcome rather than the outcome in the next round. In future work, the implications for her strategic incentives could be investigated.

To model more dynamic phenomena, it might be useful to look at richer dynamic epistemic structures that are able to deal with belief revision. Plausibility models (Baltag and Smets, 2008) are used to represent more nuanced versions of knowledge and belief. In a plausibility model, every agent composes a plausibility order over the set of states in the model, indicating how plausible the agent thinks a certain world is relative to another world. New information only increases or decreases the plausibility of a world, which is a less rigorous update procedure than the update procedure of the model presented in this thesis. Plausibility models could be useful to model more dynamics, such as a voter who changes the level of her reasoning or a voter who strategically spreads a lie.

We established that in many cases, higher-order reasoning agents are able to manipulate the election. However, it can computationally be very hard to do so. Future research could investigate the computational complexity of manipulation under higher-order reasoning.

# Appendix

**Proof of Proposition 3.16**:
Let $M$ be an epistemic model for strategic voting with actual world $s^*$. Let $U_{b_i^G}$ and $U_{b_j^H}^H$ a $b_i^G$-manipulation update model and a $b_j^H$-manipulation update model for $M$ respectively. We show that the $U_{b_i^G}; U_{b_j^H}$ is isomorphic to $U_{b_j^G}; U_{b_i^H}$. Then, by Proposition 3.8, it follows that $M, s \vDash [b_i^G][b_j^H]\varphi \iff M, s \vDash [b_j^H][b_i^G]\varphi$. The non-trivial part is to show that the postconditions coincide. Let $post_1$ denote the postconditions of $U_{b_i^G}$, $post_2$ the postconditions of $U_{b_j^H}$, $post_{12}$ the postconditions of $U_{b_i^G}; U_{b_j^H}$ and $post_{21}$ the postconditions of $U_{b_j^H}; U_{b_i^G}$. Consider $(e_1, e_1') \in U_{b_i^G}; U_{b_j^H}$. For any $x, y \in X$ and $k \neq i, j$, we have

$$post_{12}(e_1, e_1')(x \succ_k^b y) = x \succ_k^b y = post_{21}(e_1', e_1).$$

For $i$, we have that $post_1(e_1)(x \succ_i^b y) = \top$ or $post_2(e_1)(x \succ_i^b y) = \bot$, and $post_2(e_1)(x \succ_i^b y) = x \succ_i^b y$ for all $x, y \in X$. Hence,

$$post_{12}(e_1, e_1')(x \succ_i^b y) = post_1(e_1)(x \succ_i^b y) = post_{21}(e_1', e_1)(x \succ_i^b y)$$

For $j$, in the same way we can show that for every $x, y \in X$,

$$post_{12}(e_1, e_1')(x \succ_j^b y) = post_2(e_1)(x \succ_j^b y) = post_{21}(e_1', e_1)(x \succ_j^b y)$$

For the other states $(e_1, e_2')$, $(e_2, e_1')$ and $(e_2, e_2')$ we can show in a similar way that their postconditions correspond with the postconditions of states $(e_2', e_1)$, $(e_1', e_2)$ and $(e_2', e_2)$ respectively.

**Proof of Proposition 5.7**:
As a notational convention, we will write $U$ for $U_{b_i^i}$ in this proof. For $j \neq i$, we have $S^{0,j} := \{s_x^j \mid 1 \leq x \leq h\}$. For $i$, let $S^{0,i} := \{s \mid s^* R_i^0 s\}$. Then $S^{1,i,s^*} = \bigcup_{j \in N} S^{0,j}$. Recall that $S^{M^0 \otimes U} = \{(s_x, e_1) \mid s_x \in S^0\} \cup \{(s_x, e_2) \mid s_x \in S^0\}$. We define the following relation $\mathcal{R} \subseteq S^{1,i,s^*} \times S^{M^0 \otimes U}$:

$$\mathcal{R} := \{(s_x, (s_x, e_1)) \mid s_x \in S^{0,i}\} \cup \{(s_x^j, (s_x, e_2)) \mid s_x \in S^0, j \neq i\}$$

We show that $\mathcal{R}$ is a bisimulation. First, we show that if $(s, (s, e)) \in \mathcal{R}$, then $V^{1,i,s^*}(s) = V^{M^0 \otimes U}(s, e)$. For $s_x \in S^{0,i}$:

$$\begin{aligned} V^{1,i,s^*}(s_x) &= (\mathsf{proj}_1(V(s_x)), (\mathcal{S}_i(M^0, s^*), \mathsf{proj}_2(V(s_x))(-i))) \\ &= (\mathsf{proj}_1(V(s_x)), (b_i, \mathsf{proj}_2(V(s_x))(-i))) \\ &= V^{M^0 \otimes U}(s_x, e_1). \end{aligned}$$

For $s_x^j \in S^{0,j}$:

$$V(s_x^j) = V^0(s_x) = V(s_x, e_2).$$

Next, we show that the forth condition holds. First, we consider $R^{1,i,s^*}(i)$. For $s_x, s_y \in S^{0,i}$, we have that $s_x R^0 s_y$ by definition of $S^{0,i}$. Furthermore, since $e_1 R_i^U e_1$, we have $(s_x, e_1) R_i^{M\otimes U}(s_y, e_1)$. If $(s_x^j, s_y^j) \in R^{1,i,s^*}(i)$, it must hold that $s_x R_i^0 s_y$. We have $(s_x^j, (s_x, e_2)) \in \mathcal{R}$ and $(s_y^j, (s_y, e_2)) \in \mathcal{R}$. Since $(e_2, e_2) \in R^U(i)$, it follows that $(s_x, e_2) R_i^{M^0\otimes U}(s_y, e_2)$.

For $R^{1,i,s^*}(j)$, we have two options again: for $(s_x, s_y^j) \in R^{1,i,s^*}(j)$, it holds that $(s_x, (s_x, e_1)) \in \mathcal{R}$ and $(s_y^j, (s_y, e_2)) \in \mathcal{R}$. We have $(s_x, s_y) \in R^0(j)$ by definition of $R^{1,i,s^*}(j)$, and furthermore $(e_1, e_2) \in R^U(j)$. We conclude that $(s_x, e_1) R_i^{M\otimes U}(s_y, e_2)$. The second case is when for some voter $l \neq i$, $(s_x^l, s_y^l) \in R^{1,i,s^*}(j)$. In this case, it holds that $s_x R_j^0 s_y$. We have $(s_x^l, (s_x, e_2)) \in \mathcal{R}$ and $(s_y^l, (s_y, e_2)) \in \mathcal{R}$. Since $(e_2, e_2) \in R^U(j)$, it follows that $(s_x, e_2) R_j^{M\otimes U}(s_y, e_2)$.

It is left to show that the back condition holds. For $R^{M\otimes U}(i)$, since $(e_1, e_2), (e_2, e_1) \notin R^U(i)$, we either have a relation of the form $((s_x, e_1), (s_y, e_1))$ for $1 \leq x, y \leq h$ or a relation of the form $((s_x, e_2), (s, y, e_2))$ for $1 \leq x, y \leq h$. In the first case, it must hold that $s_x R_i^0 s_y$, and $(s_x, (s_x, e_1)) \in \mathcal{R}$ and $(s_y, (s_y, e_1)) \in \mathcal{R}$ for $s_x, s_y \in S^{0,i}$. By definition of $R^{1,i,s^*}(i)$, it follows that $(s_x, s_y) \in R^{1,i,s^*}(i)$. In the second case, it must hold that $s_x R_i^0 s_y$, and that $(s_x^j, (s_x, e_2)) \in \mathcal{R}$ and $(s_y^j, (s_y, e_2)) \in \mathcal{R}$ for $s_x, s_y \in S^{0,i}$ and some $j \in N$. Again, it follows that $(s_x, s_y) \in R^{1,i,s^*}(i)$ by definition.

Now take $j \neq i$. For $R^{M\otimes U}(j)$, since $(e_1, e_1), (e_2, e_1) \notin R^U(j)$, there are two cases: a relation of the form $((s_x, e_1), (s_y, e_2))$ or a relation of the form $((s_x, e_2), (s, y, e_2))$ for $1 \leq x, y \leq h$. In the first case, it must hold that $s_x R_j^0 s_y$. We have $(s_x, (s_x, e_1)) \in \mathcal{R}$ and for all voters $l \neq i$, $(s_y^l, (s_y, e_2))$, so in particular $(s_y^j, (s_y, e_2)) \in \mathcal{R}$. By definition, $(s_x, s_y^j) \in R^{1,i,s^*}(j)$. In the second case, it must hold that $s_x R_j^0 s_y$ and for all voters $l \neq i$, $(s_x^l, (s_x, e_2)) \in \mathcal{R}$ and $(s_y^l, (s_y, e_2)) \in \mathcal{R}$. Since $s_x R_j^0 s_y$, it follows that for all voters $l \neq i$, we have $(s_x^l, s_y^l) \in R^{1,i,s^*}$.

**Proof of Proposition 5.11**:
Let $M^0$ be a level-0 model with actual state $s^*$. For $k = 1$, take $i \in N$ and let $G = N$ or $G = \{i\}$. We consider a model in which $i$ is a level-1-reasoner: $M^{1,G,s^*}$. Take $s \in S^{1,G,s^*}$ with $s^{1,G,s^*} R_i^{1,G,s^*} s$. Then $s \in S^{0,i}$, and hence

$$
\begin{aligned}
\boldsymbol{b}_s(i) &= \mathsf{proj}_2(V^{1,G,s^*}(s))(i) \\
&= \big( \bigtimes_{j<i} \mathsf{proj}_2(V^0(s_x))(j) \times \mathcal{S}_i(M^0, s^*) \times \bigtimes_{j>i} \mathsf{proj}_2(V^0(s_x))(j) \big)(i) \\
&= \mathcal{S}_i(M^0, s^*) \\
&= \mathcal{S}_i(\{\boldsymbol{b}(-i) \mid \text{there exists } s' \in S^0 \text{ with } s^* R_i s' \text{ and } M^0, s' \vDash \boldsymbol{b}\}) \\
&= \mathcal{S}_i(\{\boldsymbol{p}(-i) \mid \text{there exists } s' \in S^0 \text{ with } s^* R_i s' \text{ and } M^0, s' \vDash \boldsymbol{p}\}) \text{ (by definition of } M^0) \\
&= \mathcal{S}_i(\{\boldsymbol{p}(-i) \mid \text{there exists } t \in S^{1,G,s^*} \text{ with } s^{1,G,s^*} R_i^{1,G,s^*} t \text{ and } M^{1,G,s^*}, t \vDash \boldsymbol{b}\}) \\
&\quad \text{(since the preferences that } i \text{ considers possible do not change)} \\
&= \mathcal{S}_i(\{\boldsymbol{b}(-i) \mid \text{there exists } t \in S^{1,G,s^*} \text{ with } s^{1,G,s^*} R_i^{1,G,s^*} t \text{ and } M^{1,G,s^*}, t \vDash \boldsymbol{b}\}) \text{ ($i$ is a level-1 reasoner)} \\
&= \mathcal{S}_i(M^{1,G,s^*})
\end{aligned}
$$

For $k > 1$, let $i \in N$ and consider $M^{k,G,s^*}$ with $G = \{i\}$ or $G = N$ and actual state $s^{k,G,s^*}$. Let

$s \in S^{k,G,s^*}$ such that $s^{k,G,s^*} R_i^{k,G,s^*} s$. Then we have:

$$\begin{aligned}
\boldsymbol{b}_s(i) &= \mathsf{proj}_2(V^{k,G,s^*}(s))(i) \\
&= \big(\bigtimes_{j<i} \mathcal{S}_j(M^{k-1,j,t}, s^{k-1,j,t}) \times \mathcal{S}_i\big(\bigcup_{t' \in S^{0,i}} \bigtimes_{j \neq i} \mathcal{S}_j(M^{k-1,j,t'}, t'), \boldsymbol{p}_{s^*}(i)\big) \times \bigtimes_{j>i} \mathcal{S}_j(M^{k-1,j,t}, s^{k-1,j,t})\big)(i) \\
&= \mathcal{S}_i\big(\bigcup_{t' \in S^{0,i}} \bigtimes_{j \neq i} \mathcal{S}_j(M^{k-1,j,t'}, t'), \boldsymbol{p}_{s^*}(i)\big) \\
&= \mathcal{S}_i\big(\bigcup_{t' \in S^{0,i}} \mathsf{proj}_2(V^{k,G,s^*}(t'))(-i), \boldsymbol{p}_{s^*}(i)\big) \\
&= \mathcal{S}_i(\{\mathsf{proj}_2(V^{k,G,s^*}(t'))(-i) \mid t' \in S^{0,i}\}, \boldsymbol{p}_{s^*}(i)) \\
&= \mathcal{S}_i(\{\mathsf{proj}_2(V^{k,G,s^*}(t'))(-i) \mid t' \in S^{k,G,s^*} \text{ with } s^{k,G,s^*} R_i^{k,G,s^*} t'\}, \boldsymbol{p}_{s^*}(i)) \\
&= \mathcal{S}_i(M^{k,G,s^*}, s^{k,G,s^*}).
\end{aligned}$$

So, we have that $\mathcal{S}_i(M^{k,G,s^*}, s^{k,G,s^*}) = \boldsymbol{b}_s(i)$, and hence it holds that $i$ has an incentive if and only if $\mathcal{S}_i(M^{k,G,s^*}, s^{k,G,s^*}) \neq \boldsymbol{p}_{s^*}(i)$.

**Proof of Theorem 5.16**:
Let $F$ be a social choice function and let $M^0$ be a level-0 model with actual state $s^*$ such that voter $i$ has a first-order incentive to manipulate. This means that

$$\mathcal{S}_i(M^{1,i,s^*}, s^{1,i,s^*}) \neq \boldsymbol{p}_{s^*}(i).$$

Now let $k \in \mathbb{N}$ and suppose that $F$ is immune to manipulation under level-$k$ reasoning. This means that in particular, for any $s \in S^0$ and any $j \in N$, the $j$-level-$k$ model $M^{k,j,s}$ based on $M^0$ and with actual world $s^{k,j,s}$, is immune to manipulation. In other words, for all $j \in N$, and all $s \in S^0$,

$$\mathcal{S}_j(M^{k,j,s}, s^{k,j,s}) = \boldsymbol{p}_s(j).$$

Now consider $M^{k+1,i,s^*}$ with actual world $s^{k+1,i,s^*}$. We show that $i$ has a $(k+1)^{\text{th}}$-order incentive to manipulate by proving that

$$\mathcal{S}_i(M^{k+1,i,s^*}, s^{k+1,i,s^*}) \neq \boldsymbol{p}_{s^*}(i).$$

We have

$$\begin{aligned}
\mathcal{S}_i(M^{k+1,i,s^*}) &= \mathcal{S}_i(\{(\mathsf{proj}_2(V^{k+1,i,s^*}(t))(-i) \mid t \in S^{k+1,i,s^*} \text{ with } s^{k+1,i,s^*} R^{k+1,i,s^*} t\} \\
&= \mathcal{S}_i\big(\{\bigtimes_{j \neq i} \mathcal{S}_j(M^{k,j,t}, s^{k,j,t} \mid t \in S^{k+1,i,s^*} \text{ with } s^{k+1,i,s^*} R^{k+1,i,s^*} t\}, \boldsymbol{p}_{s^*}(i)\big) \\
&= \mathcal{S}_i\big(\{\bigtimes_{j \neq i} \mathcal{S}_j(M^{k,j,t}, s^{k,j,t} \mid t \in S^{0,i}\}, \boldsymbol{p}_{s^*}(i)\big) \\
&= \mathcal{S}_i\big(\{\bigtimes_{j \neq i} \boldsymbol{p}_t(j) \mid t \in S^{0,i}\}, \boldsymbol{p}_{s^*}(i)\big) \\
&\quad \text{(since } F \text{ is strategyproof under level-}k \text{ reasoning)} \\
&= \mathcal{S}_i\big(\{\boldsymbol{p}_t(-i) \mid t \in S^{0,i}\}, \boldsymbol{p}_{s^*}(i)\big) \\
&= \mathcal{S}_i\big(M^{1,i,s^*}, s^{1,i,s^*}\big) \\
&\neq \boldsymbol{p}_{s^*}(i) \text{ (since } i \text{ has a first-order incentive to manipulate).}
\end{aligned}$$

We conclude that $i$ has a $(k+1)^{\text{th}}$-order incentive to manipulate, and hence that $F$ is susceptible to manipulation under level-$k+1$ reasoning.

# Bibliography


Ayala Arad and Ariel Rubinstein. The 11–20 money request game: A level-k reasoning study. *American Economic Review*, 102(7):3561–3573, 2012.

Zeinab Bakhtiari, Hans van Ditmarsch, and Abdallah Saffidine. How does uncertainty about other voters determine a strategic vote? *Journal of Applied Non-Classical Logics*, To appear, 2018.

Alexandru Baltag and Sonja Smets. A qualitative theory of dynamic interactive belief revision. In Giacomo Bonanno, Wiebe van der Hoek, and Michael Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory (LOFT 7). Texts in Logic and Games 3*, pages 11–58. Amsterdam University Press, 2008.

Alexandru Baltag, Lawrence S. Moss, and Slawomir Solecki. The logic of public announcements, common knowledge, and private suspicions. Technical report, Centrum voor Wiskunde en Informatica, Amsterdam, 1999.

Alexandru Baltag, Zoé Christoff, Rasmus K. Rendsvig, and Sonja Smets. Dynamic epistemic logics of diffusion and prediction in social networks. *Studia Logica*, pages 1–43, 2018.

Heinrich Anton de Bary. Die Erscheinung der Symbiose. *Verlag von Karl J. Trübner, Strassburg*, 1879.

Jean-Pierre Benoît. The Gibbard–Satterthwaite theorem: a simple proof. *Economics Letters*, 69: 319–322, 2000.

Johan van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2011.

Johan van Benthem and Ferong Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics*, 17(2):157–182, 2007.

Johan van Benthem, Jan van Eijck, and Barteld Kooi. Logics of communication and change. *Information and Computation*, 204:1620–1662, 2006.

Vittoria Bilò and Michele Flammini. Extending the notion of rationality of selfish agents: Second order nash equilibria. *Theoretical Computer Science*, 412:2296–2311, 2011.

Steven J. Brams and Donald Wittman. Nonmyopic equilibria in 2x2 games. *Conflict Management and Peace Science*, 6(1):39–62, 1981.

Simina Brânzei, Ioannis Caragiannis, Jamie Morgenstern, and Ariel Procaccia. How bad is selfish voting? In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-13)*, 2013.

Colin F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

Colin F. Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

Samir Chopra, Eric Pacuit, and Rohit Parikh. Knowledge-theoretic properties of strategic voting. In Jóse Júlio Alferes and João Leite, editors, *Logics in Artificial Intelligence*, pages 18–30. Springer Berlin Heidelberg, 2004.

Giovanni Ciná and Ulle Endriss. Proving classical theorems of social choice theory in modal logic. *Autonomous Agents and Multi-Agent Systems*, 30(5):963–989, 2016.

Vincent Conitzer and Toby Walsh. Barriers to manipulation in voting. In *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

Vincent Conitzer, Toby Walsh, and Lirong Xia. Dominating manipulations in voting with partial information. In *Proceedings of the 25th National Conference on Artificial Intelligence (AAAI-2011)*, page 787–791, 2011.

Miguel Costa-Gomes, Vincent P. Crawford, and Bruno Broseta. Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5):1193–1235, 2001.

Hans van Ditmarsch and Barteld Kooi. Semantic results for ontic and epistemic change. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the foundations of game and decision theory (LOFT 7)*, Texts in Logic and Games 3, pages 87–117. University of Amsterdam, 2006.

Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library*. Springer, 2007.

Hans van Ditmarsch, Jérôme Lang, and Abdallah Saffidine. Strategic voting and the logic of knowledge. In *Proceedings of 14th TARK – Chennai*, 2013.

Jan van Eijck. PDL as a Multi-Agent Strategy Logic. *TARK 2013, Chennai, India*, 2013.

Joseph Farrell. Communication, coordination and nash equilibrium. *Economics Letters*, 27: 209–214, 1988.

Joseph Farrell. Meaning and credibility in cheap-talk games. *Games and Economic Behavior*, 5: 514–531, 1993.

Michael Franke and Robert van Rooij. Strategies of persuasion, manipulation and propaganda: Psychological and social aspects. In Johan van Benthem, Sujata Ghosh, and Rineke Verbrugge, editors, *Models of Strategic Reasoning*. Springer-Verlag Berlin Heidelberg, 2015.

Allan Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601, 1973.

Joseph Y. Halpern. Reasoning about knowledge: An overview. In Joseph Y. Halpern, editor, *Reasoning about knowledge*, pages 1–18. Morgan Kaufmann, 1986.

Trey Hedden and Jun Zhang. What do you think i think you think?: Strategic reasoning in matrix games. *Cognition*, 85:1–36, 2002.

D. Marc Kilgour. Equilibria for far-sighted players. *Theory and Decision*, 16:135–157, 1984.

Omer Lev and Jeffrey S. Rosenschein. Convergence of iterative voting. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2012)*, 2012.

Omer Lev and Jeffrey S. Rosenschein. Convergence of iterative scoring rules. *Journal of Artificial Intelligence Research*, 57:573–591, 2016.

Reshef Meir. Plurality voting under uncertainty. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 2015.

Reshef Meir, Maria Polukarov, Jeffrey S. Rosenschein, and Nicholas R. Jennings. Convergence to equilibria in plurality voting. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, 2010.

Reshef Meir, Omer Lev, and Jeffrey S. Rosenschein. A local-dominance theory of voting equilibria. In *Proceedings of the fifteenth ACM conference on Economics and Computation*, pages 313–330, 2014.

Hervé Moulin. On strategy-proofness and single peakedness. *Public Choice*, 35:437–455, 1980.

Rosemarie Nagel. Unraveling in guessing games: An experimental study. *The American Economic Review*, 85(5):1313–1326, 1995.

Prashant Parikh. Communication and strategic inference. *Linguistics and Philosophy*, 14:473–514, 1991.

Andrés Perea. *Epistemic Game Theory.* Cambridge University Press, 2012.

Tin Perkov. Natural deduction for modal logic of judgment aggregation. *Journal of Logic, Language and Information*, 25(3-4):335–354, 2016.

Matthew Rabin. Communication between rational agents. *Journal of Economic Theory*, 5: 144–170, 1990.

Annemieke Reijngoud. Voter response to iterated poll information. Master's thesis, Institute for Logic, Language and Computation, University of Amsterdam, 2011.

Annemieke Reijngoud and Ulle Endriss. Voter response to iterated poll information. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2012)*, 2012.

Reyhaneh Reyhani and Mark C. Wilson. Best-reply dynamics for scoring rules. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI-2012)*, 2012.

Mark Satterthwaite. 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, 1975.

Arkadi Slinko and Shaun White. Is it ever safe to vote strategically? *Social Choice and Welfare*, 43:403–427, 2014.

Sonja Smets and Fernando R. Velázquez-Quesada. A logical perspective on social group creation. In Pavel Arazim and Tomáš Lávička, editors, *The Logica Yearbook 2017*, pages 271–288. College Publications, London, UK, 2018.

Dale O. Stahl. Evolution of smart$_n$ players. *Games and Economic Behaviour*, 5:604–617, 1993.

Dale O. Stahl and Paul W. Wilson. Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, 25(3):309–327, 1994.

Dale O. Stahl and Paul W. Wilson. Players' models of other players. *Games and Economic Behaviour*, 10:218–254, 1995.

Zoi Terzopoulou. Manipulating the manipulators: Richer models of strategic behavior in judgment aggregation. Master's thesis, Institute for Logic, Language and Computation, University of Amsterdam, 2017.

Nicolas Troquard, Wiebe van der Hoek, and Michael Wooldridge. Reasoning about social choice functions. *Journal of Philosophical Logic*, 40(4):473–498, 2011.