Democratic Decision Making

—

A Theoretical Analysis

Dissertation

zur Erlangung des akademischen Grades
eines Doktors der Sozial- und Wirtschaftswissenschaften (Dr. rer. pol.)

an der Fakultät Sozial- und Wirtschaftswissenschaften der Otto-Friedrich-Universität Bamberg

vorgelegt von
Simon Scheller, M.Sc, M.A,
geboren am 23. November 1988 in Bayreuth

Bamberg, 6. Dezember 2017

URN: urn:nbn:de:bvb:473-opus4-508144
DOI: https://doi.org/10.20378/irbo-50814

1. Gutachter:	Prof. Dr. Johannes Marx Otto-Friedrich-Universität Bamberg
2. Gutachter:	Prof. Dr. Thomas Rixen Otto-Friedrich-Universität Bamberg
3. Gutachter:	Prof. Dr. Kai Fischbach Otto-Friedrich-Universität Bamberg

Weiteres Mitglied der Promotionskommission:

Prof. Dr. Florian Herold
Otto-Friedrich-Universität Bamberg

Eingereicht am: 13. Oktober 2017

Table of Contents

Framework Paper	*Democratic Decision Making – A Theoretical Analysis*
Paper 1	*Mitigating the Problem of Manipulation in the 'Adjusted Winner' Procedure*
Paper 2	*Rationally Poor? – What the Emergence of Inequality can Teach us About Rational Behaviour*
Paper 3	*When do Groups get it right? – On the Epistemic Performance of Voting and Deliberation*
Paper 4	*Fear Appeals as a Political Strategy – A Theoretical Exploration*

Framework paper

Democratic Decision Making

–

A Theoretical Analysis

This is an unpublished paper that provides the thesis with an overall theoretical context.

The suggested citation is:

Scheller, Simon (2017) *Democratic Decision Making – A Theoretical Analysis.* Doctoral Dissertation. Bamberg: University of Bamberg (mimeo).

Democratic Decision Making – A Theoretical Analysis

Simon Scheller

October 11, 2017

1 Introduction

Group decisions constitute a central element of social interactions. Committees need to decide which candidate to hire, parliaments must choose laws, and groups of friends may want to pick a restaurant for a joint dinner. For such decisions, it is often claimed that they should be made democratically. My PhD-thesis revolves around the question of how such claims can be justified, and what mechanisms should govern democratic decision making [DDM] in different situations. Due to the ubiquity and importance of group decisions, the philosophical analysis of DDM deserves central attention.

This chapter discusses the philosophical foundations of these questions in order to frame and inform the analysis of specific democratic mechanisms in the individual papers. For that purpose, I first analyse different fundamental approaches to justifying DDM in section 2. On the one side I distinguish between the *market-* and the *forum* view on DDM, which is based on a distinction between preferential and epistemic group disagreements. On the other side, justifications of DDM adhere to either its *instrumental* or *final* value, and whether this value is *intrinsic* to a democratic mechanism or also depends on *extrinsic* factors. I then discuss central issues and problems of DDM mechanisms and their implications for different justificatory approaches – both for the market- and the forum-view (sections 3 and 4). I further provide a taxonomy of different group decision problems and procedures, which constitutes a reference point for the individual papers that make up this thesis (section 5). Methodologically, the individual papers revert to formal modelling techniques to address the aforementioned questions. In section 6, I argue for the appropriateness of agent-based and game-theoretic models for the study of democratic procedures. Their greatest merits in this

context are (a) that they allow for a general analysis of mechanisms regardless of the availability of empirical data, (b) the possibility to connect micro foundations to macro-level behaviour by explaining aggregate outcomes on the basis of individual behaviour, and (c) their suitability for complex social processes such as DDM. The four individual papers scrutinize and evaluate specific elements of DDM procedures in light of that framework. I give a short description of each paper in section 7, show how it fits into the overall scheme of this thesis, and describe its major contributions.

The central contribution of my thesis consists in an improved theoretical understanding of specific DDM procedures. Each individual paper sheds light on vice and virtues of certain procedures, how they could be improved or should ideally be designed, and what can be done to prevent them from being manipulated or from failure more generally. In doing so, this thesis also contributes to a better fundamental understanding of the value and properties of DDM. Finally, on a methodological level, it illustrates how agent-based models (and formal models more generally) can be fruitfully employed in social science and philosophy alike.

2 DDM: Primary Questions

2.1 Definition and Scope

A necessary condition to call a group decision process democratic is that it gives each group member the possibility of being involved in the decision making process.[1] DDM stands in contrast to hierarchical decision making, where one or a few members decide about issues concering the whole group without involving all group members. Involvement must be substantial in the sense that it goes beyond passive observation or the possibility of voicing concerns: Each individual must have an equal influence on the decision. How exactly an individual's impact should be conceptualised varies across situations and procedures, and is one of the central points of discussion throughout this entire project.

As this definition implies, this work's subject is not Democracy as a state system, but rather DDM as a procedural practice in various areas of life. Democracy as a state system is a prominent topic in political philosophy. For a thorough conceptual analysis, see Lauth (2004). Dahl (1989) discusses the fundamentals of Democratic theory and practice alike, and outlines conditions for the success of democratic systems in the world. Other authors

[1] Not all group members need to actually make use of that possibility for a procedure to be democratic.

have addressed more specific issues, such as the relation between Democracy and human rights (Beetham, 1999; De Mesquita et al., 2005), its impact on economic prosperity (Przeworski, 2004; Helliwell, 1994), inequality (Sirowy and Inkeles, 1990), peace (Gleditsch, 1992; Ray, 1998), as well as the balance of power between different democratic institutions and other constitutional questions more generally (Bellamy, 2007; Schweller, 2010). In this thesis, I disregard these issues. Instead, I focus on Democracy as a principle for making group decisions more generally. This is not to say that Democracy can be reduced to DDM mechanisms alone – after all, majority rule does not suffice to make a state democratic. Still, the results of my thesis carry strong implications also for democratic systems, as DDM constitutes a vital and necessary part of all democratic systems. In order to avoid confusion for the subsequent discussion, I also assume that the group decisions to be discussed are all situated within well-established Democratic systems. For example, this means that some issues are protected by individual rights and hence cannot be subject of group decisions to start with. Discussing democratic decision making under non-democratic circumstances would be problematic at least, potentially self-contradictory in parts, and certainly would require a contextual analysis to an extent which is neither possible nor intended here.

Finally, while one can try to infer predictions about larger classes of mechanisms, it is impossible to provide a general assessment of group decision making mechanisms as a whole. In the individual papers, I therefore inquire into the characteristics of specific democratic mechanisms, identifying their vices and virtues. When these insights are used for a comprative analysis, one may find that democratic mechanisms come with disadvantages of which non-democratic procedures are unaffected. These results can then be used to inform discussions about optimal institutional design. Yet, it is important to keep in mind that there may be other criteria apart from a mechanism's democraticness which are not addressed in this work. Keeping these considerations in mind, I now turn to the central arguments about the value of democratic procedures.

2.2 Four types of Value

The main goal of this chapter is to present and discuss different justifications of democratic decision making. Asking about the goodness of DDM, however, first requires a discussion about 'goodness' itself. For such a clarification, Koorsgard's distinctions prove useful (Korsgaard, 1983). Koorsgard identifies two decisive questions for categorising normative evaluations. First, one must ask whether a thing is valuable in itself, or whether it is merely an instrument to achieve another goal. This invites the distinction between *final* value and

instrumental value. As Anderson (2009) argues, if something has final value, "we'd still prefer to engage in it, even if the same consequences could be brought about by other (passive) means" (Anderson, 2009, p. 225). If, on the other hand, something has solely instrumental value, one would abandon a procedure if its result could be brought about more easily in other ways.

Second, one must ask whether a thing's intrinsic properties are sufficient for its value, or whether it is dependent on external conditions to realise its value. This constitutes the dychotomy between intrinsic and extrinsic values, which refers to the value's location or source (Korsgaard, 1983, p. 170). If X is intrinsically valuable, X's sole existence is enough for the value to be realised, while if the value of X is extrinsic, other conditions need to be fulfilled for X's value to be realised.[2]

In combination, these two dichotomies give rise to four categories, which I illustrate by prime examples (see also table 1): Money classically has instrumental and extrinsic value, since we value it as a means to buy other things (instrumental) and it is only valuble if others in society share its meaning of value (extrinsic). A classic example for a finally and intrinsically valuable good is physical well-being, because it is valuable *in* itself (final value) and *by* itself (intrinsic value). Third, a prime example for something with intrinsic, instrumental value is food. It is valuable as a means for survival, and its value depends solely on its intrinsic properties (i.e. nutritiousness). Lastly, something that has a final and extrinsic value could be to hold a position of power. Power may constitute someone's final goal, and is hence an end in itself. Yet, holding a position of power is naturally relational to outside conditions, namely that others are subject to the person's power. Power's value is therefore extrinsic.

Table 1: Four kinds of value: prime examples

	Final	**Instrumental**
Intrinsic	Well-being	Food
Extrinsic	Power	Money

[2]There is an extensive philosophical debate about the correct understanding of both dychotomies, whether the two dimensions naturally coincide and if they can be even distinguished in the first place. For an illuminating application and discussion of these categories to Democracy and Capitalism, see Marx and Waas (ming). For a more detailed and sophisticated discussion on the properties of value judgments generally, see Korsgaard (1983); Dorsey (2012); Rønnow-Rasmussen (2002) and the references therein.

2.3 Two types of Questions

While the two distinctions in the previous section refer to the nature of something's value more generally, justifications of DDM also fundamentally differ with regards to their central arguments.

In line with the economic tradition, some authors have argued that the goal of DDM is to aggregate individual preferences in a fair way. In that view, people's preferences are assumed to be exogenously given, they are matters of individual taste, and, accordingly, no judgment about the 'goodness' or 'correctness' of such preferences can reasonably be made. Matters of preference aggregation are characterized by competing interests between individuals, whereas everyone knows her personal valuation of different options. Elster (1986) labels this the *market-view* on democracy. Examples for typical market-view questions are 'How should a common surplus be distributed?', 'Should there be more unemployment benefits?' or 'Should we, as a group, watch a romantic movie or a thriller?'. For such market view problems, democratic mechanisms aim at reconciling competing individual interests. Exogenously given preferences are aggregated (usually by voting) in order to come to collective binding decisions.

Others attribute an epistemic role to DDM. This interpretation requires (or postulates) the existence of some objective truth, whereas individuals hold beliefs about what constitutes the best outcome for everyone. Epistemic problems are characterized by aligning individual interests, while the decision problem results from uncertainty about the evaluation of different alternatives. In Elster's terminology, this set of assumptions is called the *forum view*. Typical epistemic questions are: 'Is the defendant guilty or innocent of the accused crime?', 'Is climate change man-made or not?', or 'Which infrastructure project is most suitable for our city?'. Ideally, democratic mechanisms should reduce uncertainty about such questions and help the group to choose the correct, best, or optimal alternative. In this view, Democracy is also often associated with a deliberative component.

While the distinction between preferential and epistemic matters is always clearly drawn in theoretical approaches, they neither principally nor practically exclude one another. Various decision problems contain components of both. In some instances, they are even logically intertwined, as for example when factual assessments infuence individual moral judgments and vice versa. For matters of this work however, suffice it to recognize this complication and proceed under the assumption that problems with a clearly distinguishable (or separable) preferential or epistemic part exist, which renders this discussion useful and important.

The claim of the forum-view is also frequently phrased in terms of the

argument that deliberation leads to a change of individual preferences in light of the common good (Habermas, 1983; Elster, 1986). This interpretation, however, requires a detailed discussion of whether or not 'moral truths' exist or if such questions should rather be envisaged as preferential questions. The question 'Is abortion morally justified?' constitutes a prime example. In the view that I take here, this discussion eventually boils down to the question of how to categorize certain questions with regards to the market-forum distinction.[3]

With this basic characterisation of the market and the forum-view and the different types of values from the previous section at hand, the subsequent sections discuss different justifications of DDM in detail.

3 The Market View

3.1 Four types of Fairness

What is the value of aggregating preferences democratically? When it comes to discussing these issues, probably the most prominent terms in the literature are *fairness* and *equality* (Christiano, 2008; Rawls, 1957; Verba, 2006; Beitz, 1989; Cohen, 1986; Copp et al., 1993). For generality and simplicity, I subsume equality claims under the term fairness, since equality can be seen as a special kind of fairness requirement. There are also arguments directed at other properties of DDM. Some arguments – mainly those who talk about Democracy as a system more broadly – appeal, for example, to the resulting cooperative spirit, the protection of rights, legitimacy and acceptance of democratically made decisions as well as conformity with them. While these claims are also interesting in themselves, the focus lies on fairness as an appealing property of DDM.

The distinction between instrumental and final value of fairness in DDM translates into a distinction between valuing fair *outcomes* (instrumental) and valuing the fairness of *procedures* by themselves (final). In the instrumental interpretation, a fair mechanism is merely a tool to bring about fair outcomes. In contrast, when fairness is seen as finally valuable, the fairness of the mechanism itself is seen as valuable. The intrinsic-extrinsic-dimension of fairness refers to whether or not fairness is given by the mechanism's democratic characteristics itself, or whether other factors must be considered as well. Coleman describes an intrinsic notion as follows: "[W]hat justifies a decision making procedure is strictly a necessary property of the procedure

[3]For a further discussion of this issue, see Landwehr (2005); Elster (1998).

– one entailed by the definition of the procedure alone."(Coleman and Ferejohn, 1986, p. 7). Hence, we are left with four kinds of fairness that can be attributed to DDM.

- **Instrumental and intrinsic:** In this notion, outcomes are fair because they have been brought about by a fair procedure and this makes an outcome fair by definition. Such a view embodies a definitorial claim about DDM because democratic procedures define outcomes to be correct if and only if they had been brought about by a fair, democratic procedure.

- **Instrumental and extrinsic:** One might also reasonably claim that democratic procedures are suitable tools to bring about fair outcomes, even though these outcomes could also have been achieved by undemocratic mechanisms. Further their democraticness does not suffice to guarantee these results. Then, the value of a democratic mechanism would be external and instrumental. Such a position can, for instance, be ascribed to Arneson (2003), who argues that a democratic government is justified if and as long as it "produces better consequences for people than any feasible alternative mode of governance and [a democratic mode of government] merely happens to be a good tool in reaching a certain result."(Arneson, 2003, p. 122).

- **Final and intrinsic:** If one were to say that making a mechanism democratic is enough to make it fair, one is, similarly to the instrumental argument, left with a rather definitorial approach. Christiano, in part, advocates such a view of "equal considerations of interest" (Christiano, 1993, p. 43) in situations of irresolvable conflict. For decision making procedures, the principle 'one man, one vote' emodies this claim most prominently. At the same time, Christiano deliberately neglects a focus on outcomes (Christiano, 1993, p. 47): "Such a conception of equality requires not that everyone be equally well-off but that citizens have equal resources for advancing their interests.". For another example, see Griffin and his argument about equal distribution of political power (Griffin, 2003, p. 129).

- **Final and extrinsic:** One could reasonably dispute this claim by saying that it is not enough to give every individual the possibility to influence decisions, and instead argue that additional, more substantial conditions must be met. Such arguments usually invite a broader conception of democracy, as proposed for instance by Hyland (1995). In this argument DDM is characterised as extrinsically valuable.

In order to evaluate these positions with regards to their persuasiveness, several fundamental insights of Social Choice Theory are worth discussing.

3.2 The Challenge of Social Choice Theory

The domain of political theory that studies the aggregation of preferences through collective decision making mechanisms is Social Choice Theory [SCT]. Stirred by Condorcet's paradox (Condorcet, 1785) and Arrow's impossibility theorem (Arrow, 1963), large parts of the social choice literature discuss whether the aggregation of preferences can work in principle, and what this implies for the actual functioning of DDM mechanisms. This section gives an overview over SCT's main contributions, and what they imply for justifications of DDM.

An intuitive access to the problems of social choice is given by the 'paradox of voting', which is often ascribed to Condorcet (1785) and will be presented on the basis of Riker (1982, p.16ff). Consider a situation where three individuals need to choose between three alternatives. Assume that the individuals hold the following transitive, ordinal preference orderings over the set of alternatives $X = (x, y, z)$:

Table 2: Preference constellation 1

Person 1	$x \succ y \succ z$
Person 2	$y \succ z \succ x$
Person 3	$z \succ x \succ y$

Under pairwise majority voting (i.e. if each possible pair of alternatives is voted on individually), the group prefers x over y, y over z and z over x. However, these three pairwise preferences violate transitivity, which leads to a cyclic group preference ordering: $x \succ y \succ z \succ x$. Hence, pairwise majority voting, which is the most straight-forward way of aggregating individual preferences into group preferences, is unable to identify a clear winner. Unfortunately, this example is more than an odd outlier. In his impossibility theorem, Arrow (1962) famously showed that no voting rule can reliably fulfil certain seemingly uncontroversial basic conditions of fairness and rationality. While there is a wide debate about Arrow's conditions and the theorem's relevance, the result carries numerous practical implications for decision making

in the real world. Most famously, Riker (1982) illustrates how seemingly reasonable decision making mechanisms suffer from three severe problems: *instability*, *non-uniqueness* and *manipulability*.

The problem of *instability* has already been identified in the above paradox of voting: Some voting rules regularly fail to identify a clear winner (Coleman and Ferejohn, 1986, p. 11). This is problematic for DDM because failure to reach a decision can result to deadlock and disproportionally favours the status quo.

In reaction to the problem of instability, the requirements for decision rules can be lowered. However, this invites the problem of *non-uniqueness* (Coleman and Ferejohn, 1986, p. 11), which circumscribes situations where different voting rules lead to different results. For example, consider the following preference profiles:

Table 3: Preference constellation 2

Group 1 (35%)	$x \succ y \succ z$
Group 2 (40%)	$z \succ y \succ x$
Group 3 (25%)	$y \succ x \succ z$

Runoff-elections between the two alternatives with the most first-round votes leaves x as the winner. Under pairwise majority voting, alternative y wins. Under simple majority rule, the winner of the process would be alternative z. Neither of those three voting rules is clearly superior, and they all produce a different winner. Thus, the choice of decision making rule directly influences the outcome, and therefore renders it somewhat arbitrary. This is problematic for the status of decision rules, as one would think that outcomes should be independent of the mechanism employed and depend solely on the constellation of individual preferences.

Third and resulting from the previous problems, voting rules can be manipulated. Riker (1982) identifies two major kinds of manipulation schemes: agenda manipulation and strategic voting. An example for manipulation by strategic voting is already contained within the previous example from table 3: Under simple majority rule, group 3 could vote strategically for x instead of y in order to make x win over z, and thus at least realise its second best alternative. An agenda-manipulation example can be constructed from the preference profile in table 2. Assume a voting mechanims of pairwise elimination between two alternatives, and pitting the first-round winner against

the third alternative. If an individual were to hold the power to control the agenda (i.e. decide on the sequence of votes), she could make each option the winner by simply pitting the other two against each other in the first round. Thus in this case, control over the agenda implies direct control over the outcome. As for Arrow's result, these examples of manipulation are not merely odd outliers. Gibbard (1973) and Satterthwaite (1975) provide a mathematical proof that every voting rule is manipulable under certain basic conditions.

As a whole, these findings and examples constitute what I call the *challenge of Social Choice Theory* against DDM mechanisms. The subsequent section discusses the impact of these results on the above justifications of DDM.

3.3 SCT and the Market View

The previously described findings from SCT and the associated problems invite three major conclusions for justifications of DDM under the market view.

First, the challenge of SCT undermines claims to DDM's instrumental value. This argument departs from the simple logic that an ideal procedure results in an ideal outcome. As SCT shows, no ideal procedure exists, and hence no general standard for evaluating outcomes can be unambiguosly defined. Riker (1982) argues that this defeats a populist interpretation of Democracy, which he circumscribes as "[w]hat the people, as a corporate entity, want ought to be social policy" (Riker, 1982, p. 238). This notion, as he goes on to argue, "fails, therefore, not because it is morally wrong, but merely because it is empty." (Riker, 1982, p. 239).

While I believe Riker's conclusion is arguably too extreme, instrumental accounts advocating the fairness of outcomes must admit that there will always be reasonable disagreement about the nature of good outcomes. Additional assumptions, such as the restriction to single peaked preferences or cardinal utilities, may alleviate such concerns (Black, 1948), yet limit DDM's scope of applicability. Even then, no all-encompassing conception of fairness beyond thin or minimal conceptions of fairness is in sight. Admitting that outcome standards are always relative does not, however, entail that no judgments about the outcomes of decision procedures at all are possible. After all, one can still argue for the usefulness of certain standards under certain conditions. Further, while multiple outcomes may be considered fair, still some outcomes may be clearly unfair. Hence, judging the fairness of outcomes is better dealt with on a more concrete, context specific level.

Second, one can also interpret the results of SCT as a direct attack on

democratic decision procedures. Various examples as well as the general findings by Arrow and Gibbard and Satterthwaite illustrate serious problems for the fairness of procedures. As summarized by Coleman and Ferejohn, "any democratic voting procedure that is fair in the appropriate sense will be normatively defensible but not meaningful, that is, its outcomes will be arbitrary. Only voting that is meaningful and fair can be justified. Unfortunately, no voting procedure can be both." (Coleman and Ferejohn, 1986, p. 11f). Since an optimally fair procedure cannot be identified, either meaningfulness or fairness must be sacrificed (Christiano, 1993, p. 182ff).

This interpretation constitutes a serious challenge to procedural accounts which assert a final value of DDM procedures, since it essentially states that no such thing as a reasonable and fair procedure exists. Yet, the mere existence of equally justifiable DDM mechanisms is not a problem in itself. Since all mechnanisms are essentially flawed, the matter of choosing a decision rule becomes one of picking the lesser evil. As Coleman and Ferejohn (1986, p. 13) argue, a football game is still fair even if the rules of the game could have been designed differently as long as they had been agreed to ex ante and were not chosen in favour of or against certain individuals. In the practical context, this directs attention to the necessity of studying what mechanisms are more or less robust, open to manipulation or disproportionally often lead to imbalanced outcomes.

Third and with regards to the distinction between intrinsic and extrinsic value, the analysis presented implies that a retreat to intrinsic value arguments renders democratic mechanisms less attackable. After all, if one were content with the argument that possession of individually equal voting power sufficiently guarantees a minimal amount of fairness, then there is nothing to be said against this in principle. There is always a fundamental value in involving people in decisions and not making them mere subject to laws they have no say in (Anderson, 2009, p. 215). However, this leaves us with a rather poor notion of fairness, and hence a much weaker claim. Nothing is said about the effective impact of voting power in a wider context, let alone about the fairness of outcomes. Many of the above examples illustrate the importance of considering extrinsic factors for the study of democratic mechanism. For instance, complying with the 'one-man-one-vote'-imperative is insufficient if a procedure, in context, can be shown to be manipulable, arbitrary or structurally unfair. In practice, such findings pose serious problems for a procedure's credibility. The stronger point to be made is therefore that a detailed study of democratic mechanisms is necessary in order to get to substantial evaluations of DDM procedures. For this, extrinsic factors need to be taken into account, even though this usually requires more detailed and extensive argumentation.

As the discussion of SCT and its relation to justifications of DDM implies, neither is there an all-encompassing, unquestionable criterion to judge the *outcomes of decision making procedures*, nor is there such a criterion to judge the *fairness of procedures itself*. Yet, we are not left with a fully relativist view in the sense that no value judgments at all are possible. Rather, I argue for a contextual, case-by case evaluation of DDM mechanisms. Fairness standards for procedures and outcomes must either be defined in a specific context (i.e. claim a limited scope), or require additional assumptions with regards to people's preferences. It is a weakness of many accounts in the literature that DDM is discussed only as an abstract, one-piece entity without differentiation between different types of mechanisms.

I share Coleman and Ferejohn's conclusion that "[t]hough we are concerned by the implications of instability theorems, we think it premature to see these results as establishing the arbitrariness of collective decision making. Rather, these results demonstrate the importance of gaining a fuller understanding of the likely performance of democratic institutions." (Coleman and Ferejohn, 1986, p. 25). In either such analysis, one must be clear about the claims, context and conditions of the arguments proposed. While it may turn out that some procedures are more flawed than it sees at first sight, several procedures may be less so. Especially since many of the criticisms are laden with hidden assumptions itself, many general suspicions may only partially hold. Based on the argument of this section, the goal of my analysis in the individual papers is to discover such factors, and to assess the functioning of various procedures.

For example, manipulability often requires substantial knowledge about, for example, other actors' preferences. This may only be available on rare occasion, as the insights provided by paper 1 show. As a result, the analysis in paper 1 partially discharges the Adjusted Winner procedure from the suspicion of unfairness, and hence its instrumental, extrinsic value as a fair procedure is strengthened. Paper 2, on the other hand, studies the fairness of unregulated, iterated bargaining, and specifically analyses the tendency of such systems to produce inequality. It shows that while single components of distributive mechanisms may be considered fair, they can turn out to be severely biased when performed regularly and on a societal level. Again, this constitutes an example that detailed analysis of mechanisms is required in order to fruitfully assess the fairness value of specific procedures, as advocated in this section.

4 The Forum View

Under the influence of SCT's critique against market view justifications of DDM, democratic theory has been said to have taken a 'deliberative turn' (Bohman, 1997). In the wake of this turn, many scholars tried to locate democracy's virtue in its epistemic quality, by which they hoped to overcome the problems of the market view. In that sense, democracy is often identified not only with voting, but also with deliberation. Like the market view, forum view arguments are of crucial importantance in the literature and therefore deserve extensive attention in order to shed light on DDM's value for forum-view arguemts. For that purpose, I first discuss different interpretations of the epistemic democrat's claim. I then look at *deliberation* and *voting* as manifestations of these claims, and discuss both their merits and pitfalls in turn. Finally, I assess their relevance for justifications of DDM referring to its epistemic value.

4.1 Four kinds of Epistemic Value

Proponents of the forum view claim that DDM is epistemically valuable. As with the market view, assessing this claim requires a categorisation of what this epistemic value amounts to on the basis of Koorsgard's distinctions.

Instrumental epistemic value is found once again on the side of outcomes, whereas epistemically valuable outcomes are 'correct' outcomes. Seeing DDM only as an instrument to bring about correct outcomes, however, presupposes the existence of an external standard of correctness. In an instrumental sense, Procedures are then tools to identify correct outcomes. *Final* epistemic value, in contrast, is better described under the label of 'reasonable' or 'legitimate' procedures, which somewhat departs from a classic epistemic notion. In this understanding, the procedure itself constitutes the standard of correctness; an outcome is correct based on whether or not it came about via a legitimizing procedure, regardless of the content of the actual outcome.

- **Final and intrinsic:** In an *final* and *intrinsic* sense, DDM is epistemically valuable because it constitutes a reasonable procedure in itself. Estlund labels this claim 'rational deliberative proceduralism', where "outcomes are rational only in a procedural sense, not in any more substantive sense" (Estlund, 1997, p. 179). Deliberation is, then, valuable as a "reason-recognizing procedure" (Estlund, 1997, p. 179). Peter (2007, p. 343) extends this notion by highlighting the aspect of public examination of knowledge claims. As a consequence, reasons cannot be evaluated or compared according to an outside standard. While this

stream of argument is interally consistent and hence not attackable in principle, one is – as in the intrinsic, final argument in the market view – left with a rather thin notion of epistemic value in DDM.

- **Final and extrinsic:** A similar yet more extensive interpretation of this claim can be ascribed to Habermas' discourse theory. Most centrally characterized by his notion of an 'ideal speech situation', Habermas claims that properly conducted discourse establishes truth and legitimacy (Lumer, 1997, p. 6). Whatever outcome emerges from such a process has to be considered right and legitimate. This, again, requires no external standard of correctness and DDM should thus be taken as finally valuable. In contrast to a purely intrinsic notion however, Habermas' conditions imply an extrinsic focus: Not the mere consideration of rational arguments alone brings about legitimacy. This can only be realized in a demanding framework of equal, rational consideration of reasons in a free and undominated setting. Whatever outcome such a procedure leads to is legitimized through the procedural standards it adheres to (Habermas, 1970).

- **Instrumental and intrinsic:** In contrast to the previous claims to final value, instrumental accounts of epistemic Democracy rely on a procedure-independent, external correctness standard. In an *intrinsic* interpretation of this claim, properly conducted DDM procedures always find the best way to advance the common good. While in the previous definitorial interpretations, discourse *establishes* the content of the 'common good', the intrinsic instrumental claim states that deliberative procedures merely *identify* what the truth is. Estlund (1997) attributes such an interpretation to Rousseau's notion of a 'General Will': "Rousseau argued that properly conducted democratic procedures (in suitably arranged communities) discovered a procedure-independent answer to the moral question, 'what should we, as a political community, do?' The correct answer, he held, is whatever is common to the wills of all citizens, this being what he called every citizen's 'general will."(Estlund, 1997, p. 183). In such a view, proper procedures *always* identify correct outcomes and are thus intrinsic sources of instrumental epistemic value.

- **Instrumental and extrinsic:** Estlund himself retrieves to an extrinsic notion of value from democratic procedures, according to which a properly conducted DDM process is better able (i.e. more likely) to advance the common good. Hence, outcomes improve when decisions are made democratically, whereas 'improvements' can be measured on

the basis of an external standard. As Cohen (1986, p. 34) describes it, "[w]hat the epistemic populist claims is that, when there is a general will, and public deliberation is guided by the principles that define that will, the decision of majorities about which policies to pursue can provide good evidence about which policies are in fact best". The weak epistemic claim is therefore an empirical claim to a large extent. As List and Goodin summarize, "[t]he hallmark of the epistemic approach, in all its forms, is its fundamental premise that there exists some procedure-independent fact of the matter as to what the best or right outcome is. A pure epistemic approach tells us that our social decision rules ought be chosen so as to track that truth." (List and Goodin, 2001, p. 4). Various theoretical arguments can be made why democratic procedures should be expected to produce the hoped for results. These can be evaluated by scrutinizing actual procedures and the circumstances under which they are employed.

4.2 Moral and Empirical Questions

When assessing the arguments above, the existence of procedure-independent truth-standards deserves special consideration, since it largely constitutes the demarcation line between procedural and instrumental justifications. This question has been discussed among philosophers for centuries, and certainly has neither an easy nor a general answer. There are, however, two distinct kinds of questions discussed in the two camps of the literature, from which an answer can be derived. Scholars in the final-value camp like Habermas usually apply their frameworks to discussions with a moral focus. Examples might be: 'What constitutes a just distribution of goods and power in society?' or 'What rights should be considered universal?' For these questions, it is highly problematic to claim that there is a prima-facie given truth that only needs to be discovered (Sunstein, 2006, p. 63). Instead, one should expect reasonable pluralism with regards to the answers of these questions (Rawls, 1997). As such, it makes more sense to talk about a definitorial notion of democratic processes in such instances.

On the other hand, scholars in the instrumental camp discuss questions with a focus on factual matters. For example: Did person X commit crime Y?, What tax rate produces the maximal revenue?, Which policy proposal provides the largest utility for the general public[4]? In these cases, it makes sense to assume a procedure independent standard of correctness, as these

[4]This, of course, presupposes some measure of general utility, which is by itself a non-trivial question.

questions are essentially empirical in nature. While many of these questions are practically hard (if not impossible) to resolve, one can always try to assess the truth or falsehood of such claims by simply checking the facts – and hence independent of the decision making procedure.

Hence, even though differentiation between these two kinds of questions may be neccessarily vague, the distinction is clear enough in many cases. In the following discussion, I disregard questions with a moral focus and instead concentrate on discussing DDM for resolving empirically-based epistemic questions. This is by no means a judgment of importance or correctness itself. Rather, I do this because the individual papers of this thesis focus on such questions. As a consequence, judgments about the final epistemic value of DDM are disregarded subsequently – not because they are inferior or irrelevant in any way, but simply for the sake of focus and relevance with regards to the topics to come. Henceforth, DDM mechanisms are therefore discussed as tools to *track* the truth rather than to *establish* moral truths or legitimacy.

Analysing this epistemic claim more thoroughly requires discussing DDM's two major truth-tracking mechanisms: *deliberation* and *voting*. The subsequent sections reviews their major issues respectively, and assess central features of the arguments in favour and against their epistemic value.

4.3 The Epistemic Value of Deliberation

The central question for deliberation is nicely phrased by Fearon (1998, p. 44): "What good reasons might a group of people have for discussing matters before making some collective decision, rather than simply voting on the issue or using some other decision rule that does not involve discussion? In other words, what is the point or value of discussing things before making political decisions?". To address this question, I employ a two-part working definition of deliberation: First, deliberation is characterized by a genuine focus on the common good of all participants. This distinguishes it from strategic action in bargaining, which is often considered cheap talk (Bächtiger et al., 2010). Second, I pick up Cohen (2007, p. 224)'s assertion that deliberation is decision-oriented, which distinguishes it from public discourse more generally.

At the core of the argument, deliberation improves knowledge of its participants in different ways. It facilitates individual competency by helping them to a) dispense reasons and relevant information, b) recognize the relevance of different propositions for a decision, c) evaluate whether a certain proposition is true or not, and d) understand the logical implications of more complex propositions.[5] In its ideal form, deliberation therefore decreases un-

[5]see e.g. Elster (1998) for a broader overview.

certainty for participants and provides a better information basis for group decisions. Mill nicely illustrates the core of deliberation's value: "In the case of a person whose judgment is really deserving of confidence, how has it become so? Because he has kept his mind open to criticism on his opinions and conduct. Because it has been his practice to listen to all that could be said against him; to profit by as much of it as was just, and expound to himself, and upon occasion to others, the fallacy of what was fallacious. Because he has felt that the only way in which a human being can make some approach to knowing the whole of a subject, is by hearing what can be said about it by persons of every variety of opinion, and studying all modes in which it can be looked at by every character of mind." (Mill, 1863, p. 42). In the subsequent paragraphs, I analyse the extent to and the conditions under which these ideal concepts hold. Most prominently, I identify two streams of criticism: the detrimental impact of social influence and rational pitfalls.

Social Influence on Deliberation

One major challenge against this ideal type of deliberation centers around the multitude of social influence effects. Such psychological factors can selverely distort and undermine the epistemic value of deliberation. The following list outlines the most important types of social influence, and discusses their impact on deliberation's functioning.

- **Personal Discrimination:** A core assumption of ideal deliberation is that reasons carry their weight irrespective of who advances them. In reality, however, groups are frequently discriminated against on the basis of stereotypes and prejudices. "Some people might be ignored no matter how good their reasons are, no matter how skillfully they articulate them, and when this happens, democratic theory doesn't have an answer, because one cannot counter a pernicious group dynamic with a good reason." (Sanders, 1997, p. 354). Systematic discrimination against people on the basis of class, race and gender is often found in US-jury decisions: Especially the voices of blacks, females and people with little property are given less recognition and weight in jury decisions, while well situated white males are much more likely to be seen as more competent and to be chosen as jury chairman (Sanders, 1997, p. 351ff). For a thorough analysis of this 'epistemic injustice', see Fricker (2007). Biases of that sort hinder the efficient flow of information in social groups. This is especially problematic when relevant information are only possessed by a disenfranchised group in society,

or if the deficient opinion[6] of a vociferous majority dominates debate.

- **Disproportionate Influence:** Information dissemination is also biased in several non-personal ways. On the one side, information that is available to a large number of people is more likely to be accepted, discussed and emphasized than less commonly available information. Gigone and Hastie (1993) call this the "Common Knowledge Effect". Potentially, it results in a broad homogenization of information among groups. On the flipside, sparsely distributed information is frequently underemphasised or even rejected. Peripheral information, as Stasser and Titus (2003) find, is much less likely to prevail in group discussion. As a result, often network structures instead of correctness determine the acceptance or rejection of certain pieces of information.

- **Social desirability:** People have a desire to be accepted and liked by others. One way to achieve this is by adopting opinions of others. This 'social comparison effect' (Sunstein, 2002, p. 179) is especially strong when people exhibit a high fear of invalidity, and when they have a high need for self-confirmation (Kruglanski and Mayseless, 1987). When people form opinions on grounds of social status in comparison to others, the rational basis of deliberation is severely undermined.

- **Limited argument pools:** Individuals are usually aware of only a small subset of all available arguments and facts. Thus, exposure to new information could shift a person's opinion. However, people possessing similar sets of arguments should be expected to group around certain positions on the opinion spectrum. Deliberation, then, takes place mostly in those homogeneous sub-groups. As a result, people will much more frequently encounter arguments in support of their current position, and much less encounter new and possibly contradictory evidence. This confirmatory bias – also labeled 'persuasive argument'-effect (Vinokur and Burstein, 1974) or 'enclave deliberation' (Sunstein, 2006, p. 186) – reinforces pre-existing opinions and drives individual opinions towards the extremes.

- **Polarisation:** Combinations of the previously described effects can lead to polarisation of opinions across society. Polarisation severely hinderes the efficient exchange of information through deliberation (Sun-

[6]While the term 'opinion' is usually more closely associated with the market view, I use it interchangeably with the term 'beliefs' in this section. This is not only for matters of style, but also because some of the biases described in this section apply to preferences and beliefs alike.

stein, 2002). Rather than to converge onto a consensus, deliberation in polarised groups leads individuals to hold more extreme opinions than before. Polarisation can even become a self-enforcing phenomenon when people within homogeneous subgroups strengthen their convictions on the basis of commonly held persuasive arguments, while repulsion mechanisms and discrimination prevent between-group exchange. As a result, beliefs become homogeneous within groups, and heterogeneous between groups. As summarized by Sunstein, "The central problem is that widespread error and social fragmentation are likely to result when like-minded people, insulated from others, move in extreme directions simply because of limited argument pools and parochial influences." (Sunstein, 2006, p. 186)).

The presented pathologies of deliberation may lead to the worrisome conclusion that "when key information is unshared, groups are more likely to select a bad option after discussion than would their individual members before discussion." (Sunstein, 2006, p. 83). In reaction, one is left wondering whether deliberation is beneficial for decision making, or whether the quality of outcomes may even becomes worse.

One crucial gap in the literature regarding these questions is that empirical findings are nearly exclusively focussed on the individual level. This stems from the fact that most of the cited literature employs laboratory experiments or similar psychological research methods to study these mechanisms. An empirical analysis on a group level, in contrast, would be much more complex. By using agent-based models, as I argue in more detail in section 6, these findings can be translated to the aggregate level. This translation, as it will turn out, is far from straight-forward due to potentially complex feedback- and interaction loops. The perspective I take in the individual papers of this thesis try to take this realisation into account, and to make up for this shortfall.

Rational Pitfalls of Deliberation

Not all of the described social influences necessarily constitute instances of irrational behaviour. For example, if a majority of people proclaims a certain proposition to be true, it is perfectly rational to assume that this is in fact the case (Goodin, 2009, p. 4). In absence of further information, it is a reasonable heuristic to assume that claims reported by more people have greater credibility (Mackie, 2009, p. 33). The major difficulty for agents, however, lies in assessing the appropriate weight of socially gathered information: Is someone's statement based on independent information or on statements of

yet other individuals, which may in turn just be following the herd? Even if an individual were to know the exact sources of information on a network level, it would still be practically impossible to calculate the appropriate weight of even single pieces of social knowledge (Golub and Jackson, 2010). Hence, even rational behavior does not guarantee unbiased results.

The *information cascade phenomenon* provides a vivid example of how rational individual action can lead to suboptimal aggregate beliefs and group decisions. Consider the following situation:[7] A group of individuals faces an epistemic decision problem. Two identically looking urns A and B each contain three balls. Urn A contains two black and one white ball, Urn B contains one black and two white balls. At the beginning, one of the urns is picked randomly. Each individual gets to see one randomly drawn ball from the chosen urn; the ball is put back afterwards. The group needs to find out which of the urns has been picked.

Without any further knowledge, an individual's best guess should follow the simple rule: If I see a black ball, I should guess A; if I observe a white ball, guess B. In case no signlas are exchanged and individual decisions are kept secret, every rational actor would report her best guess according to this simple rule. However, now imagine that after receiving their private signals, people openly report their best guesses in random order. Now, apart from the first individual, every subsequent person can infer information from people's previous statements. Crucially, while this procedure improves the knowledge of each person individually, this process leads to a suboptimal group outcome.

To see this, consider the following case: The first person to be asked saw a black ball, and she therefore rationally reports A. From this, all others can infer that she saw a black ball[8]. If person two observed a black ball as well, the evidence for A is even stronger and she will also guess A. If, however, her signal was a white ball, there are now two conflicting pieces of evidence and she is indifferent between A and B. For simplicity, assume that an individual follows her own signal in such cases.

If person three observes two aligning guesses from the previous two persons, her own signal becomes irrelevant. Even if her own signal contradicts the two previous signals, the evidence from the *two* inferred signals overrule her own signal, which is based only on *one* random draw. Hence, person three should rationally guess in accordance with person one and two regardless of her private signal. This effect carries over to all people thereafter, and everyone will rationally disregard her own signal and 'follow the herd'.

[7] The example is presented on the basis of Anderson and Holt (1997).

[8] Assuming they are rational and consider each other to be rational.

As a result, while each individual's chances of giving a correct answer slightly increase, the probability that the majority comes to a correct judgment is much lower than under independent judgments – especially for larger groups. Even though individuals behave rationally, signal exchange of this form is thus epistemically harmful on an aggregate level. Such information cascades occur whenever individuals have an incentive to follow the judgments of others regardless of their own information (Bikhchandani et al., 1992, p. 992). Unfortunately, this is even the case when individuals are capable of forming higher order beliefs about the reasoning capacities of others (Baltag et al., 2013).

By showing how "a lot of people can be wrong" (Lemieux, 2003, p. 21), the crucial challenge that information cascades pose to the epistemic democrat's argument is that overwhelming majorities need not necessarily be the result of overwhelming evidence (Lemieux, 2003). This undermines the value of public opinion by preventing the efficient use of all available bits of information. Mass opinion can therefore convey a false sense of certainty. Anderson and Holt (1997) were able to reproduce information cascade phenomena in laboratory experiments, based on decision problems similar to the urn example described above. Information cascades are often brought forward to explain sudden changes in social systems. Examples range from fashion trends, stock market bubbles or sudden outbreaks of political revolutions.

All the described rational and irrational pitfalls of deliberation invite far reaching concerns about the functioning of democratic procedures, which will be discussed later. Overall, democratic procedures must aim to facilitate deliberation's appeal of common rationality and shared knowledge, while avoiding the rational and irrational pitfalls of social communication. A rather radical solution to this problem is to dispense with deliberation altogether and aggregate individual information independently. This proclaims an epistemic value in voting, to which the next section is dedicated.

4.4 The Epistemic Value of Voting

Earlier, voting has been discussed as a tool for the aggregation of preferences, which is why it may seem puzzling to now discuss voting as an epistemic tool. Understanding this claim requires a fundamentally different interpretation of what a vote expresses. In the market view, a vote expresses one's own preferences – I vote for what I prefer. In the forum view, in contrast, votes must be interpreted as judgments about what best advances the public good (Cohen, 1986, p. 29). Hence, under the forum view, I vote for what I think is best for everyone.

The claim that voting constitutes an epistemic tool for DDM is closely

associated with the Condorcet Jury Theorem [CJT].[9] Imagine a decision problem between the alternatives X and Y, and assume that X is the correct choice. Further, assume that each person has a better-than-random chance of correctly identifying X as the correct alternative, and that individual judgments are stochastically *independent*. Condorcet famously proved that larger groups are more likely to make correct group decisions if individual guesses are aggregated by simple majority rule.

The CJT stands exemplarily for the lager class of 'wisdom of crowds' phenomena. Already Galton (1907) describes an illustrative example from 18th century England where a group of 800 people at a county fair contest guessed the weight of an ox with less than 1% deviation when taking the mean of all individual guesses (Lyon and Pacuit, 2013, p. 1). Wise crowds arguably play a major role in various contexts, such as the generation and provision of knowledge through Wikipedia (Kittur and Kraut, 2008). They can also be used to predict, for example, stock market trends (Chen et al., 2014). For a broader overview and discussion, see Surowiecki (2005).

Due to the demanding assumptions of statistical independence and better-than-randomness, the CJT applies to only a very limited range of problems. To overcome this limitation, various authors generalized the theorem's applicability by relaxing its assumptions. Ladha (1992) extends the CJT result to statistically dependent individual judgments for large enough groups. Grofman et al. (1983, p. 268f) do the same for heterogeneous judgment qualities, showing that the CJT holds as long as the mean individual correctness probability is larger than 50%. List and Goodin (2001) generalize the theorem for choices between more than two options and plurality voting. Comparing different voting rules with regards to their 'truth-tracking abilities', List and Goodin (2001, p. 288ff) find the most popular voting rules roughly perform equally well.

According to List and Goodin (2001), the CJT constitutes "the jewel in the crown of epistemic democrats, many of whom offer it as powerful evidence of the truth-tracking merits of majority rule."List and Goodin (2001, p. 283). Ironically for advocates of epistemic democracy, "that effect can be achieved perfectly well without the very thing that deliberative democrats say is what they want, which is talking face-to-face." (Goodin, 2008, p. 95). In order to analyze what the CJT's findings imply for voting as an epistemic tool, it is necessary to discuss the two major criticisms against the CJT. Each criticisms attacks one of the theorem's two core assuptions: independence and individual competency. In the following paragraphs, I

[9]See e.g. Estlund (2009, p. 15ff) for an extensive presentation and discussion or the original source in Condorcet (1785).

discuss these challenges and conclude that there exists an inherent tension between independence and individual competency.

For the CJT to hold, individual judgments must be *independent* in a stochastic sense[10]. Grofman and Feld (1988) derived from this a necessity to prevent all deliberation. Every communcation potentially harms the independence of voters' judgments and therefore undermines the group's epistemic competency in light of the CJT. In a subtle experiment, Lorenz et al. (2011) show that the sole knowledge about other's positions (without direct communication) is already sufficient to make individual judgments depend on each other and bring social influences to bear[11]. Yet, an exclusion of deliberation does not guarantee independence either, as people's judgments may be be based on a common body of evidence and therefore be *causally dependent* (Dietrich and Spiekermann, 2013). When this is the case, group's aggregate guesses can only become as reliable as the commonly available evidence itself. Thus, the CJT holds only if the common body of evidence is not misleading (Dietrich and List, 2004). Consider, for example, a politician trying to assess the state of the economy before the crash of the housing bubble in 2007. In light of the CJT, one might advise her to question a large number of experts independently and adopt the majority's judgment. For the chosen example, obviously this would have been a bad idea. More independent judges would not have improved an overall judgment because their individual assessments of the situation were based on very similar methods, datasets and theories. Thus, they would have been causally dependent even if one had prevented them from deliberating with each other (Dietrich and Spiekermann, 2013, p. 99). In short, the wisdom of crowds vanishes if many individuals are misled by the same flawed evidence.

Second, individual judgments in the CJT must be *better than random*. If this is not the case, the CJT actually tips to the opposite extreme: If people are on average less likely to be correct, the probability of the crowd to make a correct decision goes to zero for larger groups. Estlund (1993) argues that this requirement invites a second-order problem: It may very well be the case that the average competency of individuals is in fact above 50%. Yet, this an only be established with access to the external truth-standard itself. After all, how can one know that people are reasonably good judges for a certain problem, without knowing the external truth value of the problem itself (Dietrich, 2008)? Arguably, knowing about the latter is even easier than knowing how much people know about the proposition. To take an overwhelming majority

[10]The likelyhood of one person guessing correctly is not correlated with any other person's probability of making a correct guess.

[11]The previously described information cascade phenomenon once again provides vivid illustration of this problem.

as evidence for the quality of individual judgments would hence be circular: All this would indicate is that the evidence is strong, not that it is good. How, then, can individuals check whether they are right or wrong? One crucial way of finding out if your beliefs are correct is by communication with others, which is why epistemology has often been considered a social endeavour (Goldman and Whitcomb, 2011, pt. II). In pure voting, people lack the possibility to revise their opinions through confrontation with good reasons and new information. Hence, the external standard of correctness remains much more unclear for them individually, while interpersonal communication constitutes one crucial way of assessing one's knowledge.

In reaction, one may be tempted to revert to deliberative ways of decision making, only to realize that the problems of social influence and rational pitfalls remain. This leaves deliberation and voting in inherent tension with each other. The benefit of pure judgment aggregation without deliberation is that opinions remain truly independent; There can be no discrimination or any other form of social influence, and all opinions are by definition weighted equally. Yet, this leaves the epistemic value of individual judgments undetermined, as there is no means of changing or checking one's individual judgments. This uncertainty, in turn, can be overcome only in exchange with others – and hence necessarily entails sacrificing voting's essential feature of independence.

4.5 The Discoursive Dilemma: Outcomes or Reasons?

Another intriguing way of comparing voting and deliberation as epistemic tools can be related to List's distinction between a *minimal liberal* account of collective decision making on the one side, and a *comprehensive deliberative* account on the other side (List, 2006). Advocates of the former would, in spirit of the CJT, hold that it suffices to aggregate individual beliefs, regardless of the reasons or information on which these beliefs are based. Decisions under such a framework can be described as 'incompletely theorized' (Sunstein, 1994). Proponents of this approach would argue that reasons are a) irrelevant in light of an actual agreement, and b) infeasible or ineffective, as agreement would be much harder or even impossible to reach (List, 2006, p. 364f). They should thus not be considered in the decision making process. The comprehensive account, in contrast, holds that reasons are essential for DDM and should thus be attributed a central role. Proponents of this framework argue that reason-giving advances individual coherence for one's beliefs regarding a certain decision as well as coherence across decisions over time (List, 2006; Pettit, 2001; Dworkin, 1986, p. 365f).

Similarly to voting and deliberation, these two approaches exhibit an

inherent tension and are to a certain degree irreconcilible. This is illustrated by a class of cases that has become known as the *discoursive dilemma*.[12]

Consider a jury that has to decide about whether or not a defendant is guilty of having broken a contract, in which she obliges herself to perform action X. A judge's evaluation of the defendant's guilt or innocence thereby depends on two conditions: first, whether or not a valid contract existed in the first place; second, if the defendant actually failed to perform action X or not. Only if both conditions are true, the defendant should be convicted. Now assume that the three judges hold the following beliefs:

Table 4: Individual beliefs in the Discoursive Dilemma

	Contract?	Failure?	Guilty?
Juror 1	Yes	No	No
Juror 2	No	Yes	No
Juror 3	Yes	Yes	Yes

If a decision about the defendant's guilt is taken through simple majority voting on the conclusion alone, the defendant goes free. However, if the judges decide on each of the two propositions individually and infers the conclusion therefrom, the court delivers a guilty verdict.

This goes to show that two conditions for a reasonable judgments are inconsistent. On the one hand, if a majority of judges beliefs the defendant to be innocent, she should go free. On the other hand, one would prefer "a collective set of judgments that is itself rational" (List and Pettit, 2002, p. 89), i.e. the group decision should be in accordance with the group's beliefs on each item. Yet, the group ends up with an inconsistent set of beliefs through itemwise aggregation:[13]

[12]The presentation of the discoursive dilemma here is based on an example by List and Pettit (2002, p. 92f), further examples are for instance provided in List (2006, p. 367ff). The problem has also been labeled the *doctrinal paradox* at other points.

[13]Note that the existence of a consistent set of beliefs can be seen as a requirement for a group to be considered an actor itself. If that were the case, collective judgments and decisions could then be treated similarly to individual judgments and decisions. For a further discussion of this issue, see Kornhauser and Sager (2004) and List and Pettit (2005).

Table 5: Collective beliefs through itemwise aggregation

	Contract?	Failure?	Guilty?
Collective Judgment	Yes	Yes	No

Similar to Arrow's impossibility theorem, List and Pettit (2002) prove that no judgment aggregation function exists that fulfils certain minimal conditions of rationality. In analogy to Arrow's theorem for the market view, this renders group judgments instable and non-unique: Outcomes either are undetermined or depend on the judgment aggregation mechanism employed. As a result, one must either accept potential collective inconsistencies or disregard reasons altogether. Van Hees (2007) generalizes the argument beyond binary belief states: Similarly for continuous propositions (e.g. Galton's example of guessing an oxe's weight), it is essentially undetermined wether, for example, mean, median or a weighted average provides the optimal truth-tracking potential.

Bovens and Rabinowicz (2004) argue that premise-based decision procedures are epistemically superior for tracking the truth in most cases, especially for larger and more competent groups. This is especially true if one is interested in correct decisions for the right reasons (as for example Pettit (2001) would demand), but also in most cases if one cares only about correct conclusions itself. In a nutshell, this is due to the fact that a premise-based decision making procedure is more likely to correct individual errors. Further, premise-based approaches are shown to be better able to incorporate new information (Bovens and Rabinowicz, 2004).

However, premise-based decision making procedures and reason-giving comes with the downside of manipulability (List, 2006; Bovens and Rabinowicz, 2004). To see this, consider the above example and assume that the judges have an interest that the group reaches the conclusion they themselves consider correct. Under conclusion-based voting, there is no incentive to deviate and the defendant would go free. Under the premise-based procedure, juror 1 would have a strategic interest in denying the correctness of proposition 1 against her actual beliefs. Likewise, juror 2 could act with regards to the second proposition. In doing so, both jurors 1 and 2 can prevent the jury from delivering a guilty verdict even under the premise-based procedure.

In sum, the discoursive dilemma sheds light on the importance of whether or not reasons for one's decisions should be required or not. While reasons potentially complicate decision processes, they may also improve the epistemic qualities of a process. However, there is also the potential for insincerity when reasons are required, which is less so under conclusion-based voting.

Studies of DDM procedures as means to find correct outcomes must consider these fundamental issues, and how they play out in specific contexts. The discussion of these issues here aimed at creating awareness for such questions, and what general effects might be expected. On this basis, paper 3 of this thesis studies voting procedures with different majority thresholds in combination with reason-based communication. It thus aims to put the major conclusions of this chapter to work, and to assess a procedure's value more concretely.

4.6 Implications for DDM's Epistemic Value

The previous discussion invites three major conclusions for justifications of DDM under the forum view.

First, definitorial accounts which claim that democratic procedures establish truth or legitimacy are well suited for moral questions. In questions of moral right and wrong, a procedure-indepent standard of correctness is questionnable, and thus, democratic discourse can constitute a tool to arrive at collectively legitimatized answers to questions of right or wrong. Due to reasons of focus, I disregard such questions, and instead focus on DDM mechanisms for empirical questions. By definition, these questions exhibit a procedure-independent standard of correctness. Thus, DDM procedures function as potential truth-tracking tools. This distinction is not a trivial insight. It sharpens the focus of discussion significantly and avoids confusion as to the claims that can be reasonably attributed to DDM. The subsequent implications, however, only apply to empirical questions and hence belong to the camp of instrumental justifications.

Second, the epistemic value of democratic procedures is very much dependent on contextual factors, as the above discussion of requirements for ideal mechanisms illuminates. Hence, epistemic value in that sense should be seen as extrinsic. Claiming that democraticness itself is sufficient for epistemic reliability would rightfully be considered absurd. No democratic procedure can always get it right, and thus one must look at specific procedures in order to identify factors that contribute to or undermine a mechanism's epistemic performance. Discussing the question "Are DDM mechanisms epistemically valuable?" has no point because it is way too general. Instead, the study of DDM can be fruitfully advanced by asking about specific properties of specific mechanisms. An example for such a specific question constitutes the research question from one of the subsequent papers, which asks: To what extent does the Adjusted-Winner Procedure lead to efficient outcomes when individuals exhibit strategic behaviour? Answers to such questions can then be fruitfully applied in practice and can also be used to inform philsophical

discussions about DDM more generally.

Third and most importantly, the discussion of this section implies that research about DDM mechanisms should focus on conditions for epistemic quality on a case-by-case basis. The previous section identifies thereby identifies crucial features of this debate generally. While deliberation's key strength lies in a group's enhanced reasoning capacity through shared information, its major weaknesses are the dangers of social influences and rational pitfalls such as information cascades. Voting, in contrast, benefits mostly from excluding these pitfalls through independence of judgments, yet ultimately lacks access to external standards for judgments, which can nearly exclusively be only realised through social interaction. Thus, deliberation and voting stand in a tense relation with each other. As the doctrinal paradox illustrates, this tension translates to the question about the appropriate role of reasons in decision processes. What needs to be shown are therefore efficient ways of reconciling these tensions. Such questions, however, must ultimately be studied for concrete cases, and cannot be convincingly achieved on a general level.

Further, the discussion in this section demands studying combined processes of deliberation and voting. I pick up this issue in the third paper of this thesis and provide such a combined theoretical model. Additionally, while I have shown that individual aspects of social influence are well studied in psychology, much less is known about how such biases play out on a group level. Addressing this questions requires a dynamic perspective which formalizes how individual behaviour interacts and aggregates. I incorporate such dynamic processes judgment aggregation under social influence in several papers. This provides a macro-level perspective on social influence for certain aspects, which has been seldom considered in the literature. As I argue in section 6, Agent-based modelling is a particularly suited tool to trace such macro-level phenomena to individual behavior on the micro-level.

5 A Taxonomy

As I have extensively argued, contextual factors are crucial for the assessment of different decision making procedures. Some of the most important contextual factors invite the development of a baseline taxonomy of DDM problems and processes for an overarching categorisation. I identify three questions as essential to guide the analysis in this work:

- **Market or forum?** The first and most crucial distinction for this work is the one between the market and the forum perspective, as already

discussed at length. In the market view, people hold fixed preferences over which a fair compromise must be found. In the forum view, people hold beliefs about epistemic questions, which must be aggregated for better group judgments.

- **Two or n participants?** The number of involved stakeholders not only plays an important practical role, especially for efficiency considerations. Obviously, one person situations are by definition trivial and hence irrelevant for this endeavour. Two person situations classically require the agreement of both participants.[14] For more participants, there are infinite possibilities of interaction modes and decision rules.

- **Communication or voting?** A further distinction adheres to the mode of interaction. The two major forms are communication and voting, as already discussed. Beyond colloquial associations, both types of interactions can be present under the market- and the forum view. Communication and voting are neither mutually exclusive, nor does either of them occur all the time.

Cases where two people need to bring their personal interests (market-view) in alignment are referred to as *bargaining problems* in the literature.[15] The classic bargaining example from economics is negotiations over prices, but the characteristics of bargaining apply to disputes over the distribution goods in the widest sense. Further, bargaining is often colloquially associated with an unstructured and unrestricted communication process that is concluded by mutual agreement on a compromise solution. For economists, this would be part of cooperative bargaining theory (Harsanyi et al., 1988), for which Nash (1950) provided a first formal representation and solution concept. Bargaining has also frequently been formalised in procedural terms, disregarding the communicative feature of bargaining for the sake of game theoretic descriptions of the process (Nash, 1953; Rubinstein, 1982). Voting, in this context, refers to either making a certain proposal or to agree or disagree with a proposal. In this latter strand, communication is usually considered cheap talk and agreements are not externally enforcible. Bargaining in that realm constitutes an example of non-cooperative game theory and aims at identifying equilibrium solutions that are stable without outside enforcement (Harsanyi et al., 1988). As paper 1 about the Adjusted-winner

[14] To distinguish between different majority thresholds always either results in a dictatorial mechanism or in a unanimity requirement.

[15] The literature also distinguishes between two-person and n-person bargaining situations, with the necessary condition of unanimity rule. For the discussion here, if not stated otherwise, bargaing shall refer to two-person bargaining.

procedure illuminates, sometimes a discrepancy between strategic approaches and fair decision making mechanisms exists. For the case of Adjusted Winner, I show how these two approaches can be harmonised.

For more than two persons in the market view, there are potentially infinite variants of decision rules. Call such situations *competitive groups*. Focussing on voting rules, the most prominent forms are majoritarian methods (simple-, absolute-, super- majorities or unanimity), positional (e.g. Borda's scheme) or utilitarian methods (e.g. range or approval voting, see Hillinger 2005) . For a thorough overview and discussion, see Riker (1982). Schemes for communication in n-person situations can formally be studied through network approaches. An explicit modeling of networks is enabled by the use of ABM (see section 6). Practical examples of such competitive groups can be found in the context of international political organisations like the EU or UN, or with political parties on the national political level. In the economic context, such cases are also referred to as n-person bargaining. Theoretical models in this category are usually very idealized in crucial regards, most prominently embodied through the assumption of perfect rationality. As I argue in section 6, the ABMs are perfectly suited to overcome this limitation and allows the building of models based on realistic behavioural rules.

Now consider the implications of these distinctions for the forum view. Two-person epistemic decision problems in the end boil down to a study of reasonable persuasion between peers. The central question is when and how someone should change her beliefs based on someone else's testimony, and when not to. In short, the subject of two-person epistemic decision problems is *persuasion*. These bilateral interactions play a crucial role as subparts in the papers to come, but are not considered as an object of study by itself. For a conceptual and philosophical introductions, see Lackey (2011) and Kelly (2011), for a psychological perspective, see Perloff (1993). Voting rules, in this context, can also be considered rules of averaging.

Epistemic decision problems for more than two individuals are classic *wisdom of crowds* situations. As already outlined in section 4, key issues of discussion revolves around the merits of mere mathematical judgment aggregation versus aggregation by deliberation (Lyon and Pacuit, 2013). Whereas plenty of research discussing either of the two principles exists, I argue that institutionalized combinations of communication *and* voting are particularly interesting for wisdom-of-crowds phenomena. This stems from the previously identified inherent tension between voting and deliberation. Again, the use of agent-based models allows for the necessary complexification in that direction.

Table 6: A taxonomy of decision situations

	Market	**Forum**
Two-person	Bargaining	Persuasion
N-person	Competitive Groups	Wisdom of Crowds

6 Methodology

Apart from the common topic of DDM, my thesis is unified by a second central theme: The use of formal models, especially the method of agent-based modelling [ABM].[16] In this section, I argue that this methodology is very well suited to analyse and evaluate DDM in light of the justificatory strategies identified above.

ABM's suitability for studying DDM mechanisms can be shown in reference to typical components of such models. For each item, I demonstrate that the features of democratic decisions that were identified in the previous chapters correspond closely with the fundamental building blocks of agent-based models. As a result, ABMs can accomodate the relevant features of DDM easily, and thus facilitate the study of DDM mechanisms as well as arguments about their justifiability.

Agent-based models usually consist of three major components: *agents*, *interaction rules*, and an *environment* (Railsback and Grimm, 2011, p. 38). Additionally, the analysis of agent-based models depends on the specification of *initial conditions* and *model parameters*. Together, one core aim is to explain and reconstruct the emergence of *macro-level phenomena*. On the basis each of these elements, one can reconstruct the suitability of formal models for the study of DDM.

- **Initial conditions:** The analysis of ABMs allows to specify a large variety of initial conditions for a model. This enables a general assessment of decision making procedures even for hypothetical scenarios, which is particularly relevant for assessing a mechanism's robustness. More specifically, this also includes the possibility to study topics where empirical data are hard or even impossible to obtain. DDM is a prime case in that regard.

- **Agents:** The study of decision making can only make sense when the

[16] While paper 1 is based on a purely game-theoretic equilibrium analysis, the remaining three papers each present agent-based models and analyse those by means of computer simulation.

individual decision makers stand in the focus of analysis. In DDM, individual decisions play a central role. In standard models, it is frequently assumed that agents behave perfectly rational, even though heuristic models of decision making are often a more appropriate description (Gigerenzer and Gaissmaier, 2011). ABM allows the specification of empirically informed micro-foundations, which makes theoretical models significantly more realistic. Other methods mostly fail to accomodate such components due to reasons of analysability.

- **Decision making rules:** The framework in which individual actors (inter-)act is centrally determined by the decision making mechanism employed. Procedural rules define what actions are available to agents, such as casting a yes-or-no-vote, making a specific offer, or different ways of uttering one's beliefs. These rules are supposed to guide individual behaviour in certain ways. From a system-design perspective, one is concerned with how those rules impact the overall process. Different decision or communcation rules can be easily incorporated into computational agent-based models and can thus be treated as explicit model parameters. This enables a detailed study of the impacts of such mechanisms in various contexts, and invites an institution-centered analysis of DDM.

- **Contextual factors:** As analysing the market- and forum-view justifications has pointed out, DDM should be seen as extrinsically valuable in order to inform substantial justifications of DDM. Thus, both agents and decision rules need to be situated within relevant contextual factors. A limited time frame, behavioural constraints outside the explicit decision mechanism (such as evolutionary pressure), or different ways of connecting are examples for factors that limit agent behaviour. They should be incorporated into models of DDM in order to evaluate how they influence processes and what boundary conditions they imply for the functioning of procedures. Using the ABM methodlogy allows for incorporating such contextual factors in an explicit way whenever they are considered relevant. While this may seem trivial, mathematical solvability often drives the choice of model features in classic models. The use of ABMs overcomes this limitations, as I argue below.

- **Emergence of macro-level phenomena:** ABM provides structures for explaining the emergence of macro-level patterns from local interactions on the micro level (Epstein, 2006). For DDM, the macro patterns of interest can be located in regularities in outcomes, such as: Does

some mechanism X result in fair (or correct) outcomes? Or on a procedural level: Does process Y structurally favour a certain group of actors? To analyse these patterns, more detailed theoretical constructs are required to deal with the complexity of group decision making systems. ABM is able to provide such a framework for analysing complexity that emerges from individual interactions and allows considering decision rules and environmental factors alike.

In the rest of section 6, I analyse various aspects of these components in more detail. Beyond the method's perfect aptness, I conclude by arguing that this thesis helps to advance applications of agent-based models generally and thus also makes a methodological contribution.

6.1 Virtual Experiments

Many questions we ask about DDM mechanisms are often *hypothetical* in nature, a generic example being: 'Given constellation c, how would decision rule r and process p interact and to what result would this combination lead?'. Rather than merely showing that mechanism x performed well in some situation y, one wants to show that mechanism x can be relied upon in types of situations similar to y due to its characteristics a, b and c. Hence, a decision mechanism's crucial aspect is its robustness. A mechanism must be able to deal with a broad range of situations, and not just those which had occurred in the past, or those which one hopes to occur in the future. Ideally, one can trust a stable mechanism regardless of what future situation one encounters, or at least in a reasonably large group of cases that can be expected to occur.

Robustness, in that sense, necessarily presupposes an analysis also of those cases which are not realised. For example, a voting rule must lead to fair outcomes no matter what preferences or beliefs people hold. This question could, for instance, be asked about the representativeness of a political system, in which the number of parties increases. Particularly agent-based models facilitate such virtual experiments (Macy and Willer, 2002). Instead of predicting certain specific outcomes under specific circumstances, such generalizing experiments "enrich our understanding of fundamental processes" (Axelrod, 1997, p. 25). Theoretical models can therefore guide our understanding of specific DDM mechanisms, which is essential for their thorough normative assessments.

Furthermore, key characteristics of DDM in reality are often *hard to observe*, or even unobservable in principle. For example, individuals' preferences or beliefs can often only be measured in connection with the current

decision making mechanism. Similar to the 'revealed preferences' approach in economics, people's preferences can often only be inferred from the choices they have made. As the above discussion has aptly shown, one's choices are often not a straight forward result from people's genuine preferences, but usually depend on various contextual factors – most importantly, the prevailing aggregation mechanism. Also, aspects like strategic decision making or social influence can severely distort the measuring of preferences and beliefs alike. Various aspects of group decision making therefore exhibit "black holes in available empirical data" (Boero and Squazzoni, 2005, sec. 2.8). In many cases, the dynamic nature of such processes renders this problem even more prominent. Thus, the functioning of a robust mechanism, as argued before, not only should be independent of its inputs. Sometimes, the empirical study of democratic mechanisms is prevented by observability problems, and they must necessarily be studied without access to these inputs.

6.2 Managing Complexity

Democratic group decisions are best viewed from a *complex-systems perspective*. Individual model components can often be described in very simple terms, yet they frequently lead to outcomes that are hard to predict. In an abstract way, the cases described above illustrate this feature: Simple majority rule can lead to incentives for strategic voting; Communication may be easy to describe on a peer-to-peer basis, yet becomes exponentially more complex already for small networks, especially when network structures are dynamic. Each of my four papers can be seen as an expression of this emergent complexity. The bounded-confidence mechanism in papers 2 and 4, the majority rule in paper 3 or the baseline bargaining game in paper 2 are all based on simple principles, yet in all of these cases, non-linear connections and feedback loops result in unpredictable patterns on the group level. Opinion dynamics under bounded confidence further exemplify the occurrence of tipping point phenomena, which constitute a common feature of complex systems: Minor changes on the individual level can lead to severe changes on the macro-level (Helbing and Lämmer, 2008). Such behaviour is, for example, observed for the models in papers 3 and 4.

ABMs explain complex phenomena on the basis of few basic principles of interaction. Holland (2000, p. 188) describes this principle as follows: "Much complexity can be generated in systems defined by a few well-chosen rules. When we observe emergent phenomena, we ought therefore to try to discover the rules that generate the phenomena. In the particular format we have developed, we need to find a constrained generating procedure that generates the emergent phenomena. In so doing, we will reduce our complex

observations of emergence to the interactions of simple mechanisms." Similarly, Epstein (2006) argues that ABM explain complex systems "from the bottom up" (Epstein, 2006, p. 15), thus tracing back the system's behaviour to simple interaction rules.[17]

While I argue below that ABMs allows for the complexification of models, a major part of their explanatory power for complex phenomena is derived from reductions of complexity: ABMs enable a researcher not only to include empirically informed components, but at the same time, irrelevant system components can (and should) be disregarded. This way, it becomes possible to carve out causal mechanisms more clearly (see e.g. Epstein 2006; Holland 2000).

Further, the use of ABMs comes with the advantage of ontological correspondence (Gilbert, 2007, p. 14): Each object in the model can be clearly assigned to the object in the target system it is supposed to represent. Other methodologies, such as equation-based modelling or statistical analysis, lack this correspondence. Ontological correspondence is not just a virtue in itself, but significantly simplifies model building and analysis (Gilbert, 2007, p. 14). Simiarly, Casini and Manzo (2016) argue that "[t]he detour through the deep, object-oriented infrastructure of ABM is not a purely technical digression. It also helps to better see the source of ABM's flexibility for designing models of mechanisms that aim to directly represent aspects of social life that statistical methods, as well as other simulation-based modeling strategies, are not able to handle with similar easiness." (Casini and Manzo, 2016, p. 15f).

To a significant extent, DDM's complexity stems from its dynamic nature, for example when communication and voting appear in a temporal process. As for opinions more generally, many applications demand a dynamic rather than a static perspective. ABM can provide such a perspective, since they are inherently dynamic (Miller and Page, 2009, p. 83). As a result, time becomes an explicit model component. Studies of DDM can therefore address questions of *efficiency*, which is highly relevant for information aggregation and decision making. This is done, for example, in paper 3.

With a system-centred perspective that focusses on the target system's crucial features and their mutual dependence, one is able to reasonably capture DDM. Classic empirical approaches, in turn, mostly look at isolated system components, for example by studying individuals' beliefs, or by focussing on single instances of bargaining. Oftentimes, this misses out on important connections between different system components. The essence of this argument can also be phrased in terms of Coleman (1990), who argued that social interactions must be analysed within an interconnected micro-

[17]For further discussion of the issue of complexity and ABM, see also Scheller (2016).

macro-framework. Macro-level observation can make sense only in light of explanations that trace the observed phenomenon back to the behaviour of individuals. Individuals, in turn, are situated in a broader context to which they react and adapt. Looking at isolated parts of this system can be useful, but the connections and interaction effects between these steps must not be underestimated. Classic models fall short of capturing this complex, dynamic nature. As Epstein (1999) argues, "even perfect knowledge of individual decision rules does not always allow us to predict macroscopic structure" (Epstein, 1999, p. 48). In light of this challenge, ABMs have often been considered an appropriate tool, especially for managing the logic of aggregation in these processes. ABM allows studying micro-macro mappings, which are crucial for understanding DDM mechanisms. ABMs thus provide "techniques for 'projecting up"' (Epstein, 1999, p.48) from micro-foundations to macro-level effects.

An example for the role of emergent complexity in DDM is discussed in paper 4, which analyses the impact of emotional appeals on party support. To update their opinions, voters simply average over the opinions of all voters with similar opinions. Parties, on the other hand, can choose whether or not to employ emotional appeals. A simple yet intricate feedback loop exists between these two components: A party's emotional appeals influence the degree to which voters become more or less closed-minded when communicating with others. On the aggregate level, these intertwined processes result in a non-linear relationship, in which stronger emotional appeals sometimes lead to more, and sometimes to less voters being attracted by the party. Without this complex-systems perspective, especially the dynamics of this process could not be studied in detail. For the case presented in paper 4, the dynamic analysis points to the importance of addressing voters who are on the verge of being attracted by an extreme party at key moments in time.

6.3 Empirically-informed Microfoundations

Crucial to modelling in general is an appropriate degree of simplification, depending on the model's purpose. A modeler's goal must be to depict no more and also no less than those components of a target system which she considers relevant for understanding a system's behaviour. Ideally, relevance should be the major determinant of which factors to include, and which factors to neglect. In practice, however, mathematical solvability imposes strong restrictions for model design. Especially for game theoretic equilibrium solutions, mathematical solvability often requires more simplicity than would be appropriate (Railsback and Grimm, 2011, p. 9). Through the use of ABMs, the choice of model entities can more directly be guided by the

question 'what is relevant?', rather than by the question 'what is possible?'.

Realistic micro foundations are essential for modelling DDM. Gilbert (2008, p. 41ff) classifies ABMs into abstract, middle-range and facsimile models. In essence, these model types vary in the degree of abstraction, and to what extent they are informed by empirical data. Usually the more empirically driven they are, the more precise and the less general their applicability. To use Gilbert's terminology, the models I present in the individual papers constitute middle-range models. On the one side, the above problems of data availability prohibit the setup of facsimile models that depict specific cases in very much detail. Even so, such models would not be general enough for the overall assessment of a DDM mechanism anyway. On the other side, knowledge about general empirical characteristics can be obtained. For example, one can employ findings about distributions of opinions for initializing an agent-based DDM model. One can also use psychological findings about pattersn of human behaviour. To use an example from paper 4, individual responses to emotional stimuli are well studied in the literature on Affective Intelligence Theory (Marcus et al., 2000). The identified behavioural pattern that fearful individuals rely less on partisan habits can be used to inform an ABM of political opinion formation.

Middle-range models employ these kinds of input in order to "describe the characteristics of a particular social phenomenon, but in a sufficiently general way" (Gilbert, 2008, p. 42), which is the mixture of specificity and generality that is necessary for evaluating specific DDM mechanisms. Thus, "experimental data and empirical knowledge [can be used] to support theoretical abstractions" (Boero and Squazzoni, 2005, sec. 4.57). The use of empirical data can thus be seen as a kind of model calibration (Boero and Squazzoni, 2005, pg. 2.6), and constitutes a "fundamental ingredient to support mechanism-based theoretical explanations" (Boero and Squazzoni, 2005, sec. 2.14).

Another assumption that standard equilibrium models usually require for solvability is the homogeneity of agents. As already argued by Kirman, "the reduction of the behavior of a group of heterogeneous agents *even if they are all themselves utility maximizers*, is not simply an analytical convenience as often explained, but is both unjustified and leads to conclusions which are usually misleading and often wrong." (Kirman, 1992, p. 117, Kirman's emphasis). Employing the ABM methodology overcomes this shortfall because heterogeneous agents are easy to implement, simply because one can program the initial characteristics of agents, and simulation has no problem dealing with all kinds of agent types (Casini and Manzo, 2016, p.16); (Bandini et al., 2002). In the models of papers 2, 3 and 4, agents are all heterogeneous in at least one crucial feature. Especially the different player types in paper 2

constitute a vivid example of this heterogeneity, and how it can be fruitfully employed.

The discussion of empirical model foundations must not be mistaken for the statement that realistic agent descriptions are valuable in itself. More realistic models are not necessarily 'better'. I agree with Doran (2006) that a model must be sufficiently detailed for addressing the model's purpose, and at the same time leave out as many irrelevant components as possible, in oder to carve out decisive causal mechanisms more clearly.

Also, this discussion must not be mistaken for the unrealistic claim that ABM allows the study of models of infinite complexity. Manageability of model parameters must always be guaranteed, and still provides a strong limiting factor on the possibilities of ABM analysis. Still, while there usually exists a trade-off between descriptive accuracy and solvability, one can sometimes have a bit more of both by using ABMs.

In summary, agent based models help to describe and analyse DDM processes because they allow a more appropriate description of such problems with regards to agents, rules and environment, and the connections between those three components. An appropriate degree of empirical adequacy is most vital when it comes to modelling individual agent behavior because it constitutes the most important component in decision-focussed models. For this reason, the following section discusses this issue further by focussing specifically on the notion of rationality.

6.4 On Rationality

In the discussion on empirical micro-foundations, special attention must be devoted to the rationality of agents' behaviour. Rational choice models have a long tradition all across social science, and are classically expressed through the use of game theory. The analysis of such models usually involves the identification of equilibria. While rational behaviour is often assumed for the sake of mathematical solvability, Kahneman (2011) and many others have aptly shown its empirical inadequaciy for a large class of cases. Instead, Gigerenzer and Gaissmaier (2011) emphasize the importance of heuristic decision making mechanisms for describing individual behavior.

The empirical inadequacy of rational choice severely undermines many a model's credibility.[18] After all, there is no prima-facie reason why rational choice models should in any way behave similarly to the real world, when individual behaviour can be shown to consistently and significantly diverge from the requirements of rationality. This is true for justifications of deom-

[18]Not all models, however, claim descriptive accuracy.

cratic mechanisms in particular, since decision making almost by definition constitutes a core aspect of democratic group decisions. An illustritative case of such arguments is the frequent attack against deliberative accounts of Democracy due to its demanding assumptions on the side of individuals (see e.g. Somin 2010 for such a criticism), which can be summarized by the question: Why would deliberative procedures produce good results when performed by non rational actors if their merits are shown only for groups of rational actors?

As Gilbert and Terna (2000) argue, modelers often make these assumptions "not because they believe the economic world really does have these characteristics, but because otherwise their models cannot be analysed." (Gilbert and Terna, 2000, p. 58). Hence, the driving force behind this empirical inadequacy is mathematical tractability. As a result of the omnipresence of perfect rationality in many areas of social science, Simon (1955) argues that many modelers have been asking the wrong questions: Instead of studying the interactions of perfectly rational subjects, one should ask: "How do human beings reason when the conditions for rationality postulated by the model of neoclassical economics are not met?" (Simon, 1955, p. 377). In contrast to classic game-theoretic solution concepts, ABM is "agnostic on the kind of micro-foundations a modeler should subscribe to" (Casini and Manzo, 2016, p. 16), and thus enables a modeler to accomodate all kinds of agent-behaviours. Hence, also the heuristics of human behaviour as identified, among others, by Kahneman (2011) or Gigerenzer and Gaissmaier (2011) can inform agent behaviour in ABMs. Paper 2 does so by incorporating a number of agents with different heuristically informed bargaining strategies. Likewise, the behaviour of agents in papers 3 and 4 is informed by the heuristics of social influence from section 4.

Despite these criticisms against the rational choice approach, papers 1 and 2 illustrate that it can be fruitfully applied for certain types of questions. In the case of the Adjusted-Winner procedure, the problem of its original formulation lies not in the fact that agents are unrealistically described as too rational, but rather that the procedure is intended only for honest agents. The original AW-model disregards the possibility of rational, strategic exploitation of the mechanism. My contribution closes this gap by analysing the consequences of (mutual) strategic behaviour. Thus, the model I provide is not based on the claim of descriptive accuracy. Instead, I argue that the possibility of such behaviour exists, and thus the potential threat of such behaviour must be evaluated. Furthermore, complexity of the two-person bargaining situation is still analytically manageable, and hence rational behaviour does not demand unrealistically strong computational capacities from agents.

Apart from the fact that rational behaviour is often empirically inadequate, various strategic situations exist, for which "individually optimal behavior is uncomputable in principle" (Epstein 1999, p. 50). This is, for example the case, when other actors must be assumed to act irrationally. Hence, strategies can only be *ecologically rational* in the sense that they are "adapted to the structure of the environment' (Gigerenzer and Gaissmaier, 2011, p. 457). The environment, in turn, is crucially constituted by other agents and their characteristic behaviour.

In paper 2, we explicitly discuss the necessity for a broader notion of rationality. More complex situations require a more complex conceptualisation of rationality. In the case depicted in paper 2, this is due to the intricate connection between rational behaviour, information acquisition and survival. One of the paper's key conclusions is that the standard notion of rationality as utility maximisation does not constitute an optimal strategy to choose in the non-trivial bargaining situation. Hence, rationality is context-senstive, as also argued by Galeazzi and Franke (2017). As a result, paper 2 does not take rational choice as a descriptive component, but considers rationality as a normative behavioural rule, asking under what conditions standard notions of rationality in fact lead to optimal behaviour.

The analysis of ABMs also illustrates how individual rationality can be decoupled from collective rationality (Epstein 1999, p. 48). As shown for the information-cascade example, individually rational behaviour does not necessarily lead to collectively rational outcomes. While this phenomenon is already known to occur in collective action problems (see e.g. Ostrom (2015)), using ABMs can help to identify this property also for more complex systems. We show this for the emergence of inequality in paper 2. The use of ABMs makes it easier to engage with a system-centric perspective, which allows to shift the focus of attention between micro- and macro-level more easily.

Vice versa, although individual rationality is sometimes considered a requirement for e.g. fairness or epistemic value on the aggregate level, this is not necessarily so. Paper 3 shows how seemingly irrational behaviour (in that case: bias against contrasting evidence) can have a positive effect on group performance. Thus, a system can be considered collectively rational (i.e. perform well according to relevant dimensions), even though individual behavioural components cannot reasonably be described as rational.

Under these considerations, the agent-based methodology can be employed to identify rational behaviour in complex environments. This is evidenced by the analysis in papers 2 and 4. Paper 2 sheds light on the exploration-exploitation trade-off that rational actors face as a result of the inherent tension between short term payoff-maximisation and the long-term

goal of collecting reliable information. Paper 4 analyses the efficiency of different party strategies, and how they translate into electoral success. The paper thus addresses the question whether or not it is beneficial to employ emotional appeals in party propaganda. Without a dynamic perspective, this question could not have been sufficiently addressed. Ironically, one core theme is the impact of emotional states on the side of voters, and thus paper 4 also illustrates how empirical findings about affective human behaviour can go hand in hand with an analysis of rational behaviour: While parties are assumed to choose their strategy rationally, individual voters are modeled as affect-driven. With regards to both papers, this extended notion of rationality clearly goes beyond a thin notion of rationality as a internal consistency, and prescribes context-specific optimal behaviour.

6.5 A contribution to Agent-based Modelling

Apart from its contribution to justifications of DDM generally, and insights into various decision making procedures in specific, my thesis also offers a number of methodological insights. As I argue above, agent based models are a suitable tool for the analysis of democratic group decisions because formal models are very well suited for capturing the key elements of group decisions. Beyond this argument of ABMs suitability, my thesis illuminates new possibilities of applying agent-based modelling in social science and philosophy.[19]

Agent-based modelling is still a relatively recent scientific methodology, mostly due to its demanding computational requirements that have only become available through the broad dissemination of computers (Macal and North, 2005). As a result, its applications are still rare and are often met with a significant degree of skepticism. Thus, ABM is not a mainstream research method (yet). Without any doubt, there is a plentytude of legitimate criticisms against it. Most centrally, ABMs are criticised for their lack of external validity. This criticism argues that the results which are derived from the models carries little relevance for real world processes because these models are simply too remote from the actual systems they try to depict.

For some models, this is deliberately so because their core aim is not a realistic description of a certain system. Rather, an abstract model's goal may be to illustrate core causal interactions, or help to spell out a certain theoretical concept in a formal and detailed way (Gilbert, 2007, p. 41f). Such models can be improved by incorporating empirical findings from the individual level. Papers 3 and 4 do this by referring to a large spectrum of

[19]This claim mainly concerns the works in papers 2, 3 and 4, in which the ABM methodology is centrally applied.

psychological research to justify the individual decision rules they prescribe. This helps to solve the problem that such individual-level findings are often hard to transfer to the aggregate level because of non-linear relations between different system components. Similarly, ABM has been recommended for this task by Smith and Conrey (2007).

In the case of paper 2, we thoroughly argue that the abstract situation which the model depicts is in fact a core building block of economic interaction. Such a model must necessarily remain abstract, since one needs to reduce complexity in order to understand the fundamental mechanisms of these procedures. In this case, the degree of abstraction is necessary for the model's purpose, which mainly consists in making an argument about a certain distributive system as a whole. The model thus constitutes a crucial building block in a larger normative argument by showing how individually justifiable bargaining structures can lead to massively unequal outcomes on a societal level.

In summary, the applications in my papers therefore provide further examples how ABMs can be fruitfully employed in social science and philosophy alike. They show how the use of empirical findings can augment the credibility of agent-based models, and, in turn, how empirical findings on the individual level can be transferred to the macro level. My papers also illustrate how philosophical arguments benefit from the system-centered perspective provided by ABMs.

7 The four Papers

My cumulative PhD consists of four individual papers. Each of these four papers discusses DDM mechanisms either under the market- or the forum perspective. Based on my previous argument that DDM cannot be evaluated categorically, each paper addresses specific questions about the advantages and disadvantages of different mechanisms for different purposes.

Further, all papers employ formal modelling techniques to do so – one paper employs classic equilibrium solution concepts, three papers are based on ABM. As I have extensively argued above, formal and Agent-based modelling are highly suitable tools for studying DDM.

The taxonomy from section 5 helps to categorize and locate the four papers in the larger frame of this project. Subsequently, I give a brief description of each paper's content, how it fits into the previous taxonomy, and what it contributes to the literature in general and to justifications of DDM more specifically.

These are the titles of my four papers, numbered in the order in which

they are discussed below:[20]

1. *Mitigating the problem of Manipulation in the 'Adjusted Winner' Procedure*

2. *Rationally Poor? – What the Emergence of Inequality can Teach us About Rational Behaviour*

3. *When do Groups get it right? – On the Epistemic Performance of Voting and Deliberation*

4. *Fear Appeals as a Political Strategy – A Theoretical Exploration*

7.1 Paper 1

Adjusted Winner (AW) is a fair division procedure by which a set of objects can be divided between two persons. Each person allocates 100 utility points to a fixed set of objects according to their individual preferences. Objects are then distributed and partly redistributed until each person receives objects worth the same utility to her so as to guarantee a fair outcome, while at the same time maximizing each individual's utility. Thus, AW claims to provide a *fair* and an *efficient* solution.

In it's original form, AW works under the assumption that people state their preferences honestly and not strategically. In the paper, I envisage AW as a strategic rather than a cooperative game. I scrutinize what impact this has on the appealing properties of AW's original solution under honesty. In the analysis, I find that manipulation has not as severe consequences as one might initially think, especially because potential losses through failed manipulation are much more severe than potential gains. Further, as long as both players manipulate equally strongly, there is no impact on the fair and egalitarian solution.

Referring to the taxonomy of decision problems in section 5, AW is applicable to classic two-person bargaining problems. The study of fair division procedures in general and that of AW in the model of paper 1 constitute an interesting case with regards to the market-forum distinction. On the one hand, I introduce the assumption of rational utility maximisation for both agents. This implies a clear market perspective since both players, care only about their individual gains. On the other hand, Brams and Taylor (2000)

[20]With these four papers, I fulfill all the point-system requirements imposed on cumulative Dissertations by the Fachgruppe für Politikwissenschaften of Bamberg University. For further details, see page 61.

intended their procedure to work on the assumption of honesty, proclaiming the identification of both a fair and efficient procedure as its central goal.

What makes the situation even more interesting is the fact that – except for the trivial case when both players have the exact same preferences – the allocation of the heterogeneous set of goods in AW is not a zero-sum-game but a positive-sum-game. Hence, from a general welfare perspective, there are better or worse outcomes to be found and the game exhibits a cooperative component. This subpart of the decision problem can be interpreted as an epistemic problem: Agents must discover a solution that is better for everyone. Hence, subjects both compete with the opponent for shares of goods, but could also partly benefit from finding an efficient solution. AW therefore constitutes a mechanism that settles disputes over competing interests, but also implicitly tackles the epistemic problem of reaching a pareto-optimal solution.

My analysis shows that both these properties are robust to strategic behaviour under certain conditions. As argued above, if a decision making mechanism relies on individuals' honesty, the procedure's trustworthiness is severely undermined in practice. By showing the relative immunity of AW against strategic action, this challenge can be countered. Even if actors are assumed to act strategically, the appealing properties of AW's fair and efficient solution are upheld. For this result, the consideration of contextual factors such as imperfect information and the resulting risk calculations are of crucial importance.

In abstract terms, my paper studies the extrinsic conditions for the instrumental value of AW as a democratic procedure. Following the above argument, manipulative behaviour stands in the center of the analysis. This implicitly requires an extrinsic justification of AW as a democratic procedure. For definitorial claims, the argument about manipulability would not even be necessary. Further, the paper's argument adheres to an instrumental justification of AW because the procedure clearly aims at fair and efficient *solutions* to the bargaining problem.

7.2 Paper 2

Paper 2 was written in co-authorship with Dominik Klein and Johannes Marx.[21] It presents a model in which agents play a simple bargaining game

[21] Johannes Marx is the first supervisor of this thesis and is professor for Political Theory at Bamberg University. Dominik Klein is a postdoctoral researcher at the chair of Political Theory, University of Bamberg. With regards to my personal contribution to this paper, I was substantially involved in all steps of the paper's development, and was primarily responsible for developing the model's game theoretic setup and for model programming.

against a randomly chosen partner multiple times. Players have to strategically choose how long to insist on a large share of payoffs, and when to concede a larger share to her opponent. Players collect information about potential opponents from previous bargains and choose their bargaining strategy according to different heuristic strategies.

Centrally, we find that inequality can arise from iterated interactions of rational agents. This macro-level result is directly related to the fact that round-by-round expected utility maximisation is outperformed significantly by other bargaining strategies – both in terms of long term payoffs and in terms of survival. To a large part, this can be attributed to the complex interaction between information acquisition and strategy choice: Aggressive bargaining strategies are structurally advantaged by the more precise information they obtain, while weaker strategies systematically gather not only less, but also biased information.

Similar to paper 1, this paper also envisages two-person bargaining situations in a market view situation. However, the bargaining process incorporates several decisive contextual factors for the analysis, most prominently iterated exchanges, random partner selection and anonymity. This mechanism is not deliberately designed to bring about fair solutions. Rather, the bargaining game in paper 2 depicts an unregulated exchange situation in the economy. Accordingly, the model asks what outcomes result from this 'natural' mechanism in the absence of state intervention. We show that massive inequalities consistently result from such exchanges under the given conditions. Hence, the paper provides an argument for the necessity of regulation for such cases, since inequalities arise despite the prima-facie (intrinsic) fairness of the process itself.

The paper contributes to justifications adhering to the extrinsic, instrumental value of DDM under the market view. It shows that fair outcomes can only be achieved when regulations are imposed in iterated, two person bargaining encounters. The paper supports the claim that unregulated bargaining is unfair, even if distributive mechanism itself provides no structural advantages. On the one hand, this calls for increased attention to hidden structural factors in DDM. It is not enough to design a structurally symmetric process if strategic complexity severely undermines its fairness. On the other hand, the unregulated system is clearly not robust in terms of fair outcomes. Implicitly, the paper again assumes an instrumental view on unregulated bargaining as a DDM procedure. It poses a challenge to arguments that locate the causes of inequality in individual differences in mental or physical capacities. In consequence, the model's insights shift the burden of proof upon those who want to justify existing inequalities. In short, the model shows that group decisions about distributive questions by un-

regulated bargaining can be considered fair only in a trivial, but not in a substantial sense.

In the standard literature, bargaining is usually discussed under a classic rational choice perspective. As a result, many contributions depart from unrealistic assumptions about agent-behaviour and neglect important contextual factors, such as iterated exchanges and learning. Enabled by the use of ABM, our paper overcomes this shortfall and explicitly incorporates various heuristic strategies that are in part empirically informed, and in part embody classic strategic rationales from rational choice theory.

Paper 2 contributes not just to discussions of DDM or inequality, but also provides crucial insights for theoretical conceptions of rationality and individual behaviour. We show that a thin notion of rationality as maximisation of expected utility falls short in several regards. A thicker notion of rationality should, for example, take into account the endogeneous relation between information acquisition and strategic action, which can (as in the paper's model) trigger unforseeable dynamics that drive both individual behaviour and aggregate outcomes. In consequence, such a notion of rationality may be ambiguos – different goals may stand in irresovable conflict to each other. For the study of democratic mechanisms, this realisation in itself may complicate the analysis at times: If not even rational behaviour can be reliably predicted, it is much harder to infer outcomes from democratic procedures in non-linear environments.

Paper 3

In contrast to papers 1 and 2, the model in paper 3 deals with classic wisdom of crowds- situations. Also, paper 3 explicitly considers the interaction between communication processes and voting. In the model, 50 agents each receive an imprecise signal about which one of four discrete alternatives A, B, C and D is the best choice for the group. Their interests in finding the right solution perfectly align, and hence they need to identify the correct alternative based on their individual information. To do so, they communicate and vote according to simple procedures. The model analysis addresses the simple question: When are groups more or less likely to come to correct decisions? Larger majority thresholds are found to improve decision making quality, but only under the condition that people are open to listening to others and gridlock is prevented. However, if independent sources of information are available, a reasonable scepticism against socially acquired information can be beneficial, since it helps to avoid overly quick convergence onto a false consensus – something that stricter majority requirements are not capable of hindering.

The paper provides insight into the vice and virtues of different structural features when groups face epistemic tasks, with a focus on combinations of voting and communication. Combined voting and deliberative-procedures have not received much attention in the literature, despite the fact that real world applications are seldom limited to either component. Paper 3 aims to provide a theoretical construct to overcome this shortfall of the literature. In the model, specific features of the communication process and the voting mechanism interact in complex ways. As a result, they have a non-linear impact on the quality and efficiency of the group's epistemic performance. For example, larger majority requirements do not necessarily make groups perform better. This is only the case if communication is reasonably open-minded. Openness for communication, in turn, is also not generally beneficial for outcomes. It can make decision-making more prone to herding effects rather than foster the arrival at rational consensus. Insights of this type are helpful for the design of decision making procedures in reality, or when flaws in existing mechanisms must be identified.

By describing those intricate discoveries, the paper is directly relevant to justifications of DDM. The results vividly illustrate the importance of contextual factors for the wisdom of crowds. Implicitly, the model picks up the distinction between conclusion-aggregation in voting and the exchange of reason in communcation, which has been discussed in abstract terms for the discoursive dilemma in section 4. The paper discusses the truth-tracking capabilities of a certain DDM mechanisms, and studies its requirements, limitations and complications. It thereby addresses the core claim of epistemic democracy directly, providing a nuanced answer to non-trivial questions for a clearly identified scope.

Paper 4

Paper 4 analyses the role of emotional appeals (specifically appeals to fear) in political opinion dynamics, and how they can be employed strategically by political actors (specifically by extreme parties). Findings from Affective Intelligence Theory suggest that people's views can be more or less influenced by other people, depending on their associated emotional state. By influencing people's emotions, political actors can thus influence the dynamic process of public opinion formation. In the model, they are assumed to do so in order to increase their electoral support. Voters, in turn, influence each other by updating their opinions under a setting of bounded confidence. According to the paper's conclusions, fear appeals can be an effective yet dangerous tool for political actors: On the one hand, they increase a party's reach for new potential supporters. On the other hand, they also increase the risk of losing

former core supporters to more moderate groups. Extreme parties should therefore clearly differentiate their political position from others. This can be seen as a micro-level rationale for poltical product differentiation, which is a popular strategy among radical parties (Kitschelt and McGann, 1997). Moderate parties, in turn can counter such attemps by somewhat moving towards the extreme end of the opinion space themselves, thereby targeting the voters in between.

The model deliberately refrains from answering whether voters' positions are epistemic or preferential. Parties, on the other hand, try to maximise their number of supporter by using emotional appeals in different ways (or not at all). Hence, this component belongs to a market view on DDM. With regards to party behaviour, the paper therefore discusses political communcation as an extrinsically fair procedure among competing rational actors: Emotional appeals are an extrinsically influential component of the DDM process. The process of how political representation comes about must be fair in order to be normatively justifiable. Parties compete for shares of voters, and are hence utility-driven.

While research regarding responses to emotional appeals is fairly advanced on an individual level, the implications of these findings for system dynamics are unclear due to a lack of empirical observability. Brader (2006), for instance, studies the impact of emotional appeals in a laboratory setting, and hence only looks at an influence on isolated individuals. Other studies, such as Marcus et al. (2000), only look at the correlation between emotional state and political behiavour in a static way. However, capturing both aspects at the same time seems to be practically impossible. As I argue in paper 4, individual behaviour must be 'added up' by taking the compliexities of opinion dynamics into account. My paper provides a framework that shows how affective responses play out for larger groups, and what this implies for the political context. From a methodological perspective, paper 4 therefore contributes by translating empirical psychological research on the individual level to the macro level of political opinion formation. This makes psychological research available for application in the political context.

By studying how political parties can strategically influence people's opinions through the use of emotional appeals, one learns about a crucial feature of political discourse. Understanding its impact is important for designing a political system. This can be essential for avoiding distortions in representative decision making procedures. As the model results show, the democratic process is somewhat resilient against the influence of emotional appeals. Still, countermeasures to balance such attempts need to be available to moderate groups. This knowledge is highly relevant for the design and assessment of mechanisms of political representation.

8 Conclusion

In the first part of this framework chapter, I discussed various approaches to justifying DDM. Central to these approaches is the distinction between the market and the forum view. Further distinctions refer to the philsophical basis of different value judgments. These considerations help to place my contribution in the literature and elaborate the larger context for the individual arguments of each paper.

By discussing crucial problems of market and forum-view justifications, I identify the most important aspects of both fields alike. As a result, many of the aspects from these sections are taken up in the individual papers. For example, paper 3 discusses the reconcilability of deliberation and voting, making use of the abstract considerations of section 4. Papers 3 and 4, in turn, incorporate features from the discussion about deliberation under social influences and rational pitfalls, and thus provide models that are based on a more empirically informed account of deliberation, which I advocate in section 6. Papers 1 focusses on the aspect of manipulabilty in the Adjusted Winner procedure. Since manipulability has been identified as centrally relevant in the discussion of DDM under the market view (see section 3), paper 1 constitutes a case study of the practical consequences of manipulation for a specific DDM mechanism. Paper 2 picks up the discussion of fairness in the market view, and gives a detailed argument for the necessity of a broader consideration of environmental factors for a substantial evaluation of fairness in bargaining. In summary, the first part of this chapter provides a broader context for the specific discussions in the individual papers. It also constitutes a necessary prerequisite for identifying the right kinds of questions and important model components, and also enables an informed interpretation of the papers' results.

Many debates in social choice theory and in the discussion of wisdom of crowds suffer from a lack of empirical data with regards to individual preferences, beliefs, and other characteristics of the situation. The use of ABM provides a useful vehicle for overcoming this obstacle, as computational methods allow for testing all kinds of hypothetical scenarios (see section 6). The methodology of ABM also allows to account for complex system dynamics with regards to belief-updating processes (as in papers 2, 3 and 4). This dynamic component is lacking in standard equilibrium models as well as in statistical approaches alike, as they are naturally restricted to a static perspective. The applications in my papers show how the use of ABM helps to overcome these shortfalls in the study of DDM processes.

On a theoretical level, standard models of DDM frequently suffer from a descriptive inaccuracy with regards to individual behaviour. Knowledge on

individual behavioural regularities exists (Gigerenzer and Gaissmaier, 2011; Kahneman, 2011), yet has only rarely been employed to inform aggregate models of social interaction. Due to the complexity of social dynamics, aggregation of individual behaviour is far from trivial. My contribution aims at closing this gap, as most vividly illustrated by paper 4: It accounts for the macro-level implications of individual affective behaviour in political opinion formation. Similarly, paper 2 assesses the impact of heuristic decision making on the distribution of incomes, and to what extent such behavioural rules can be considered ecologically rational in the sense of Gigerenzer and Gaissmaier (2011).

With this latter point comes the recognition that rationality plays a central role in these debates. I contribute to that debate, most explicitly by the considerations in paper 2. While rationality can function as a normative standard, it should not serve as a descriptive modelling tool. As I have argued, model credibility requires an appropriate description of individual model components. Decision making models must therefore rest on realistic assumptions regarding individual agent capacities. This becomes particularly important when considering the fact that group decision making is a social process, and hence one's own behaviour must be situated in a game theoretic rather than a decision theoretic context. Still, the rational choice paradigm can be fruitfully employed as a component of philosophical arguments, as in papers 1 and 2.

All four papers of my cumulative thesis contribute to a more thorough understanding of DDM procedures. This is useful for addressing questions of institutional design in decision making bodies. The analysis informs debates about claims to normative values such as fairness generally, but also shows how specific design features are most likely to play out in practice. This also carries implications for scholarly debates about justifications of Democratic systems, because the principles of decision making I discuss here constitute a necessary fundamental building block of democratic systems.[22]

[22]Dahl (2008, p. 130) considers electoral systems as probably the most influential institution in democratic systems. These systems work on the basis of the same decision making procedures that are subsumed under the label of DDM mechanism in this work. Similarly, Lijphart et al. (2007, p. 12) asserts that a discussion of sub-features of majority rule (in his terms: bare-majority rule and broad-majority rule) is crucial for a discussion of power-sharing in democracy. Needless to say, there is a direct parallel between majority requirements as an element in democratic systems, and as an element of DDM principles more broadly, as discussed in this work. In the forum view, Hong and Page (2008) assert wisdom of crowds to be a necessary condition for democracies to function: "[I]nstitutional structures such as democracies and markets rests substantially on the emergence of collective wisdom. Without a general tendency for groups of people to make reasonable appraisals and decisions, democracy would be doomed." (Hong and Page, 2008, p. 2).

Crucially, these contributions do not consist of a general, overarching assessment of democratic mechanisms. Instead, I advocate that the value of DDM procedures – at least for substantive, normative accounts – should be considered extrinsic to democratic procedures, and therefore requires a context-specific discussion of DDM. Regardless of whether one subscribes to a market or forum perspective, and no matter if one's argument appeals to the outcomes of a process or the process itself: As soon as claims about the value of DDM go beyond being definitorial, it is essential to envisage decison problems and decision processes not as isolated entities, but as embedded in a context of complex individual interactions and constraints. I derive the necessity of this view in sections 3 and 4, and show how the ABM methodology is equipped to face this task in section 6. This maxim carries through all four papers of this thesis.

Such assessments of specific mechanisms are crucially relevant for practice. Take, for instance, public opinion formation on issues like environmental protection. Should one insist on public referenda, follow expert committee judgments or leave it up to the standard political representation process? Experts might disagree, and we are left with the need for a group decision once again. Votes in a referendum may be based on poorly informed individual judgments as a result of malfunctioning public deliberation and group polarisation. In the political arena, strategic incentives may overrule the force of the better argument and even incentivise actors to acively spread doubt on core issues. In choosing a mechanism in such a situation, one obviously aims for the lesser of many evils, making trade-offs along the way. Yet, not evils are alike, and a careful context-based analysis, as attempted in this work, can help to minimize the pathologies of public deliberation or voting schemes. Political science in general and my thesis in particular should not claim to give definite answers to these questions. Instead, the goal of such analyses should be to provide a firm basis of understanding of complex social processes generally, and DDM in specific. My contribution aspires to provide such a better understanding with regards to the Adjusted Winner procedure, unregulated bargaining, combined deliberation and voting, as well as democratic competition under the impact of emotional appeals.

Hence, also for the forum view, the principles underlying group decision making in epistemic problems are the same principles on which also accounts for the epistemic value of Democracy are based.

9 Bibliography

Anderson, E. (2009). Democracy: Instrumental vs. non-instrumental value. In Christiano, T. and Christman, J., editors, *Contemporary Debates in Political Philosophy*, pages 213–227. Wiley-Blackwell.

Anderson, L. R. and Holt, C. A. (1997). Information cascades in the laboratory. *The American economic review*, pages 847–862.

Arneson, R. J. (2003). Defending the purely instrumental account of democratic legitimacy. *Journal of Political Philosophy*, 11(1):122–132.

Arrow, K. J. (1963). *Social choice and individual values*, volume 2. Yale university press.

Axelrod, R. M. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton University Press.

Bächtiger, A., Niemeyer, S., Neblo, M., Steenbergen, M. R., and Steiner, J. (2010). Disentangling diversity in deliberative democracy: Competing theories, their blind spots and complementarities. *Journal of Political Philosophy*, 18(1):32–63.

Baltag, A., Christoff, Z., Hansen, J. U., and Smets, S. (2013). Logical models of informational cascades. *Studies in Logic*, 47:405–432.

Bandini, S., Manzoni, S., and Simone, C. (2002). Heterogeneous agents situated in heterogeneous spaces. *Applied Artificial Intelligence*, 16(9-10):831–852.

Beetham, D. (1999). *Democracy and human rights*. Polity.

Beitz, C. R. (1989). *Political equality: An essay in democratic theory*. Princeton University Press.

Bellamy, R. (2007). *Political Constitutionalism: A Republican Defence of the Constitutionality of Democracy*. Cambridge University Press.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026.

Black, D. (1948). On the rationale of group decision-making. *Journal of political economy*, 56(1):23–34.

Boero, R. and Squazzoni, F. (2005). Does empirical embeddedness matter? methodological issues on agent-based models for analytical social science. *Journal of Artificial Societies and Social Simulation*, 8(4).

Bohman, J. (1997). *Deliberative democracy: Essays on reason and politics*. MIT press.

Bovens, L. and Rabinowicz, W. (2004). Voting procedures for complex collective decisions. an epistemic perspective. *Ratio Juris*, 17(2):241–258.

Brader, T. (2006). *Campaigning for hearts and minds: How emotional appeals in political ads work*. University of Chicago Press.

Brams, S. J. and Taylor, A. D. (2000). *The win-win solution: guaranteeing fair shares to everybody*. WW Norton & Company.

Casini, L. and Manzo, G. (2016). Agent-based models and causality: A methodological appraisal. *The IAS Working Paper Series*, 7(80).

Chen, H., De, P., Hu, Y., and Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5):1367–1403.

Christiano, T. (1993). Social choice and democracy. In Copp, D., Hampton, J., and Roemer, J. E., editors, *The idea of democracy*. CUP Archive.

Christiano, T. (2008). *The constitution of equality: Democratic authority and its limits*. Oxford University Press.

Cohen, J. (1986). An epistemic conception of democracy. *Ethics*, 97(1):26–38.

Cohen, J. (2007). Deliberative democracy. In Rosenberg, S. W., editor, *Deliberation, Participation and Democracy: Can the People Govern?*, pages 219–236. Palgrave Macmillan UK, London.

Coleman, J. (1990). Foundations of social theory. *Cambridge, MA: Belknap*.

Coleman, J. and Ferejohn, J. (1986). Democracy and social choice. *Ethics*, 97(1):6–25.

Condorcet, N. d. (2014[1785]). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Cambridge University Press.

Copp, D., Hampton, J., and Roemer, J. E. (1993). *The idea of democracy*. CUP Archive.

Dahl, R. A. (1989). *Democracy and its Critics*. Yale University Press.

Dahl, R. A. (2008). *On democracy*. Yale University Press.

De Mesquita, B. B., Cherif, F. M., Downs, G. W., and Smith, A. (2005). Thinking inside the box: A closer look at democracy and human rights. *International Studies Quarterly*, 49(3):439–458.

Dietrich, F. (2008). The premises of condorcet's jury theorem are not simultaneously justified. *Episteme*, 5(1):56–73.

Dietrich, F. and List, C. (2004). A model of jury decisions where all jurors have the same evidence. *Synthese*, 142(2):175–202.

Dietrich, F. and Spiekermann, K. (2013). Epistemic democracy with defensible premises. *Economics & Philosophy*, 29(1):87–120.

Doran, J. (2006). Agent design for agent-based modelling. In Billari, F. C., Fent, T., Prskawetz, A., and Scheffran, J., editors, *Agent-based computational modelling: applications in demography, social, economic and environmental sciences*. Taylor & Francis.

Dorsey, D. (2012). Can instrumental value be intrinsic? *Pacific Philosophical Quarterly*, 93(2):137–157.

Dworkin, R. (1986). *Law's empire*. Harvard University Press.

Elster, J. (1986). The market and the forum: Three varieties of political theory. In Jon Elster, A. H., editor, *Foundations of social choice theory*. New York: Cambridge University Press.

Elster, J. (1998). *Deliberative democracy*, volume 1. Cambridge University Press.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, 4(5):41–60.

Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modeling*. Princeton University Press.

Estlund, D. (1993). Making truth safe for democracy. In D. Copp, J. Hampton, J. E. R., editor, *The idea of democracy*, pages 71–100. Cambridge University Press.

Estlund, D. (1997). *Beyond fairness and deliberation: The epistemic dimension of democratic authority*. MIT Press Cambridge, MA.

Estlund, D. M. (2009). *Democratic authority: A philosophical framework*. Princeton University Press.

Fearon, J. D. (1998). Deliberation as discussion: New directions for democratic reform. In Elster, J., editor, *Deliberative democracy*. Cambridge University Press.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Galeazzi, P. and Franke, M. (2017). Smart representations: Rationality and evolution in a richer environment. *Philosophy of Science*, 84(3):544–573.

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451.

Gibbard, A. (1973). Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, pages 587–601.

Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, 62:451–482.

Gigone, D. and Hastie, R. (1993). The common knowledge effect: Information sharing and group judgment. *Journal of Personality and social Psychology*, 65(5):959.

Gilbert, N. (2007). *Agent-Based Models (Quantitative Applications in the Social Sciences)*. SAGE Publications, Inc.

Gilbert, N. (2008). *Agent-based models*. Sage.

Gilbert, N. and Terna, P. (2000). How to build and use agent-based models in social science. *Mind & Society*, 1(1):57–72.

Gleditsch, N. P. (1992). Democracy and peace. *Journal of Peace Research*, 29(4):369–376.

Goldman, A. and Whitcomb, D. (2011). *Social epistemology: essential readings*. Oxford University Press.

Golub, B. and Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149.

Goodin, R. E. (2008). *Innovating democracy: Democratic theory and practice after the deliberative turn*. Oxford University Press.

Goodin, R. E. (2009). Rationalising discursive anomalies. *Theoria*, 56(119):1–13.

Griffin, C. G. (2003). Democracy as a non–instrumentally just procedure. *Journal of political philosophy*, 11(1):111–121.

Grofman, B. and Feld, S. L. (1988). Rousseau's general will: a condorcetian perspective. *American Political Science Review*, 82(2):567–576.

Grofman, B., Owen, G., and Feld, S. L. (1983). Thirteen theorems in search of the truth. *Theory and Decision*, 15(3):261–278.

Habermas, J. (1970). Towards a theory of communicative competence. *Inquiry*, 13(1-4):360–375.

Habermas, J. (1983). *Moralbewußtsein und kommunikatives Handeln*, volume 422. Suhrkamp Frankfurt.

Harsanyi, J. C., Selten, R., et al. (1988). *A general theory of equilibrium selection in games*. MIT Press.

Helbing, D. and Lämmer, S. (2008). Managing complexity: An introduction. *Managing complexity: Insights, concepts, applications*, pages 1–16.

Helliwell, J. F. (1994). Empirical linkages between democracy and economic growth. *British journal of political science*, 24(2):225–248.

Hillinger, C. (2005). The case for utilitarian voting. *Homo Oeconomicus*, 23:295–321.

Holland, J. H. (2000). *Emergence: From chaos to order*. Oxford University Press.

Hong, L. and Page, S. E. (2008). Some microfoundations of collective wisdom. *Collective Wisdom*, pages 56–71.

Hyland, J. L. (1995). *Democratic theory: the philosophical foundations*. Manchester University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kelly, T. (2011). Peer disagreement and higher order evidence. In Goldman, A. and Whitcomb, D., editors, *Social epistemology: Essential readings*, pages 183–217. Oxford University Press.

Kirman, A. P. (1992). Whom or what does the representative individual represent? *The Journal of Economic Perspectives*, 6(2):117–136.

Kitschelt, H. and McGann, A. J. (1997). *The radical right in Western Europe: A comparative analysis*. University of Michigan Press.

Kittur, A. and Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46. ACM.

Kornhauser, L. A. and Sager, L. G. (2004). The many as one: Integrity and group choice in paradoxical cases. *Philosophy & public affairs*, 32(3):249–276.

Korsgaard, C. M. (1983). Two distinctions in goodness. *The Philosophical Review*, 92(2):169–195.

Kruglanski, A. W. and Mayseless, O. (1987). Motivational effects in the social comparison of opinions. *Journal of Personality and Social Psychology*, 53(5):834.

Lackey, J. (2011). Acquiring knowledge from others. In Goldman, A. and Whitcomb, D., editors, *Social epistemology: Essential readings*, pages 71–91. Oxford University Press.

Ladha, K. K. (1992). The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, pages 617–634.

Landwehr, C. (2005). Rational choice, deliberative democracy and preference transformation. *Studies in Social and Political Thought*, 11:40–68.

Lauth, H.-J. (2004). *Demokratie und Demokratiemessung*. Springer.

Lemieux, P. (2003). Following the herd. *Regulation*, 26(4):16–21.

Lijphart, A. et al. (2007). *Thinking about democracy: power sharing and majority rule in theory and practice*. Routledge.

List, C. (2006). The discursive dilemma and public reason. *Ethics*, 116(2):362–402.

List, C. and Goodin, R. E. (2001). Epistemic democracy: generalizing the condorcet jury theorem. *Journal of Political Philosophy*, 9(3):277–306.

List, C. and Pettit, P. (2002). Aggregating sets of judgments: An impossibility result. *Economics & Philosophy*, 18(1):89–110.

List, C. and Pettit, P. (2005). On the many as one: a reply to kornhauser and sager. *Philosophy & public affairs*, 33(4):377–390.

Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025.

Lumer, C. (1997). Habermas' diskursethik. *Zeitschrift für philosophische Forschung*, 51:42–64.

Lyon, A. and Pacuit, E. (2013). The wisdom of crowds: Methods of human judgement aggregation. In *Handbook of human computation*, pages 599–614. Springer.

Macal, C. M. and North, M. J. (2005). Tutorial on agent-based modeling and simulation. In *Simulation Conference, 2005 Proceedings of the Winter*, pages 14–pp. IEEE.

Mackie, G. (2009). Astroturfing infotopia. *Theoria*, 56(119):30–56.

Macy, M. W. and Willer, R. (2002). From factors to factors: computational sociology and agent-based modeling. *Annual review of sociology*, 28(1):143–166.

Marcus, G. E., Neuman, W. R., and MacKuen, M. (2000). *Affective intelligence and political judgment*. University of Chicago Press.

Marx, J. and Waas, J. (forthcoming). Gut und günstig? über den wert von demokratie und kapitalismus. In *Jahrbuch Normative und institutionelle Grundfragen der Ökonomik*. Metropolis Verlag.

Mill, J. S. (1863). *On Liberty*. Boston: Tickner and Fields.

Miller, J. H. and Page, S. E. (2009). *Complex adaptive systems: An introduction to computational models of social life*. Princeton university press.

Nash, J. F. (1950). The bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 155–162.

Nash, J. F. (1953). Two-person cooperative games. *Econometrica: Journal of the Econometric Society*, pages 128–140.

Ostrom, E. (2015). *Governing the commons*. Cambridge university press.

Perloff, R. M. (1993). *The dynamics of persuasion*. Lawrence Erlbaum Associates, Inc.

Peter, F. (2007). Democratic legitimacy and proceduralist social epistemology. *politics, philosophy & economics*, 6(3):329–353.

Pettit, P. (2001). Deliberative democracy and the discursive dilemma. *Philosophical Issues*, 11(1):268–299.

Przeworski, A. (2004). Democracy and economic development. In Mansfield, E. D. and Sisson, R., editors, *The evolution of political knowledge. democracy, autonomy, and conflict in comparative and international politics*, pages 300–324. Ohio State University Press.

Railsback, S. F. and Grimm, V. (2011). *Agent-based and individual-based modeling: a practical introduction*. Princeton university press.

Rawls, J. (1957). Justice as fairness. *The Journal of Philosophy*, 54(22):653–662.

Rawls, J. (1997). The idea of public reason revisited. *The University of Chicago Law Review*, 64(3):765–807.

Ray, J. L. (1998). Does democracy cause peace? *Annual Review of Political Science*, 1(1):27–46.

Riker, W. H. (1982). *Liberalism against populism: A confrontation between the theory of democracy and the theory of social choice*. WH Freeman.

Rønnow-Rasmussen, T. (2002). Instrumental values–strong and weak. *Ethical Theory and Moral Practice*, 5(1):23–43.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109.

Sanders, L. M. (1997). Against deliberation. *Political Theory*, 25(3):347–376.

Satterthwaite, M. A. (1975). Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2):187–217.

Scheller, S. (2016). *Simulating Bargaining Processes with Agent-based Modelling*. Tectum Verlag Marburg.

Schweller, R. L. (2010). *Unanswered threats: Political constraints on the balance of power*. Princeton University Press.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118.

Sirowy, L. and Inkeles, A. (1990). The effects of democracy on economic growth and inequality: A review. *Studies in Comparative International Development*, 25(1):126–157.

Smith, E. R. and Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and social psychology review*, 11(1):87–104.

Somin, I. (2010). Deliberative democracy and political ignorance. *Critical Review*, 22(2-3):253–279.

Stasser, G. and Titus, W. (2003). Hidden profiles: A brief history. *Psychological Inquiry*, 14(3-4):304–313.

Sunstein, C. R. (1994). *Political conflict and legal agreement*. Tanner Lectures on Human Values, Harvard.

Sunstein, C. R. (2002). The law of group polarization. *Journal of political philosophy*, 10(2):175–195.

Sunstein, C. R. (2006). *Infotopia: How many minds produce knowledge*. Oxford University Press.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Van Hees, M. (2007). The limits of epistemic democracy. *Social Choice and Welfare*, 28(4):649–666.

Verba, S. (2006). Fairness, equality, and democracy: Three big words. *Social Research*, 73(2):499–540.

Vinokur, A. and Burstein, E. (1974). Effects of partially shared persuasive arguments on group-induced shifts: A group-problem-solving approach. *Journal of Personality and Social Psychology*, 29(3):305.

Paper 1

Mitigating the Problem of Manipulation in the 'Adjusted Winner' Procedure

This is article was published in the *Jahrbuch für Handlungs- und Entscheidungstheorie* following peer review.

The suggested citation is:

Scheller, Simon (2017). *Mitigating the Problem of Manipulation in the 'Adjusted Winner' Procedure.* In: Linhart E., Debus M., Kittel B. (eds) Jahrbuch für Handlungs- und Entscheidungstheorie. Jahrbuch für Handlungs- und Entscheidungstheorie. Springer VS, Wiesbaden

This version is also available online at:
https://link.springer.com/chapter/10.1007%2F978-3-658-16714-1_5

Mitigating the Problem of Manipulation in the 'Adjusted Winner' Procedure

Simon Scheller

Abstract

The 'Adjusted Winner' procedure (AW) is a mechanism to reach fair agreements in bargaining situations over a fixed set of objects. A major shortfall of AW for both mediators and participants is that it relies on participants' honesty, which makes it open for manipulation. In a general model of AW for two objects and continuous manipulation strategies, I show that (a) manipulation is always risky since potential losses are always larger than potential gains; and (b) there exists an equilibrium of symmetric manipulation and with equal threat of potential losses that leads to exactly the same outcome like truthful behavior. These findings imply that the problem of manipulation in AW is mitigated.

Keywords

Adjusted winner · Manipulation · Bargaining · Negotiation · Fairness

1 Introduction

Bargaining is a ubiquitous feature of social interaction. Politicians negotiate agreements frequently—be it the post-election bargaining over offices or the settlement of international disputes. In people's daily lives, bargaining and conflict

S. Scheller (✉)
Bamberg Graduate School of Social Sciences, Otto-Friedrich-Universität Bamberg, Feldkirchenstr. 21, 96047, Bamberg, Deutschland
E-Mail: simon.scheller@uni-bamberg.de

E. Linhart et al. (Hrsg.), *Jahrbuch für Handlungs- und Entscheidungstheorie,*
Jahrbuch für Handlungs- und Entscheidungstheorie,
DOI 10.1007/978-3-658-16714-1_5

play an equally important role: Take for example haggling at a car dealership or pay-raise negotiations with an employer.

A central issue in bargaining is the question of fairness. Fairness in general has stirred theoretical-philosophical enquiry across many disciplines; from Plato's formula 'that each should get what he deserves' to Rawls' (1985) 'Justice as Fairness'. In addition, economists like Nash (1950) tried to find a rational and, as he claimed, fair solution in bargaining situations (Nash 1950, p. 158).

The question of fairness in bargaining is mostly asked from an *outside* point of view, which asks: Given the characteristics of the situation and the players, what can be considered a just distribution according to external principles? In contrast, individuals *within* such situations are often assumed to only care about benefitting individually. While it seems that these questions are separate from each other, one main point of this paper is to argue that—especially for practical purposes—they should be thought about together.

Evidence from psychological experiments suggests that some people exhibit a preference for fairness. As outlined for example by Fehr and Schmidt (1999), a small number of people is willing to sacrifice their own gains for the sake of the gains of others in simple ultimatum bargaining experiments. The benefit of posing the additional question of incentive compatibility is that, instead of trusting people's adherence to fair solutions, distributional mechanisms in which individual incentives lead to fair outcomes can be found (for another example, see Sauermann and Beckmann 2017, in this book). I discuss this problem of incentive-proof fair division for the 'Adjusted Winner' (AW) procedure by Brams and Taylor (1999) because AW is a practically interesting fair division procedure and a vivid example of the described dilemma.

In part 2, I describe how the AW procedure works, why it can be said to lead to a fair outcome and give an example of a practical application of the method. I go on to argue that, while practical issues limit the scope of AW's applicability, the problem of manipulation poses a more severe and even fundamental threat to it: If players manipulate, the fairness properties of the AW solution can no longer be guaranteed.

The main contribution of this paper is the description of a model that envisages AW as a strategic game with continuous manipulation strategies for the case of distributing two objects (part 3). Through the introduction of continuous strategies, the model in this paper fundamentally diverges from previous approaches, such as Schüssler's (2007) who discusses a situation with a binary strategy choice. The model yields two remarkable findings that mitigate the problem of manipulation in AW. First, the model shows that the maximal potential gains from manipulation are always smaller than the minimal losses in case manipula-

tion fails. This implies that a manipulative strategy is dangerous for a manipulating player if she is not entirely sure about her opponent's valuation or strategy. Second, if both players simultaneously announce their manipulated preferences, a Nash equilibrium exists which results in the same payoffs as in the case where both players announce their valuations truthfully. This specific Nash equilibrium is an appealing solution because it is symmetric and both players face the same potential loss through deviation from their strategy. Furthermore, as long as both manipulate equally strongly and do not change the initial distribution of objects, the appealing properties of the AW solution are preserved. This suggests that the problem of manipulation in AW is mitigated. Section 4 concludes.

2 The AW Procedure

2.1 How Does AW Work?

AW can be employed to allocate a fixed number of objects between two[1] players fairly. The objects at stake are assumed to be arbitrarily divisible and linear in their utility when divided. Also, the utility of having one object is independent from having any other objects. These requirements are admittedly strong and lead to rather strict limitations for AW, which will be further discussed in light of the example of the Camp David negotiations below.

AW proceeds in three steps. The first step for the players is to assign a total of 100 points to the objects at stake, whereas their allocation of points must reflect their preferences about these objects. Through that step, AW surveys and normalizes the players' preferences in an easy and understandable way. However, note that this step requires a cardinal interpretation of utility: If a solely ordinal utility scale was assumed, any allocation of points that puts the objects in the same order would be indistinguishable.

In the second step, the distribution of the objects takes place: Each object goes to the player who gave the object more points in the first step. In the case where an object has received equal points from both players, the object goes to the player who received fewer points so far.

In a last step, the gains from step two are 'adjusted': The object which is most similar in valuation is partly transferred from the 'richer' to the 'poorer' player

[1]Note that the generalization to n players has consequences for AW. The procedure becomes more complicated and loses some of its appealing properties.

Table 1 An exemplary point allocation for AW

Object	Player 1	Player 2
A	50	40
B	20	30
C	15	10
D	10	10
E	5	10

until both are in the possession of (shares of) objects to which they have assigned the same amount of points.

The term 'most similar in valuation' refers to the ratio of the players' point allocations to the same objects o, i.e. $r_o = \frac{v_o^{rich}}{v_o^{poor}}$. The closer this ration is to one, the more similar in valuation is the object for the two players. Of course, only objects with $v_o^{rich} > v_o^{poor}$ can be redistributed. If one object's complete redistribution does not suffice to equal out the player's points, the procedure continues analogously with the good which is now the most similar in valuation and still belonging to the richer player. In summary, the three steps are as follows:

Step 1: Both players allocate 100 points truthfully to the objects at stake
Step 2: Each object goes to the player who allocated more points to it. Objects with equal points go to the player who received fewer points in total so far
Step 3: The object for which the valuation is most similar is partially redistributed so that both receive objects worth the same amount of points

A simple example illustrates the steps of AW (see Brams and Taylor 1999, p. 72). Consider two players facing the task of distributing five not further specified objects A, B, C, D and E. In step 1, the players allocate their 100 points as shown in Table 1. The higher valuation for each object is underlined. In step 2, player 1 receives objects A and C, while player 2 receives B and E. Object D also goes to player 2, because she has only received objects worth 40 points to her, while player 1 has received 65 points.

Step 3: Player 1 is richer (65 points vs. 50 points so far). The object closest in valuation that has been given to player 1 is A with a ratio of $r_A = 1.2$, while for object C, $r_C = 1.5$. Therefore, A must be redistributed so that both players have objects worth equal points to them afterwards.

To calculate how much of A must change hands, let x be the percentage of object A that will be transferred from player 1 to player 2. The condition that player 1 and 2 must have objects of the same point-value after the transfer translates into the following equation:

$$u_1(C) + (1 - x) \cdot u_1(A) = u_2(B\&D\&E) + x \cdot u_2(A)$$

This leads to the following equation via plugging in the actual numbers:

$$15 + (1 - x) \cdot 50 = 50 + x \cdot 40.$$

Solving for x yields $x = 1/6$. This means that player 1 has to give $1/6$ of A to player 2. The final allocation is: Object C and $5/6$ of A go to player 1; B, D, E and $1/6$ of A go to player 2. Each player therefore receives objects worth 56.67 points to her.

2.2 Why Should the Outcome of an AW Process Be Considered Fair?

Raith (2000) shows that AW yields the Kalai-Smorodinsky Bargaining Solution (KSBS). The KSBS is the only solution that fulfills a certain set of axiomatic requirements (see Kalai and Smorodinsky 1975). That AW or the KSBS carries a central notion of fairness is expressed in the following axioms.

The *symmetry* axiom demands that indistinguishable players receive the same payoffs. Or put shortly: Equals should be treated equally. The *monotonicity* axiom accounts for how to treat players with different preconditions. It requires that a player should not be worse off if the space of feasible solutions is increased in her favor, or intuitively: 'new options for a player should never be a disadvantage'.

The KSBS can be derived as the unique solution when *symmetry*, *monotonicity* as well as *Pareto-efficiency* and *positive linear transformability* of utility functions are required as further axioms. In the KSBS, each player has her best outcome satisfied to the same extent, and it is the solution that benefits each player the most.

Since AW yields the only solution that embodies these axioms of fairness all at the same time, it can be considered a fair mechanism that results in a fair outcome. Certainly, one can disagree with the particular notion of fairness employed here and how it is captured in the axioms. Nonetheless, it is a reasonable and analytically sound argument that formalizes those principles and gives a clear derivation of their implication.

Those abstract properties translate into the practical merits of envy-freeness, efficiency and equitability (Brams and Taylor 1999, p. 69). AW's solution is *envy-free* because in the resulting allocation, neither player would want to exchange her final bundle of objects with the other player. It is *efficient* because no Pareto-improvements are possible in the final allocation. It is *equitable* (or fair) in the sense of the KSBS, that each player's optimal preference is realized to the same degree. Based on this theoretical argument, it is reasonable to adhere to the AW solution as a fair outcome.

2.3 An Example: AW and the Camp David Accords

Brams and Taylor (1999, p. 69) also argue that AW can be performed in a relatively simple way. While this may be true for the steps of the mechanism itself, the range of applicability is a crucial challenge for AW. Without doubt, there are severe limitations that come along with its requirements. Divisibility, linear and independent utility are met only by very few objects. However, even though this may render AW inapplicable in a variety of situations, there are still cases where those requirements are at least reasonable approximations of reality.

The example of the negotiations between Egypt and Israel, which took place in Camp David, USA, in 1978 illustrates this. The Camp David negotiations were the conclusion of a long process of peace talks after several violent conflicts between the two countries. After 13 days of negotiations, leaders of both parties finally signed an agreement, later known as the 'Camp David Accords'.

Brams and Taylor (1999, p. 89 ff.) use AW to assess whether the outcome of the Camp David negotiations was fair for both parties. They identify six major issues in the process and reconstruct the Israeli and Egyptian preferences based on expert judgments (see Table 2). Issues 1, 3 and 6 stand for having control over the respective area. Issue 2 stands for *diplomatic recognition* of Israel, which

Table 2 Issues and implied point allocation in the Camp David negotiations

Issue	Israel	Egypt
Sinai	35	55
Diplomatic recognition	10	5
West Bank/Gaza strip	20	10
Linkage	10	5
Palestinian rights	5	20
Jerusalem	20	5

Israel favored and Egypt opposed. Similarly, *Palestinian rights* have been advocated by Egypt but disfavored by Israel. *Linkage* embodies Egypt's claim that the success of the negotiations at hand must be formally linked to the progress of recognition of Palestinian autonomy. In these three latter cases, 'winning' would mean to get one's way in the decision (see Brams and Taylor 1999, p. 91 ff., for a more elaborate description and discussion of the issues).

AW's solution would prescribe that issues 1 and 5 go to Egypt, while issues 2, 3, 4 and 6 are ascribed to Israel. To even out point gains, 1/6 of the issue 'Sinai' must be redistributed to Israel. With the final distribution, each player receives objects worth 66.7 points.

Brams and Taylor argue that the actual outcome after the talks closely resembles the solution that is prescribed by AW. All issues were allocated accordingly; the division of the issue 'Sinai' was accomplished through the following compromise: Israeli military bases and civil settlements were removed, but the Sinai Peninsula was demilitarized and U.S. troops were stationed to monitor the enforcement of the agreement. According to Brams and Taylor (1999, p. 97), this can be envisaged as a redistribution of 1/6 of the issue. Thus, they argue, the analysis using AW allows to qualify the Camp David agreement as fair.

Note that Brams and Taylor employ AW in a manner of a 'hypothetical procedure'. They do not argue that AW has actually been used during the Camp David talks; they only demonstrate that the actual outcome looks like the AW solution.

2.4 General Problems for the Application of AW

There are several practical issues that could be discussed at this point: Are the identified issues really independent of each other? Can the agreement on the Sinai issue be considered a 1/6- split? Could the other issues have been split as well if the procedure had required it? How reliable is the point allocation based on expert judgments?

Certainly, if one wants to apply AW as an actual negotiation tool, it is always questionable if these or similar conditions are fulfilled. Certainly, there will be cases where the issues at stake do not allow the application of AW because of dependencies or non-linearity. Yet, there are cases that sufficiently fulfill the conditions of AW as in the Camp David case. Researchers, therefore, have to carefully check the necessary preconditions before applying AW.

However, there is a more severe problem for the application of AW, namely the problem of manipulation. Given that the involved actors know how the mechanism works, a rational behavior includes the try to shift outcomes in their

favor. Therefore, to convince actors of the usefulness of AW, one should appeal to individual incentives rather than benevolence or a desire for fairness. Instead of a cooperative mechanism, AW then becomes a strategic game between rational actors. AW (like other fair division mechanisms) must be able to work under the assumption of selfish utility maximizers. If actors were cooperative anyways, there would be no need for a dispute settling mechanism.

2.5 Optimal Manipulation in the Camp David Example

Schüssler (2007, p. 290 f.) scrutinizes the Camp David example and assesses strategies of manipulation in the case where the true valuations are known by the other side. An intuitively optimal manipulation strategy, given that the opponent announces truthful valuations, is to win the same issues as in AW, but to win each issue only by a slight margin. The manipulating player can then allocate the 'saved' points to the issues she loses. This leads to either higher shares of redistributed goods to the manipulating player or to a situation in which the manipulating player has to give up less of her goods through redistribution.

For instance, following the above strategy intuition most closely, Israel could announce the point scheme 53, 6, 11, 6, 18, 6 in the Camp David example, as depicted in Table 3. In that case, the first round distribution of AW would remain the same as before. However, now Israel receives issues worth 27 points only (according to the manipulated point allocation), while Egypt still gains 75 first-round points. Through the AW redistribution mechanism, 12/27 of Sinai must be given to Israel now. Israel's payoff (for which its true valuation from the table above is used) is then a satisfaction of interests worth 75 points compared to 66.7 points for truth telling. Egypt's satisfaction would be reduced to 51 points.

Table 3 Truthful and manipulated point allocations in the Camp David example

Issue	Israel (true)	Israel (manipulated)	Egypt (true)
Sinai	35	53	55
Diplomatic recognition	10	6	5
West Bank/Gaza strip	20	11	10
Linkage	10	6	5
Palestinian rights	5	18	20
Jerusalem	20	6	5

Table 4 AW as a strategic game in the Camp David example with binary strategy choices

		Israel	
		Truth-telling	Manipulation
Egypt	Truth-telling	66.7; 66.7	51; 71
	Manipulation	79; 51	34; 34

Schüssler shows that with a similar strategy, Egypt could even gain 79 points and reduce truth-telling Israel's satisfaction to 51 points. Yet, if both parties announce their manipulated valuations, both would receive only 34 points: The efficient allocation of objects would be reversed and each would get those objects she likes less.

Given a binary choice between manipulating (in this particular way) or being truthful, this would leave the players in a 'chicken game': Whoever convinces the other that she will manipulate will force the other to comply by stating her true preferences (since this is the best response to the described manipulation strategy). There are two pure strategy equilibria (one manipulator, one truth-teller) and one mixed-strategy Nash equilibrium. For applications of AW, this would be a problem because it is not clear how players would act, and manipulation must be expected to occur at least with a certain probability. The situation is shown in Table 4 (see also Schüssler 2007, p. 290).

Schüssler's description simplifies the situation with regard to one central point: It models the actors' strategy choices as binary. Both players can either manipulate (in an optimal way, given that the other player is truthful) or be truthful. This reduces the space of options considerably. According to AW, each player can announce any valuation she wants. Hence, also partial manipulation is possible. In the following section, I account for this possibility of continuous manipulation for both players. However, this modeling approach limits itself to a generic case with only two goods. Generalizing those results to the case of n goods is far from straightforward and involves manifold mathematical complications. Thus, the analysis here is limited for the sake of clarity and solvability in order to get first insights before tackling the more complicated n items case.

3 A General Model for Manipulation in AW

This section develops a model of manipulation in AW with continuous strategies for cases with two objects. In the model, both players can choose to announce any valuation for the objects at stake. The model shows that the threat of manipulation is mitigated because manipulation is risky and mutual manipulation can cancel out.

Due to the complexity of the mathematical analysis, all results are derived for the case of two players (1 and 2) with only two objects (A and B). This, of course, has an impact on the generalizability of the findings, which will be discussed afterwards. Further, all the practical requirements for AW will be assumed to be fulfilled. This means that A and B's utility are linear for shares of objects, and the utility from having a share of A is independent from one's share of B (and vice versa). While the discussion of those aspects is also of high relevance, they are more suited for empirical analysis rather than an abstract-mathematical approach as the one taken here; this paper focusses on the threat through manipulation.

3.1 Setup and Payoff Functions

Let $\widetilde{v_i}$ denote player i's true valuation for object A (i.e. the share of points she would truthfully allocate to object A in AW). It is assumed that $0 < \widetilde{v_i} < 1$, what means that both players have at least some interest in each individual good. It follows that player i's valuation for object B is $1 - \widetilde{v_i}$. $\widetilde{v_i}$ is a continuous variable, so that any real fraction of whole points can be allocated. Further, assume $\widetilde{v_1} > (1 - \widetilde{v_1})$ without loss of generality. This is just the convention to choose object A to be the object that player 1 likes more.

Let v_i denote the valuation which player i *announces* for object A. This is the value that affects the distribution from AW. Further, $v_i \in [0, 1] \forall i$, and v_i is also continuous. For a truthful player, the announced valuation is equal to her real valuation, $\widetilde{v_i} = v_i$. Both players can also choose to manipulate by announcing a valuation $v_i \neq \widetilde{v_i}$. Again, it is assumed that all points are allocated, meaning that player i announces $1 - v_i$ for object B.

The first important step is identifying the player's payoff functions. The following section will reconstruct this for player 1, since this is the point of view taken henceforth. Player 2's payoff function can be obtained simply by switching indices.

Player 1's payoff function depends on her true valuation $\widetilde{v_1}$, which is a not further specified random parameter. The payoff function is further determined by both players' strategic announcements v_1 and v_2, which they can choose freely. Two factors must be distinguished for calculating the payoff function:

First, did player 1 announce a higher valuation for object A than player 2 did? This question decides whether player 1 receives object A or object B initially. In mathematical terms, is $v_1 > v_2$? If yes, she will be given object A in step 2 of AW. If not, she gets object B.

The second question is: Which player realized more points in step 2 of AW? This question decides whether player 1 has to give something to player 2, or vice versa. For instance, if $v_1 > v_2$ and therefore player 1 received A and player 2 received B, the question is: Is $v_1 < (1 - v_2)$, or the other way around? The player who received more will have to give up a share of her object in order to equal out points. In total, this results in four cases that stem from these two questions. Figure 1 visualizes these four cases which can be described in colloquial language as follows:

(a) Player 1 receives A and gives parts of A to player 2
(b) Player 1 receives B and receives parts of A from player 2
(c) Player 1 receives A and receives parts of B from player 2
(d) Player 1 receives B and gives parts of B to player 2

Note that even though it is assumed that $\widetilde{v_1} > 1 - \widetilde{v_1}$, this must not necessary hold for v_1. Player 1 can very well chose to announce a valuation below 0.5 for object A, even if her true valuation for A is assumed to be larger than 0.5.

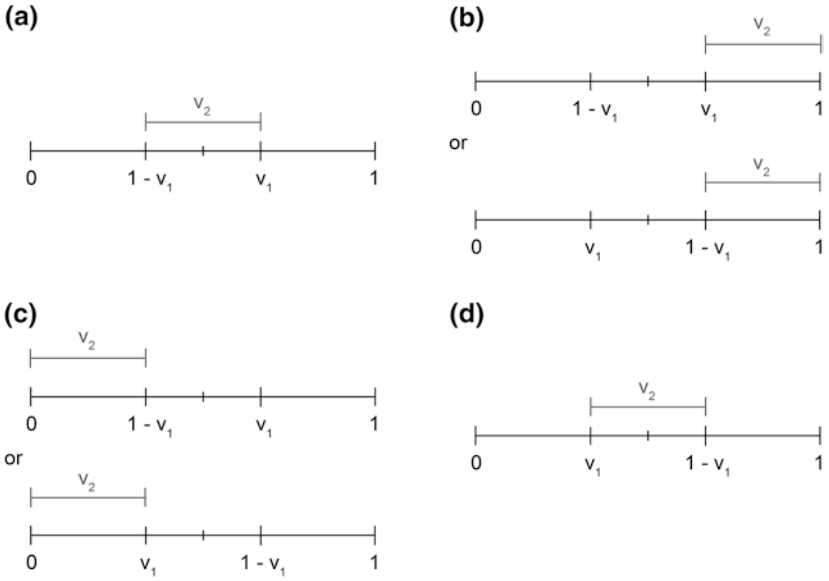


Fig. 1 Visualization of all possible constellations of v_1 and v_2

Consider case (a) where $v_1 > v_2$ and $v_1 > 1 - v_2$. In the initial allocation of objects, player 1 receives object A; player 2 receives object B. Since by assumption $v_1 > 1 - v_2$, player 1 realized more points in step 2, hence x percent of good A must be redistributed from 1 to 2.

The relevant condition for the calculation of x is that both players have equal overall points after redistribution. Player 1 gives up x percent of A, while Player 2 receives x percent of A. Note that for the calculation of x, only the *announced* valuations are relevant. This leads to the following equation:

$$(1 - x) \cdot v_1 = (1 - v_2) + x \cdot v_2$$

$$x = \frac{v_1 + v_2 - 1}{v_1 + v_2} = 1 - \frac{1}{v_1 + v_2}$$

For the calculation of the payoff, the *true* valuations must be employed since these are the values that determine a player's actual payoff. In the final allocation, player 1 holds $(1 - x)$ of good A, therefore her payoff is

$$P_1^{(a)} = (1 - x) \cdot \widetilde{v_1}$$

$$P_1^{(a)} = \left(1 - \left(1 - \frac{1}{v_1 + v_2}\right)\right) \cdot \widetilde{v_1}$$

$$P_1^{(a)} = \frac{\widetilde{v_1}}{v_1 + v_2}$$

The calculations for the other three cases run along similar lines and are given in Appendix A. Table 5 gives the payoff function for player 1 for all four cases. Note that all P_i's are a function of variables v_1 and v_2, hence $P_i(v_1, v_2)$. This is omitted for the sake of brevity and will be denoted only as P_i.

3.2 Optimal Manipulation Against a Truthful Player

Now assume that player 2 will always announce her true valuation, i.e. $v_2 = \widetilde{v_2}$. First, the optimal manipulation strategy for player 1 in this case is discussed. It can be generally shown that player 1's optimal manipulation strategy is to let v_1 approach $\widetilde{v_2}$, thus proving for two objects what Schüssler (2007) described on an intuitive level for more than two objects. However, as the subsequent section shows, this strategy is very dangerous because the maximal gains from optimal

Table 5 Payoff functions for player 1

	$v_1 > v_2$	$v_1 < v_2$
$1 - v_1 < v_2; v_1 > 1 - v_2$	(a) $P_1^{(a)} = \frac{\tilde{v_1}}{v_1 + v_2}$	(b) $P_1^{(b)} = 1 - \frac{\tilde{v_1}}{v_1 + v_2}$
$1 - v_1 > v_2; v_1 < 1 - v_2$	(c) $P_1^{(c)} = 1 - \frac{(1-\tilde{v_1})}{(1-v_1)+(1-v_2)}$	(d) $P_1^{(d)} = \frac{1-\tilde{v_1}}{(1-v_1)+(1-v_2)}$

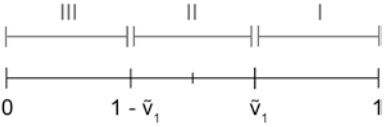


Fig. 2 Three possible constellations of true valuations

manipulation (compared to an honest strategy) is always smaller than the loss player 1 would suffer from only the slightest over-manipulation.

Technical hint: It is implicitly assumed that, if v_1 is exactly on the border of two payoff functions, player 1 can 'choose' which of the two bordering payoff functions is used. The correct notation for this would be for instance $v_1 = \lim_{\varepsilon \to 0} \left(\tilde{v_2} - \varepsilon\right)$ for $\varepsilon > 0$ to show that v_1 approaches $\tilde{v_2}$ from below, so that still the payoff function for the case $v_1 < \tilde{v_2}$ is applicable. This detailed notation is omitted here for the sake of brevity. Originally, AW prescribes that in case of equal points, the object goes to the player with fewer total points, and is then potentially redistributed. The problem of 'choosing a payoff function' does not occur in the original AW-procedure, since payoffs are the same under both functions.

Under the assumption $\tilde{v_1} > (1 - \tilde{v_1})$, three different constellations for $\tilde{v_1}$ and $\tilde{v_2}$ can occur. $\tilde{v_2}$ can be either in region I, II or III in relation to $\tilde{v_1}$, as depicted in Fig. 2. In the following, the optimal manipulation strategy for constellation I will be shown. The same results can be obtained for constellations II and III, for which the calculations are in Appendix B. For this part, assume therefore that $\tilde{v_2}$ lies in area I, and hence $1 - \tilde{v_1} < \tilde{v_1} < \tilde{v_2}$. Depending on player 1's choice of v_1, either payoff function (a), (b), or (d) is applicable. Figure 3 depicts the regions in which the different payoff functions apply.

- Player 1's payoff function is $P_1^{(a)} = \frac{\tilde{v_1}}{v_1 + \tilde{v_2}}$ for $v_1 > \tilde{v_2}$ (since it follows $v_1 > 1 - \tilde{v_2}$) Function (a) strictly decreases with v_1, hence the local maximum is reached for $v_1 = \tilde{v_2}$. The maximum payoff is then $P_1^{(a)max} = \frac{\tilde{v_1}}{2\tilde{v_2}}$.

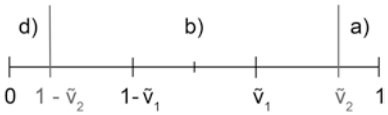


Fig. 3 Applicable payoff function depending on choice of v_1

- Player 1's payoff function is $P_1^{(b)} = 1 - \frac{\tilde{v_1}}{v_1 + \tilde{v_2}}$ for $1 - \tilde{v_2} < v_1 < \tilde{v_2}$ (since it follows $1 - v_1 < \tilde{v_2}$). Payoff function (b) strictly increases with v_1, hence the local maximum is reached for $v_1 = \tilde{v_2}$, which is the largest value of v_1 in (b). The maximum payoff is then $P_1^{(b)max} = 1 - \frac{\tilde{v_1}}{2\tilde{v_2}}$.
- Player 1's payoff function is $P_1^{(d)} = \frac{(1-\tilde{v_1})}{(1-v_1)+(1-\tilde{v_2})}$ for $v_1 < 1 - \tilde{v_2}$ (since it follows $1 - v_1 > \tilde{v_2}$). Payoff function (d) strictly increases with v_1, hence the local maximum is reached for $v_1 = 1 - \tilde{v_2}$, again the largest value of v_1 in (d). The maximum payoff is then $P_1^{(d)max} = (1 - \tilde{v_1})$.

The local maxima can now be compared. Since we are still in case I where $\tilde{v_1} < \tilde{v_2}$, it follows that $P_1^{(b)max} > P_1^{(a)max}$. Therefore, player 1 prefers being in (b) over being in (a) in constellation I. Also, $P_1^{(b)max} > P_1^{(d)max}$, which means that player 1 also prefers (b) over (d).

Looking at the complete payoff function, the global maximum is $P_1^{max} = P_1^{(b)max}$. Therefore, player 1's optimal strategy is to chose $v_1 = \tilde{v_2}$ but still stay in (b) and hence a little below $\tilde{v_2}$ (technically correct: $v_1 = \tilde{v_2} - \varepsilon$ with $\varepsilon \to 0$ and $\varepsilon > 0$). Figure 4 illustrates the payoff function for an example of constellation I with $\tilde{v_1} = 0.6$ and $\tilde{v_2} = 0.7$. Generally (the constellations II and III are shown in Appendix B), the optimal strategy for player 1 is to let v_1 approach $\tilde{v_2}$.

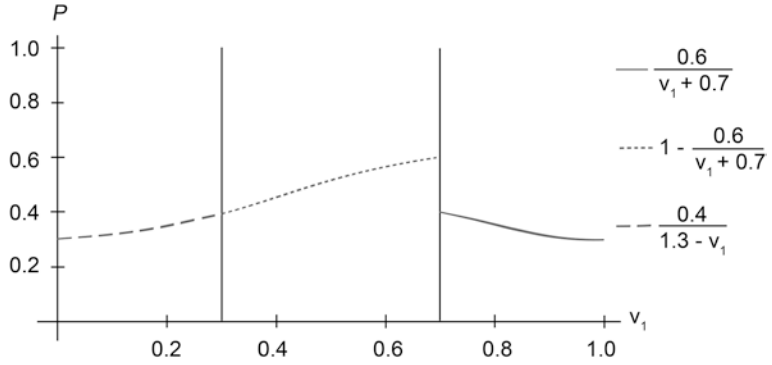


Fig. 4 Exemplary payoff function for constellation I with $\tilde{v_1} = 0.6$ and $\tilde{v_2} = 0.7$

More specifically, if $\widetilde{v_1} > \widetilde{v_2}$, v_1 should remain larger than $\widetilde{v_2}$, and if $\widetilde{v_1} < \widetilde{v_2}$, v_1 should remain smaller than $\widetilde{v_2}$. This is the optimal manipulation strategy proposed by Schüssler (2007).

3.3 Optimal Manipulation and the Risk of Over-Manipulation

In the case above, if player 1's optimal manipulation strategy works, her payoff against the honest player 2 is $P_1^{(b)max} = 1 - \frac{\widetilde{v_1}}{2\widetilde{v_2}}$. However, if she misjudges $\widetilde{v_2}$ and therefore manipulates 'too far', the payoff function switches to $P_1^{(a)}$, and her payoff for failed manipulation becomes $P_1^{fail} < \frac{\widetilde{v_1}}{2\widetilde{v_2}}$ since $P_1^{(a)}$ decreases with v_1. Therefore, $P_1^{fail} = \frac{\widetilde{v_1}}{2\widetilde{v_2}}$ is the most player 1 can expect from failed manipulation. More over-manipulation reduces her payoff further.

Compared to the strategy 'honesty' with $P_1^{honest} = 1 - \frac{\widetilde{v_1}}{\widetilde{v_1}+\widetilde{v_2}}$, player 1 could at most gain $g = P_1^{(b)max} - P_1^{honest}$. On the other hand, if her manipulation fails, she will at least lose $l = P_1^{honest} - P_1^{fail}$. Calculating the difference between maximal potential gains and minimal potential losses results in the term $g - l = P_1^{max} + P_1^{fail} - 2P_1^{honest} = \frac{2\widetilde{v_1}}{\widetilde{v_1}+\widetilde{v_2}} - 1$. It is easy to see that $g - l = \frac{2\widetilde{v_1}}{\widetilde{v_1}+\widetilde{v_2}} - 1 < 0$, since in I, $\widetilde{v_1} < \widetilde{v_2}$.

This is a crucial result because it shows that the manipulation strategy which seeks to maximize player 1's gains through letting v_1 approach $\widetilde{v_2}$ has a higher potential loss through failure than can maximally be gained by successful manipulation. The same result is also found for constellations II and III, as shown in Appendix B.

This result can have a strong impact on a player's incentives to manipulate. If there is uncertainty about one's opponent's valuation, a player should be reluctant to follow the above optimal manipulation strategy. In real bargaining situations, at least some uncertainty can always be expected. Manipulation therefore becomes less attractive.

Even if there is no uncertainty about valuations, a truthful player could protect herself by announcing a randomization strategy, which lets her announce her true valuation plus some error term. This could partly prevent the other player from manipulating, since she would increase the risk to lose more than she could gain. Such a strategy could be employed only if a player can credibly communicate such a strategy or in repeated games.

To extend this type of reasoning, one would have to calculate the optimal strategy for settings of imperfect information. For example, one could assume a uniform (or normal) distribution of $\widetilde{v_2}$ over the interval [0, 1] and check what player

1's optimal strategy would be for a given $\widetilde{v_1}$. However, this would go beyond the scope of this paper. The reasoning presented thus far shall nonetheless hint upon the fact that truth-telling might be a promising candidate for an optimal strategy in certain settings.

3.4 Mutual Manipulation

So far, I assumed player 2 to announce her valuation truthfully. This section considers the situation of two manipulative players, which is in a way the worst case scenario when using AW. Yet, it is probably not only the most realistic assumption but also a crucial test for AW.

Schüssler (2007) considers the case of n objects and two players able to choose binarily between truth-telling and the optimal manipulation strategy. For this setting, he argues that the binary choice between manipulation and truth-telling renders the players in a chicken game. In this section, I argue that this changes once manipulation is characterized in the continuous fashion proposed in this paper. Both players can choose not only whether or not to manipulate; they can also choose *how much* they want to manipulate, i.e. what valuation they want to announce.

The optimal strategy from above was derived only as the best response to a truthful player. The same result would also occur if one player had a first-mover's advantage: Through committing to the above optimal manipulation strategy, she could force the other player to state her true preferences, which is the best response to said strategy.

If the players have to announce their valuations simultaneously, the situation is very similar to Nash's demand game (Nash 1953). The announced valuation is the analog to the announced demand. This is an appropriate way of looking at AW as a strategic game because when valuations are supposed to be announced truthfully, open haggling should not be expected to occur. Mediators should further be able to enforce simultaneous valuation announcement, since one runs into much deeper troubles if this is not the case. From a theoretical point of view, this is also the simplest form of describing a symmetric situation in which neither player has a structural advantage.

To begin with the strategic analysis, consider again constellation I from the previous section, where $\widetilde{v_2} > \widetilde{v_1} > 1 - \widetilde{v_2}$, depicted in Fig. 2. As long as $v_1 < v_2$, the players haggle about how much of object A is redistributed from player 2 to 1. These will be called the 'compatible' cases, because the initial distribution of objects remains efficient. As soon as $v_1 > v_2$, the manipulation strategies fail

because payoff functions switch from (b) to (a) for player 1, and from (a) to (b) for 2. The outcome becomes inefficient.

Formalizing these intuitive arguments, a Nash equilibrium in the valuation-demand game must fulfill the requirement that the two valuations are the same in the limit. For the case when equal valuations are announced, it is assumed that the players can pick who receives which object in step 2 of AW. Then, they will always choose to allocate objects in accordance with their true valuations, meaning that they will choose the better payoff function to apply for each of them. If they did not, both players would be worse off.

Hence every Nash equilibrium must fulfill the condition $v_1 = v_2 = v$. If, for instance in a situation of constellation I, player 1 would lower v_1 in comparison to the Nash equilibrium with $\widetilde{v_1} > v > \widetilde{v_2}$, she would render demands incompatible and decrease her payoff. By increasing v_1 back towards $\widetilde{v_1}$, she would also decrease her payoff since the amount of object A that she has to redistribute to 2 would become larger. Thus, no unilateral deviation from this strategy can be profitable for her, and a beneficial deviation is always possible if $v_1 = v_2 = v$ does not hold.

The main difference to this game and the original Nash demand game is the structure of payoffs. In the original demand game, payoffs are the respective demands themselves for compatible demands, and zero for both players in case of incompatibility. The payoffs here (for constellation I) are

- $P_1^{(b)} = \frac{\widetilde{v_1}}{v_1+v_2}$ and $P_2^{(a)} = 1 - \frac{\widetilde{v_2}}{v_1+v_2}$ for compatible demands with $v_1 \leq v_2$, and
- $P_1^{(a)} = 1 - \frac{\widetilde{v_1}}{v_1+v_2}$ and $P_2^{(b)} = \frac{\widetilde{v_2}}{v_1+v_2}$ for incompatible demands with $v_1 > v_2$.

The incompatibility payoff is not zero, it is not even constant in the AW-manipulation-game. Therefore, the question of which Nash equilibrium will be selected cannot be answered unambiguously.

3.5 The Threat-Equivalent Equilibrium

However, there is one Nash equilibrium with a special appeal. This is the equilibrium where both players face the same loss if demands become incompatible. Call this the *threat-equivalent equilibrium*. In that equilibrium, both players have the same capacity to threaten the other player into behaving compatibly. In every other Nash equilibrium, one player faces a higher potential loss through incompatibility than the other player.

If any other equilibrium were to be chosen, one player could threaten the other and argue as following: 'If you do not reduce your demand, I will render mutual

demands incompatible. You would lose more than me through my move, therefore I urge you to reduce your demand.' She could make this argument exactly up to the point where both could threaten each other with the same potential loss. This is the threat-equivalent equilibrium.

To calculate this equilibrium strategy, the equilibrium payoff of player 1 minus her potential loss at this equilibrium point must be equal to player 2's equilibrium payoff minus 2's potential loss.

$$P_1^{(a)}(v,v) - P_1^{(b)}(v,v) = P_2^{(b)}(v,v) - P_2^{(a)}(v,v)$$

$$\frac{\widetilde{v_1}}{2v} + \left(\frac{\widetilde{v_1}}{2v} - 1\right) = 1 - \frac{\widetilde{v_2}}{2v} - \frac{\widetilde{v_2}}{2v}$$

$$\frac{\widetilde{v_1}}{v} - 1 = 1 - \frac{\widetilde{v_2}}{v}$$

$$v = \frac{\widetilde{v_1} + \widetilde{v_2}}{2}.$$

This solution has a simple graphic interpretation: The point where both face equal potential losses is exactly the point in the middle between the two true valuations: $\frac{\widetilde{v_1}+\widetilde{v_2}}{2}$. Both players shift their valuation the same distance from their true valuation towards the other's true valuation in the threat-equivalent equilibrium.

The very important feature of this equilibrium is that the payoffs for both players are exactly the same as in the case where both players play truthful strategies:

$$P_1^{(b)}\left(\frac{\widetilde{v_1} + \widetilde{v_2}}{2}, \frac{\widetilde{v_1} + \widetilde{v_2}}{2}\right) = 1 - \frac{\widetilde{v_1}}{2\left(\frac{\widetilde{v_1} + \widetilde{v_2}}{2}\right)} = 1 - \frac{\widetilde{v_1}}{\widetilde{v_1} + \widetilde{v_2}} = P_1^{(b)}\left(\widetilde{v_1}, \widetilde{v_2}\right)$$

$$P_2^{(a)}\left(\frac{\widetilde{v_1} + \widetilde{v_2}}{2}, \frac{\widetilde{v_1} + \widetilde{v_2}}{2}\right) = \frac{\widetilde{v_2}}{2\left(\frac{\widetilde{v_1} + \widetilde{v_2}}{2}\right)} = \frac{\widetilde{v_2}}{\widetilde{v_1} + \widetilde{v_2}} = P_2^{(a)}\left(\widetilde{v_1}, \widetilde{v_2}\right).$$

This result is truly remarkable. $P_2^{(a)}\left(\frac{\widetilde{v_1} + \widetilde{v_2}}{2}, \frac{\widetilde{v_1} + \widetilde{v_2}}{2}\right)$ is the payoff under the prescribed threat equivalent manipulation. $P_2^{(a)}\left(\widetilde{v_1}, \widetilde{v_2}\right)$ is the payoff if both are truthful. Those two payoffs are identical what means that if both players manipulate in this way, manipulation does not matter, and the solution preserves all the appealing properties of the ideal AW procedure. In particular, it is efficient and fair under the definition of Kalai and Smorodinsky.

The justification that this equilibrium will be the outcome of the game is not without counterarguments, and it is by no means the claim of this paper that the threat-equivalent equilibrium is the only feasible equilibrium outcome here. Nevertheless, the comparison of losses in that way is a reasonable argument to justify an equilibrium—especially since AW normalizes the maximal payoffs of both players, which means that they are comparable between the players. Thus, the argument of equivalent losses is a solid argument if players have similar risk preferences.

Furthermore, the threat-equivalent equilibrium has the appealing property that it is the symmetric point between the two true valuations. This could be another argument for a player to choose the according strategy, since both would be manipulating equally strong here.

3.6 Symmetric Manipulation

In the threat-equivalent equilibrium, both players manipulate equally strongly *and* they are in a Nash equilibrium. What if that latter characteristic is dropped, and it is only assumed that both players manipulate to the same extent?

To answer this question, first assume that both shift their true valuation t points towards the other's valuation, and demands remain compatible ($t < \frac{\widetilde{v_2} - \widetilde{v_1}}{2}$). As the calculation below shows, the payoffs as under truth telling are preserved. The equality $P_1^{(b)}\left(\widetilde{v_1} + t, \widetilde{v_2} - t\right) = P_1^{(b)}\left(\widetilde{v_1}, \widetilde{v_2}\right)$ means that the payoffs where both manipulate with t are the same as when both announce truthfully. This is true for both players. Thus, as long as demands remain compatible and manipulation is symmetric, manipulation does not matter either.

$$P_1^{(b)}\left(\widetilde{v_1} + t, \widetilde{v_2} - t\right) = 1 - \frac{\widetilde{v_1}}{\widetilde{v_1} + t + \widetilde{v_2} - t} = 1 - \frac{\widetilde{v_1}}{\widetilde{v_1} + \widetilde{v_2}} = P_1^{(b)}\left(\widetilde{v_1}, \widetilde{v_2}\right)$$

$$P_2^{(a)}\left(\widetilde{v_1} + t, \widetilde{v_2} - t\right) = \frac{\widetilde{v_2}}{\widetilde{v_1} + t + \widetilde{v_2} - t} = \frac{\widetilde{v_2}}{\widetilde{v_1} + \widetilde{v_2}} = P_2^{(a)}\left(\widetilde{v_1}, \widetilde{v_2}\right).$$

Now consider the case where both manipulate equally strongly, but too strong to keep demands compatible ($t > \frac{\widetilde{v_2} - \widetilde{v_1}}{2}$) . The obtained payoffs are as follows:

$$P_1^{(a)}\left(\widetilde{v_1} + t, \widetilde{v_2} - t\right) = \frac{\widetilde{v_1}}{\widetilde{v_1} + t + \widetilde{v_2} - t} = \frac{\widetilde{v_1}}{\widetilde{v_1} + \widetilde{v_2}}$$

$$P_2^{(b)}\left(\widetilde{v_1} + t, \widetilde{v_2} - t\right) = 1 - \frac{\widetilde{v_2}}{\widetilde{v_1} + t + \widetilde{v_2} - t} = 1 - \frac{\widetilde{v_2}}{\widetilde{v_1} + \widetilde{v_2}}.$$

Those payoffs are certainly smaller than before. Comparing those payoffs with what they would have gotten under mutual truth-telling (or in the threat-equivalent-equilibrium, or any case where they manipulate equally strong but remain compatible), one obtains:

$$P_1^{(a)}\left(\widetilde{v_1}+t,\widetilde{v_2}-t\right)-P_1^{(b)}\left(\widetilde{v_1},\widetilde{v_2}\right)=\frac{\widetilde{v_1}}{\widetilde{v_1}+\widetilde{v_2}}-\left(1-\frac{\widetilde{v_1}}{\widetilde{v_1}+\widetilde{v_2}}\right)=\frac{2\widetilde{v_1}}{\widetilde{v_1}+\widetilde{v_2}}-1=\frac{\widetilde{v_1}-\widetilde{v_2}}{\widetilde{v_1}+\widetilde{v_2}}$$

$$P_2^{(b)}\left(\widetilde{v_1}+t,\widetilde{v_2}-t\right)-P_2^{(a)}\left(\widetilde{v_1},\widetilde{v_2}\right)=1-\frac{\widetilde{v_2}}{\widetilde{v_1}+\widetilde{v_2}}-\left(\frac{\widetilde{v_2}}{\widetilde{v_1}+\widetilde{v_2}}\right)=1-\frac{2\widetilde{v_2}}{\widetilde{v_1}+\widetilde{v_2}}=\frac{\widetilde{v_1}-\widetilde{v_2}}{\widetilde{v_1}+\widetilde{v_2}}.$$

Since $\widetilde{v_1}<\widetilde{v_2}$, both terms are negative, hence losses occur through over-manipulation (as before). Therefore, losses compared to honesty are the same for both players if both players over-manipulate to the same extent, namely $\frac{\widetilde{v_1}-\widetilde{v_2}}{\widetilde{v_1}+\widetilde{v_2}}$. The same results are also found for the remaining constellations, described in Appendix C.

As a conclusion, effects are symmetric, if both players manipulate equally strongly. If they remain compatible, the payoffs from truthfulness and hence AW's appealing solution is preserved. If manipulation renders payoffs incompatible, both players face equal losses when they manipulate equally strongly.

To summarize the main results from Sect. 3: In the threat-equivalent equilibrium, both players manipulate. Yet, the same payoffs as in the AW-solution and hence all its properties are preserved. Due to the strategic interaction, manipulation cancels out, and the payoffs are exactly the same as under truth-telling. Further, even if players are not in equilibrium but manipulate symmetrically, the AW properties are still preserved as long as manipulation does not change the initial distribution of objects.

Thus, the calculations for this model show that there is quite a range of cases of unproblematic manipulation. There are good arguments why those situations are more likely to occur (symmetry, equal threats, equal losses).

4 Conclusion: Assessing the Problem of Manipulation for AW

The previous part has analyzed manipulation strategies for AW in the two objects case. Even though the problem of manipulation is not negligible in all cases, there are circumstances where the problem is mitigated. First, since manipulation carries the danger of over-manipulation, imperfect information can be beneficial for the applicability of AW. The less knowledge a player has about an opponent's

valuation, the riskier it is for her to manipulate. A player can create such uncertainty for her opponent by randomizing her own valuation, thereby creating the danger of over-manipulation for the opponent.

Second, if both players manipulate, there are still cases where manipulation has no negative effects on the outcome. As I showed, there is an equilibrium with the same payoffs like truthful AW. The same is true for all cases where both manipulate equally strongly as long as the initial allocation of objects does not change.

Thus, AW is not as severely threatened by manipulation as it seems at first sight. Imperfect information, randomization of one's own strategy and mutual strategic manipulation mitigate the problem. Still, further theoretical and practical research is needed to deepen the understanding of manipulative strategies in theory and practice. The approach taken here can serve as a starting point and a baseline approach for further modeling.

What do these findings imply for the AW method and its applicability? The main insight from these findings is that manipulation may not be as ubiquitous as one might expect at first sight. The reason for this is that manipulation performs badly in a cost-benefit-analysis. Whilst this point was already partly implied by Schüssler's model, a new finding in this paper is that manipulation does not matter in all cases where it occurs and that solutions produced by manipulative behavior can still be fair. Thus, the AW mechanism can be seen as resilient against certain kinds of untruthful behavior. This improves AW's applicability to dispute settlements in areas where agents must be assumed to be pure egoists, for instance in the realm of international politics. Admittedly, conflicts of a political importance comparable to the Israel-Egypt case above will probably not be settled by applying the AW procedure any time soon. Yet, even for those cases, AW may provide an interesting perspective on what a solution could look like. For other cases, such as the negotiation of coalition agreements, AW can provide a reasonable guideline, especially under time pressure, even when manipulative behavior most certainly occurs.

The next crucial theoretical step is the generalization of the approach to cases with more than two goods. This complicates matters more than one might imagine at first sight. It is possible that the additional complication of the situation favors truthful behavior as more uncertainty occurs. Unfortunately, analytical results in that direction are hard to obtain, as exemplified by the paper by Aziz et al. (2015): The authors provide some general insights for the n-goods case, most prominently they prove the existence of Nash equilibria under certain conditions. Yet, they are only able to characterize the impact on general welfare of those equilibria, which they show to be at least ¾ of the welfare from the original solution. They do not identify characteristics of those solutions with regards to

how this welfare is split. Thus, the question about fairness in AW with manipulation in the n-goods case remains unanswered.

A solution to those difficulties might be found in the computer simulation of AW. For example, an agent-based model in which manipulating and truthful players compete in multiple issue negotiations could be employed to assess which strategies prove to be most successful.

Acknowledgements For their valuable feedback, I am truly grateful to Rudolf Schüssler, Eric Linhart, Florian Herold, Johannes Marx and two anonymous reviewers. This research is based on parts of a master's thesis in economics at Otto-Friedrich University of Bamberg, but underwent substantial changes and was completely revised for this publication. This work was supported by the Bamberg Graduate School of Social Sciences which is funded by the German Research Foundation (DFG) under the German Excellence Initiative (GSC1024).

Appendix A: Calculation of the Payoff Functions of AW

This section gives the calculations for all cases that can occur for the payoff function of a player in AW with two objects. These results are referred to in Sect. 3.

Case (b)

Relevant constraints: $v_1 < v_2$ and $1 - v_1 < v_2$.

Initial allocation of objects: Player 1 receives object B; Player 2 receives object A.

Redistribution: Since $1 - v_1 < v_2$, redistribution of x percent of A from 2 to 1.

Calculation of x:

$$(1 - v_1) + x \cdot v_1 = (1 - x) \cdot v_2$$

$$x = \frac{v_1 + v_2 - 1}{v_1 + v_2} = 1 - \frac{1}{v_1 + v_2}.$$

Payoff for player 1:

$$P_1^{(b)} = \left(1 - \widetilde{v_1}\right) + x \cdot \widetilde{v_1}$$

$$P_1^{(b)} = \left(1 - \widetilde{v_1}\right) + \left(1 - \frac{1}{v_1 + v_2}\right) \cdot \widetilde{v_1}$$

$$P_1^{(b)} = 1 - \frac{\widetilde{v_1}}{v_1 + v_2}.$$

Case (c)

Relevant constraints: $v_1 > v_2$ and $v_1 < 1 - v_2$.

Initial allocation of objects: Player 1 receives object A; Player 2 receives object B.

Redistribution: Since $v_1 < 1 - v_2$, redistribution of x percent of B from 2 to 1.

Calculation of x:

$$v_1 + x \cdot (1 - v_1) = (1 - x) \cdot (1 - v_2)$$

$$x = \frac{1 - v_1 - v_2}{2 - v_1 - v_2}.$$

Payoff for player 1:

$$P_1^{(c)} = \widetilde{v_1} + x \cdot (1 - \widetilde{v_1})$$

$$P_1^{(c)} = \widetilde{v_1} + \frac{1 - v_1 - v_2}{2 - v_1 - v_2} \cdot (1 - \widetilde{v_1})$$

$$P_1^{(c)} = \widetilde{v_1} \cdot \left(1 - \frac{1 - v_1 - v_2}{2 - v_1 - v_2}\right) + \frac{1 - v_1 - v_2}{2 - v_1 - v_2}$$

$$P_1^{(c)} = \widetilde{v_1} \left(\frac{1}{2 - v_1 - v_2}\right) + \frac{1 - v_1 - v_2}{2 - v_1 - v_2}$$

$$P_1^{(c)} = \frac{\widetilde{v_1} + 1 - v_1 - v_2}{2 - v_1 - v_2} = \frac{\widetilde{v_1} - 1}{2 - v_1 - v_2} + 1 = 1 - \frac{1 - \widetilde{v_1}}{(1 - v_1) + (1 - v_2)}.$$

Case (d)

Relevant constraints: $v_1 < v_2$ and $1 - v_1 > v_2$.

Initial allocation of objects: Player 1 receives object B; Player 2 receives object A.

Redistribution: Since $1 - v_1 > v_2$, redistribution of x percent of B from 1 to 2.

Calculation of x:

$$(1 - x) \cdot (1 - v_1) = v_2 + x \cdot (1 - v_2)$$

$$1 - x - v_1 + x \cdot v_1 = v_2 + x - x \cdot v_2$$

$$1 - v_1 - v_2 = 2x - x \cdot v_1 - x \cdot v_2$$

$$x = \frac{1 - v_1 - v_2}{2 - v_1 - v_2}.$$

Payoff for player 1:

$$P_1^{(c)} = (1 - x) \cdot (1 - \widetilde{v_1})$$

$$P_1^{(c)} = \left(1 - \frac{1 - v_1 - v_2}{2 - v_1 - v_2}\right) \cdot (1 - \widetilde{v_1})$$

$$P_1^{(c)} = \frac{1 - \widetilde{v_1}}{2 - v_1 - v_2} = \frac{1 - \widetilde{v_1}}{(1 - v_1) + (1 - v_2)}.$$

Appendix B: Optimal Manipulation, Potential Gains and Losses

This section completes the calculations for the result that the optimal strategy for player 1 against an honest player 2 is to let v_1 approach $\widetilde{v_2}$ for constellations II and III. Further, the result that the potential gains from optimal manipulation are always smaller than the potential losses from over-manipulation are shown for constellations II and III.

Optimal strategy in Constellation II

$\widetilde{v_2}$ lies in II, hence $1 - \widetilde{v_1} < \widetilde{v_2} < \widetilde{v_1}$. Depending on 1's choice of v_1, either payoff function (a), (b), or (d) is applicable. The payoff functions and the local maxima are the same as in constellation I, but the global maximum is different:

- Local maximum in (a): $P_1^{(a)max} = \frac{\widetilde{v_1}}{2\widetilde{v_2}}$ for $v_1 = \widetilde{v_2}$
- Local maximum in (b): $P_1^{(b)max} = 1 - \frac{\widetilde{v_1}}{2\widetilde{v_2}}$ for $v_1 = \widetilde{v_2}$
- Local maximum in (d): $P_1^{(d)max} = (1 - \widetilde{v_1})$ for $v_1 = 1 - \widetilde{v_2}$

From $1 - \widetilde{v_1} < \widetilde{v_2} < \widetilde{v_1}$ follows
$P_1^{(a)max} > P_1^{(b)max}$ and $P_1^{(a)max} > P_1^{(d)max}$.

The global maximum is $P_1^{max} = P_1^{(a)max}$. Therefore, player 1's optimal strategy is to choose $v_1 = \widetilde{v_2}$ but still a little above $\widetilde{v_2}$ to stay in (a) (technically correct: $v_1 = \widetilde{v_2} + \varepsilon$ with $\varepsilon \to 0$ and $\varepsilon > 0$). Figure 5 illustrates the payoff function for the example $\widetilde{v_1} = 0.8$ and $\widetilde{v_2} = 0.7$.

Gains and losses from manipulation in constellation II

If player 1's optimal manipulation strategy works, her payoff against the honest player 2 is

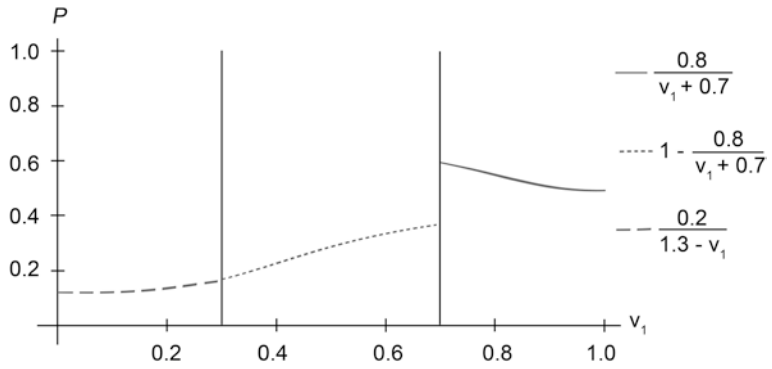


Fig. 5 Exemplary payoff function for constellation II with $\widetilde{v_1} = 0.8$ and $\widetilde{v_2} = 0.7$

$$P_1^{(a)max} = \frac{\widetilde{v_1}}{2\widetilde{v2}}.$$

If she only slightly manipulates too much, the payoff function switches to $P_1^{(b)}$, and her maximal payoff for failed manipulation is

$$P_1^{(b)fail} = 1 - \frac{\widetilde{v_1}}{2\widetilde{v_2}}.$$

The payoff from strategy 'honesty' is

$$P_1^{(a)honest} = \frac{\widetilde{v_1}}{\widetilde{v_1} + \widetilde{v_2}}.$$

Player 1 could at the most gain $g = P_1^{(a)max} - P_1^{(a)honest}$. On the other hand, if her manipulation fails, she will at least lose $l = P_1^{(a)honest} - P_1^{(b)fail}$. Calculating the difference between maximal potential gains and minimal potential losses results in the term

$$g - l = P_1^{(a)max} + P_1^{(b)fail} - 2P_1^{(a)honest} = 1 - \frac{2\widetilde{v_1}}{\widetilde{v_1} + \widetilde{v_2}}.$$

In II, $\widetilde{v_1} > \widetilde{v_2}$, and therefore

$$g - l = 1 - \frac{2\widetilde{v_1}}{\widetilde{v_1} + \widetilde{v_2}} < 0.$$

Optimal strategy in constellation III

$\widetilde{v_2}$ lies in III, hence $\widetilde{v_2} < 1 - \widetilde{v_1} < \widetilde{v_1} < 1 - \widetilde{v_2}$. Depending on 1's choice of v_1, either payoff function (a), (c), or (d) is applicable. The payoff functions are different from constellations I and II. Also local and global maxima differ.

- For $v_1 > 1 - \widetilde{v_2}$: $P_1^{(a)} = \frac{\widetilde{v_1}}{v_1 + \widetilde{v_2}}$; which is strictly decreasing with v_1; local maximum in (a): $P_1^{(a)max} = \widetilde{v_1}$ for $v_1 = 1 - \widetilde{v_2}$
- For $\widetilde{v_2} < v_1 < 1 - \widetilde{v_2}$: $P_1^{(c)} = 1 - \frac{(1-\widetilde{v_1})}{(1-v_1)+(1-\widetilde{v_2})}$; which is strictly decreasing with v_1; local maximum in (c): $P_1^{(c)max} = 1 - \frac{(1-\widetilde{v_1})}{2(1-\widetilde{v_2})}$ for $v_1 = \widetilde{v_2}$
- For $v_1 < \widetilde{v_2}$: $P_1^{(d)} = \frac{1-\widetilde{v_1}}{(1-v_1)+(1-\widetilde{v_2})}$, which is strictly increasing with v_1; local maximum in (d): $P_1^{(d)max} = \frac{(1-\widetilde{v_1})}{2(1-\widetilde{v_2})}$ for $v_1 = \widetilde{v_2}$.

From $\widetilde{v_2} < 1 - \widetilde{v_1} < \widetilde{v_1} < 1 - \widetilde{v_2}$ follows $P_1^{(c)max} > P_1^{(d)max}$ and $P_1^{(c)max} > P_1^{(a)max}$.

To see the latter, calculate $P_1^{(c)max} - P_1^{(a)max} = \frac{(1-\widetilde{v_1})*(1-2\widetilde{v_2})}{2(1-\widetilde{v_2})} > 0$ (since $\widetilde{v_2} < 0.5$, all factors are < 0. The global maximum is therefore $P_1^{max} = P_1^{(c)max}$. This means that player 1's optimal strategy is $v_1 = \widetilde{v_2}$ but still a little above $\widetilde{v_2}$ to stay in (c) (technically correct: $v_1 = \widetilde{v_2} + \varepsilon$ with $\varepsilon \to 0$ and $\varepsilon > 0$). Figure 6 illustrates the payoff function for $\widetilde{v_1} = 0.8$ and $\widetilde{v_2} = 0.3$.

Gains and losses from manipulation in constellation III

If player 1's optimal manipulation strategy works, her payoff against the honest player 2 is $P_1^{(c)max} = 1 - \frac{(1-\widetilde{v_1})}{2(1-\widetilde{v_2})}$.

If she only slightly manipulates too much, the payoff function switches to $P_1^{(d)}$, and her maximal payoff for failed manipulation is

$$P_1^{(d)fail} = \frac{(1-\widetilde{v_1})}{2(1-\widetilde{v_2})}.$$

The payoff from strategy 'honesty' is

$$P_1^{(c)honest} = 1 - \frac{(1-\widetilde{v_1})}{(1-\widetilde{v_1})+(1-\widetilde{v_2})}.$$

Player 1 could at the most gain $g = P_1^{max} - P_1^{honest}$. On the other hand, if her manipulation fails, she will at least lose $l = P_1^{honest} - P_1^{fail}$. Calculating the difference between maximal potential gains and minimal potential losses results in the term

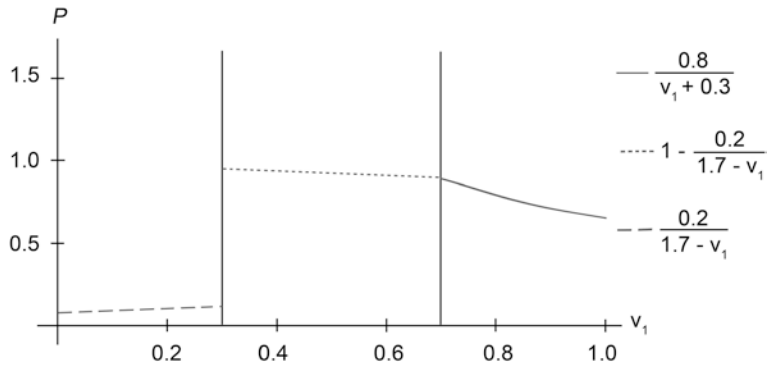


Fig. 6 Exemplary payoff function for constellation III with $\widetilde{v_1} = 0.8$ and $\widetilde{v_2} = 0.3$

$$g - l = P_1^{(c)max} + P_1^{fail} - 2P_1^{honest} = \frac{2(1-\widetilde{v_1})}{(1-\widetilde{v_1}) + (1-\widetilde{v_2})} - 1.$$

In III, $1 - \widetilde{v_2} > 1 - \widetilde{v_1}$, and therefore

$$g - l = \frac{2(1-\widetilde{v_1})}{(1-\widetilde{v_1}) + (1-\widetilde{v_2})} - 1 < 0.$$

Appendix C: Calculations for the Threat-Equivalent Equilibrium for Other Constellations

Constellation III

Consider first true valuation constellation III with $\widetilde{v_1} > \widetilde{v_2}$ and $1 - \widetilde{v_2} > \widetilde{v_1}$ (see Fig. 2). Again, both players manipulate their valuation towards the other's true valuation, and also for every Nash equilibrium $v_1 = v_2 = v$. However, two types of equilibria need to be distinguished in III, namely those $v > 0.5$ from those with $v < 0.5$.

In the cases of $v > 0.5$, the payoff functions are the same as in constellation I, only with reversed roles of the players. The proof for the threat-equivalent equilibrium and the resulting payoff therefore remains the same.

In the cases of $v < 0.5$, the payoff functions are

- $P_1^{(c)} = 1 - \frac{1-\widetilde{v_1}}{(1-v_1)+(1-v_2)}$ and $P_2^{(d)} = \frac{1-\widetilde{v_2}}{(1-v_1)+(1-v_2)}$ for compatible demands $v_1 > v_2$
- $P_1^{(d)} = \frac{1-\widetilde{v_1}}{(1-v_1)+(1-v_2)}$ and $P_2^{(c)} = 1 - \frac{1-\widetilde{v_2}}{(1-v_1)+(1-v_2)}$ for incompatible demands $v_1 < v_2$.

The threat-equivalent equilibrium is then calculated along similar lines:

$$P_1^{(c)}(v,v) - P_1^{(d)}(v,v) = P_2^{(d)}(v,v) - P_2^{(c)}(v,v)$$

$$1 - \frac{1-\widetilde{v_1}}{2(1-v)} - \frac{1-\widetilde{v_1}}{2(1-v)} = \frac{1-\widetilde{v_2}}{2(1-v)} - \left(1 - \frac{1-\widetilde{v_2}}{2(1-v)}\right)$$

$$1 - \frac{1-\widetilde{v_1}}{1-v} = \frac{1-\widetilde{v_2}}{(1-v)} - 1$$

$$v = \frac{\widetilde{v_1}+\widetilde{v_2}}{2}.$$

The payoffs are again the same as from truth-telling.

$$P_1^{(c)}\left(\frac{\widetilde{v_1}+\widetilde{v_2}}{2}, \frac{\widetilde{v_1}+\widetilde{v_2}}{2}\right) = 1 - \frac{1-\widetilde{v_1}}{2-2\left(\frac{\widetilde{v_1}+\widetilde{v_2}}{2}\right)}$$
$$= 1 - \frac{1-\widetilde{v_1}}{(1-\widetilde{v_1})+(1-\widetilde{v_2})} = P_1^{(c)}\left(\widetilde{v_1}, \widetilde{v_2}\right)$$

$$P_2^{(d)}\left(\frac{\widetilde{v_1}+\widetilde{v_2}}{2}\right) = \frac{1-\widetilde{v_2}}{2-2\left(\frac{\widetilde{v_1}+\widetilde{v_2}}{2}\right)} = \frac{1-\widetilde{v_2}}{(1-\widetilde{v_1})+(1-\widetilde{v_2})} = P_2^{(d)}\left(\widetilde{v_1}, \widetilde{v_2}\right).$$

Now consider symmetric manipulation. For compatible manipulation ($t < \frac{\widetilde{v_2}-\widetilde{v_1}}{2}$) in constellation III:

$$P_1^{(c)}\left(\widetilde{v_1}+t, \widetilde{v_2}-t\right) = 1 - \frac{\left(1-\widetilde{v_1}\right)}{\left(1-\left(\widetilde{v_1}+t\right)\right)+\left(1-\left(\widetilde{v_2}-t\right)\right)}$$
$$= 1 - \frac{\left(1-\widetilde{v_1}\right)}{\left(1-\widetilde{v_1}\right)+\left(1-\widetilde{v_2}\right)} = P_1^{(c)}\left(\widetilde{v_1}, \widetilde{v_2}\right)$$

$$P_2^{(d)}\left(\widetilde{v_1}+t, \widetilde{v_2}-t\right) = \frac{1-\widetilde{v_2}}{(1-(\widetilde{v_1}+t))+(1-\widetilde{v_2}-t))} = \frac{1-\widetilde{v_2}}{(1-\widetilde{v_1})+(1-\widetilde{v_2})}$$
$$= P_2^{(d)}\left(\widetilde{v_1}, \widetilde{v_2}\right).$$

For incompatible manipulation ($t > \frac{\widetilde{v_2} - \widetilde{v_1}}{2}$) in constellation III, payoff functions are once again reversed.

$$P_1^{(d)}\left(\widetilde{v_1} + t, \widetilde{v_2} - t\right) = \frac{1 - \widetilde{v_1}}{(1 - (\widetilde{v_1} + t)) + (1 - (\widetilde{v_2} - t))}$$

$$P_2^{(c)}\left(\widetilde{v_1} + t, \widetilde{v_2} - t\right) = 1 - \frac{\left(1 - \widetilde{v_2}\right)}{\left(1 - \left(\widetilde{v_1} + t\right)\right) + \left(1 - \left(\widetilde{v_2} - t\right)\right)}$$

$$= 1 - \frac{\left(1 - \widetilde{v_2}\right)}{\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right)}.$$

Comparing this with the payoffs from mutual truthfulness

$$P_1^{(d)}\left(\widetilde{v_1} + t, \widetilde{v_2} - t\right) - P_1^{(c)}\left(\widetilde{v_1}, \widetilde{v_2}\right) = \frac{1 - \widetilde{v_1}}{\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right)} - \left(1 - \frac{1 - \widetilde{v_1}}{\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right)}\right)$$

$$= \frac{2 \cdot \left(1 - \widetilde{v_1}\right)}{\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right)} - 1 = \frac{2 \cdot \left(1 - \widetilde{v_1}\right) - \left(\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right)\right)}{\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right)}$$

$$= \frac{\left(1 - \widetilde{v_1}\right) - \left(1 - \widetilde{v_2}\right)}{\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right)}$$

$$P_2^{(c)}\left(\widetilde{v_1} + t, \widetilde{v_2} - t\right) - P_2^{(d)}\left(\widetilde{v_1}, \widetilde{v_2}\right) = 1 - \frac{1 - \widetilde{v_2}}{\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right)} - \left(\frac{1 - \widetilde{v_2}}{\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right)}\right)$$

$$= 1 - \frac{2 \cdot \left(1 - \widetilde{v_2}\right)}{\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right)} = \frac{\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right) - 2 \cdot \left(1 - \widetilde{v_2}\right)}{\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right)}$$

$$= \frac{\left(1 - \widetilde{v_1}\right) - \left(1 - \widetilde{v_2}\right)}{\left(1 - \widetilde{v_1}\right) + \left(1 - \widetilde{v_2}\right)}.$$

Therefore, potential losses are also of the same size in the case of symmetric over-manipulation.

Constellation II

What is now left to complete are the results for all cases in constellation II with $\widetilde{v_1} > \widetilde{v_2}$ and $1 - \widetilde{v_2} < \widetilde{v_1}$. No new calculations are necessary: If, in equilibrium, $v > 0.5$, payoff functions are the same as in constellation I, but with reversed roles of the players. If $v < 0.5$, the payoff functions are the same as in the case $v < 0.5$ in constellation III. If not in equilibrium, due to the assumed symmetry of manipulation, no switches in payoff functions occur and hence the same logic can be applied also to those cases.

References

Aziz, Haris, Simina Brânzei, Aris Filos-Ratsikas, and Søren K. S. Frederiksen. 2015. *The Adjusted Winner Procedure: Characterizations and Equilibria, arxiv.org [*arXiv:1503.06665*].*

Brams, Steven, and Alan Taylor. 1999: *The win-win solution. Guaranteeing fair shares to everybody.* New York: W.W. Norton.

Fehr, Ernst, and Klaus Schmidt. 1999. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics* 114 (3): 817–868.

Kalai, Ehud, and Meir Smorodinsky. 1975. Other solutions to Nash's bargaining problem. *Econometrica* 43: 513–518.

Nash, John. 1950. The bargaining problem. *Econometrica* 18: 155–162.

Nash, John. 1953. Two-Person Cooperative Games, *Econometrica* 21 (1): 128–140.

Raith, Matthias. 2000. Fair-negotiation procedures. *Mathematical Social Sciences* 39 (3): 303–322.

Rawls, John. 1985. Justice as Fairness: Political not Metaphysical. *Philosophy and Public Affairs* 14: 223–251.

Sauermann, Jan, and Paul Beckmann. 2017. 'Divide the dollar' using voting by veto, in this book.

Schüssler, Rudolf. 2007. "Adjusted Winner" (AW) Analyses of the 1978 Camp David Accords—Valuable Tools for Negotiators? In *Diplomacy games. Formal models and international negotiations*, eds. Rudolf Avenhaus, and I. William Zartman, 284–296. Berlin: Springer.

Paper 2

Rationally Poor?

–

What the Emergence of Inequality can Teach us About Rational Behaviour

---

This is an unpublished manuscript. A strongly modified version of this paper has been published in *Synthese* (ISSN: 0039- 7857).

The suggested citation is:

Klein, Dominik, Johannes Marx and Simon Scheller (2017). *Rationally Poor? – What the Emergence of Inequality can Teach us About Rational Behaviour.* Bamberg: University of Bamberg (mimeo).

Noname manuscript No.
(will be inserted by the editor)

Rationally Poor?
What the Emergence of Inequality can Teach us About Rational Behaviour

Dominik Klein · Johannes Marx · Simon Scheller

Received: date / Accepted: date

Abstract The emergence of economic inequality has often been linked to individual differences in mental or physical capacities. By means of an agent-based simulation this paper shows that neither of these is a necessary condition. Rather, inequality can arise from iterated interactions of fully rational agents. This bears consequences for our understanding of both inequality and rationality. In a setting of iterated bargaining games, we find that expected utility maximizing agents perform suboptimally in comparison to other strategies. The reason for this lies in a complex interaction between this strategy and the quality of beliefs used to calculate expected utility. Consequentially, we argue that the standard notion of rationality as maximizing expected utility is insufficient, even for certain standard cases of economic interaction.

Keywords Inequality · Rationality · Bargaining · agent-based modelling · Rational Choice

1 Introduction

Inequality has been increasing over the last decades, as has been amply shown by Piketty (2014). In recent work, inequality has been discussed as a source of conflict between countries of the global south and north Wood (1995) but also as a problematic issue within developed western democracies Piketty (2014); Gottschalk and Smeeding (2000). An extensive literature identifies a broad range of social, cultural, educational and long-term-economic effects of inequality and poverty McLeod et al. (2014).[1]

In the present paper, however, we are concerned with the *causes* of inequality rather than its consequences. When inquiring into the origins of in-

[1] While poverty is a property that can be ascribed to one person, inequality is relational in nature. By inequality we refer to a concept that captures one's relative position compared to a group of relevant others Haughton and Khandker (2009).

equality, most approaches employ a macro level perspective. Others take a decision-theoretic perspective and focus on the quality of the individual's decisions Neckerman and Torche (2007); Dabla-Norris et al. (2015). We hold that both miss out on an important source of inequality. In an economy where agents interact with each other, revenue maximization must be situated in a game theoretic - rather than a decision theoretic - context. The present paper complements existing literature with a simulational approach, studying the emergence of inequality from a *game theoretic* perspective.

We start from the assumption that bargaining constitutes a ubiquitous (if not defining) feature of economic interaction – especially when distributional issues are settled. Additionally, modern economy is best described as a highly interconnected system, where agents frequently interact with each other. As a result, inequality may thus emerge through the interaction of agents in a complex system, and cannot be reduced to individual psychological traits of agents alone. This system of strategic dependencies should be expected to have an effect on the distribution of incomes. We therefore ask to what extent strategic behaviour in bargaining processes can be a source of lasting inequality in societies.

In analyzing our model, we put a special focus on the relationship between inequality and aspects of rationality of the agents involved. Literature on the connection between poverty, inequality and rationality is scarce, but that which exists provides some crucial insights. We refer to Sheehy-Skeffington and Rea (2017) for a comprehensive overview. A common result of the recent empirical literature on poverty and rationality is that "the poor often behave differently from the non-poor" Carvalho et al. (2016) and that the former frequently fail to meet the high requirements of rationality. Two competing rationales are frequently put forward for this finding. One strand of literature argues that poverty impedes mental capacities and that being poor affects an agent's goals, their taste for risk or the choices made in strategic situations (Mishra et al., 2015; Mani et al., 2013). These changes in mental capacity can, a fortiori, diminish the economic performance of agents in the long run, thus furthering existing inequalities. A second strand of literature takes the differential behaviours of poor and rich agents as reflecting rational adaptations to the different environment agents are exposed to Carvalho et al. (2016). In this interpretation, poorer agents do not suffer from a lack of rationality, but behave optimally in light of their specific set of constraints.

We side with the second of these two strands of literature. Poorer agents in their adoption of certain strategies do not necessarily fail to perform rationally. Rather, the standards of rational choice may be context sensitive and the notion of rationality as maximising expected utility is not sufficient for evaluating strategic behaviour. This translates to two specific questions: First, on an individual level, which strategies should agents choose in order to maximise their gains and, more generally, their short and long term prospects? Second and on a more abstract level: Is the *thin* notion of rationality as maximizing expected utility sufficient for evaluating strategic behaviour or do certain situations demand an extension of the concept of rationality? Our simulation

demonstrates that, depending on circumstances and the agent's background, different decision rules can be identified as rational. In section 4, we relate these findings to a thicker notion of rationality, taking the agents' beliefs and, more generally, their informational economy into account. More particularly, we show that an agent's action might indirectly influence her future beliefs and thus her future strategic choices. Also, strategic possibilities for agents are shown to be largely dependent on short term needs, which in turn are directly linked to sufficient initial endowments. We argue that the standard notion of rationality as expected utility maximisation falls short of capturing epistemic aspects of rational action as well as side constraints on maximisation. Thus, with the present simulation, we argue for the need of a substantive thicker theory of rationality.

2 Bargaining as a Generative Mechanism for Inequality: Theory and Model

In a highly interactive and integrated economy, inequlity should be conceptualised as an emergent property of a complex system. By focussing on bargaining games, we aim to capture the distributional component of joint production processes relevant to the emergence of inequality.[2] Bargaining games are directly related to distributional matters and hence to the emergence of equality or inequality.

We are interested in situations where a mutually beneficial economic endeavour is feasible, for example a joint production process or an employer hiring an employee for a certain job. We do not assume that both agents have symmetric or interchangeable roles in the process, as one agent may represent a potential employer and the other an employee. All we assume is that a successful production creates a surplus that is to be divided among the two agents. However, before production can start, agents have to agree on the division of the expected benefits. Only after having done so will they engage in the production process and distribute the benefits according to their prior agreement. An extended bargaining process is costly in itself. More specifically, we assume that the time spent on bargaining cannot be used for production. The longer the bargaining process lasts, the less time remains for the actual production. This leaves actors in a situation where both have a strong interest in shifting the outcome in their own favour. One crucial way of doing so is to bargain long and hard to obtain a favourable deal. That is, *not to give in* to the opponent's demands, but wait until she accepts one's own terms. Imagine, for example, a situation in which two partners are bound together by an incomplete contract. Such an assumption is, for example, put forward by Hart and Moore (1999), who argue that in fact all contracts can be seen

[2] In general, economic interaction consists of first-level economic problems addressing the coordination problems surrounding the production of goods (i.e. positive-sum-games) and bargaining games as zero sum games representing the second-level economic problems of allocating a surplus generated.

as incomplete and are hence subject to renegotiation after initiation. In such a setting, agents could possibly generate large of gains, but only if they succeed in agreeing to a potential distribution. Lipman (1986) asserts that within many such situations "each party prefers a poor agreement to no agreement" (Lipman, 1986, p. 317). Hence, agents involved are faced with the challenge to establish a common bargaining solution within a competitive environment. The crucial decision they have to make is *when* (if ever) to agree on a successful distribution. We hold that these *endurance competitions* depict a relevant feature of bargaining problems in the real world.

Classically, bargaining behaviour is determined by when and how far to adjust one's own demands in reaction to the opponent's behaviour. Ceteris paribus, the later and the smaller the adjustments an agents makes, the better the bargaining solution will be for her, once found. The downside to a tough bargaining strategy, obviously, is that it increases the expected time and hence the cost until a common solution is found. With the present simulation, we limit ourselves to a simplified bargaining process. For reasons of tractability, we represent the agents' toughness in bargaining by a single parameter, denoting after how many rounds of unsuccessful coordination they are willing to adjust their demands. By a slight idealization, we assume that the agent who adjusts first does so in such a way that a common ground is found. If both agents adjust simultaneously, they split the difference and meet in the middle.[3]

The Bargaining Model

To represent these situations formally, we constructed an agent based model of iterated bargaining encounters. At the beginning of each simulation run, players are matched together in pairs randomly for a fixed number of bargaining rounds rounds. Whenever this maximal interaction length is reached, all matchings are dissolved, new random pairs are formed and interact for the same fixed number of rounds.[4] Players can neither choose their partner freely, nor can they leave a partner prematurely or stay longer than the fixed number of rounds.

In the bargaining process, both agents have to decide between making a *high* or a *modest* demand on the surplus generated. They make their claims simultaneously and without knowledge of the other player's action. We assume that a high demand from both players is *incompatible*. In this case, no pay-offs are generated and they need to enter into a further round of bargaining of the same game structure. Bargaining is continued until one or both agents lower their demands so that compatibility is reached or until the maximal interaction time is up.

[3] Hence, we allow for egalitarian solutions. However, in light of the competitive environment, these are not reached through egalitarian offers, but they arise if both partners have a similar bargaining strength.

[4] Within each run, the number of interaction rounds is constant. Between different runs, we varied the values of this parameter systematically between 10 and 20.

Every other combination, i.e. a high and a modest or two modest demands is admissible. In this case, the players start producing the surplus and divide all benefits produced by the distribution agreed upon. They will use each subsequent round to produce a surplus, which is distributed according to the agreed-upon solution. That is, once agents agree on a distribution, they will not renegotiate, but keep producing as long as they are matched together.

The one-round bargaining situation is represented by a game similar to a chicken game, which can be seen as a simplified version of Nash's demand game Nash (1953). In this game a total common resource worth a utility of 4 for either agent is to be distributed. A high demand is represented by a utility of 3 and a modest demand by a utility of 1. Both players make their demands simultaneously. The demands are compatible if their sum does not exceed the value of the resource. If demands are incompatible (i.e. larger than 4 in sum), neither player receives anything. If demands are compatible, each agent receives what she had demanded. Additionally, by a small deviation from the original framework, if both agents make a modest demand, we assume the remaining resource is divided equally between both agents, i.e. both receive a utility of 2.

		Player Y	
		high	*modest*
Player X	*high*	$(0,0)$	$(3,1)$
	modest	$(1,3)$	$(2,2)$

Fig. 1 Normal form of the baseline game

Given that there are only two possible moves in the one-shot game, one of which guarantees successful cooperation, a player's strategy is determined entirely by specifying how long a player is willing to maintain a high demand when no cooperation is achieved.[5] For obvious reasons, we call this parameter an agent's 'toughness'. A toughness of 0 thus denotes agents who start off with a modest demand, while agents whose toughness equal to the number of interaction rounds will always place high demands. Hence, one could describe the bargaining process between two players as the player with the higher toughness holding out until the other player gives in. The lower of the

[5] Skyrms (2014) emphasizes the existence of a fair solution as a special focal point. However, in most economic interactions, no saliently egalitarian distribution exists. When an employer bargains about wages with an employee, the different possible agreements might benefit either the employer or the employee more. Consequently, no possible wage offer could be regarded as obviously egalitarian in any way. Notably, we do not preclude that some distribution actually divides the achieved surplus fairly. We merely claim that this property is by no means transparent enough to both parties involved to qualify as a focal point. We hold that this is a common feature of many bargaining situations, from wage negotiations to partners with complimentary skills setting up a joint endeavor. While a numerical representation of pay-offs in the chicken game seems to suggest that there is an egalitarian solution, this might often be an artifact of the modelling.

two toughness values determines after how many rounds this will be the case. Thus, the player giving in first determines how much time is spent bargaining, how many rounds are left for the production of a surplus and who gains what from the interaction. Finally, if both agents have the same level of toughness, this will result in an egalitarian distribution of the surplus generated in the remaining time.

The rationale behind the choice of toughness can be intuitively interpreted as follows: A high value aims at making the opponent give in eventually, so that the player receives a high pay-off from that point on. A high toughness, however, bears the risk of forgoing many rounds without income, before either player gives in. In the worst case, such a strategy can result in minimal gains if an agreement is found too late or the opponent turns out to have an even higher level of endurance. An agent with low toughness, in contrast, rather gives in earlier in order to avoid enduring too many rounds without income. Accepting a lower pay-off in each successive round after having given in is thereafter the price that must be paid.

Another way to describe the situation is based on the recognition that the one-shot-game offers two Nash-Equilibria in pure strategies. These are when one player chooses the 'high' strategy while the other player opts for 'modest'. Each player has an interest in establishing the equilibrium with pay-offs in her favour. For this, it is necessary to start with the high demand, and then to wait until the other player switches to 'modest'. The rationale of the bargaining game therefore symbolises a classic war of attrition.

Learning and Information Processing

Agents can use the experience gained in previous games for choosing their level of toughness. Crucially, we do not assume that agents learn specific information about certain individual others. Each pairing is resolved after a finite amount of time and agents do not expect to encounter the same opponent again. They can, however, form expectations about how unknown others in society behave or how frequent the different values of toughness exist. These expectations are formed purely on the basis of agents' experiences in previous interactions.

To understand the mechanics of the game better, we need to distinguish between two different types of information a player can gain. First, if the agent 'wins' a game, i.e. if the opponent concedes first or both give in at the same time, she receives an exact signal about the toughness of her opponent: If an opponent gives in after 3 rounds, the agent knows for certain that the opponent's toughness was 3. The second type of information concerns the case where a player "loses" her current game, i.e. she decides to give in before her opponent does. In this case, she receives only an imprecise signal about the opponent's toughness, namely that it is larger than her own.

In both cases, the information gained is incorporated into the agent's subjective probability distribution of toughness by means of a weighted update.[6]

[6] This learning rules represents statistical learning with temporal discounting. Note that the object our agents learn about is not a fixed distribution of toughness, as the distribu-

That is, each agent has a prior probability distribution p_{prior} about how likely the different levels of toughness are. After each encounter, the agent can then calculate her posterior distribution p_{op} of the strategies played by her current opponent. She does so by conditioning her prior distribution p_{prior} in accordance with the observed behaviour:

$$p_{op} = p_{prior}|\text{Observation}.$$

This updated information about the opponent's actions then feeds back into the agent's general probability distribution of what agents in society do. She calculates her updated probability distribution $p_{updated}$ by means of a weighted average:

$$p_{updated} = 0.9 \cdot p_{prior} + 0.1 \cdot p_{op}.$$

At the beginning of a simulation run, agents start with a uniform prior, i.e. they consider all values of toughness equally likely.

Strategies

Toughness is an agent's key strategy parameter. The choice of toughness reflects a variety of different strategic considerations, such as reducing possible losses, maximising possible gains, maximising expected utility, or, on the more epistemic side, finding out when an opponent is likely to give in. In general, agents can adapt their toughness from game to game, depending on which value of this parameter they find most promising. However, in choosing the level of toughness, different strategies might be cognitively more or less demanding. Normally, the choice of strategy may depend on the information available to the agents, their current financial situation and their available cognitive resources.

In the current model, we explore five different strategy types that guide the choice of toughness. Later, we will add two further types for the sake of conceptual exploration. By a strategy, we mean the way an agent updates her toughness after having encountered another agent. By no means do we claim that these choices exhaust the realm of possible strategies, even in this already simplified scenario. Nevertheless, we seek to capture intuitions, arguments and heuristic rationales about how the game could reasonably be played, and to incorporate the strategies that are assumed to be the most prominent in such situations.

– **MaxEU**: This first type always chooses the toughness that maximises her subjective expected utility. Naturally, the utility gained depends on the

tion might change itself over time as others adapt their toughness. Consequently, an agent might rationally apply some temporal discounting to the information collected. That is, after observations $o_1, \ldots, o_n,$, with o_1 the most recent, an agent might infer to a temporally discounted distribution $d_{disc} = \frac{1}{\sum q^i} \sum_1^n q^i o_i$ rather than the classically frequentist $d_{freq} = \frac{1}{n} \sum_1^n o_i$. When the number of previous data points converges to infinity, this converges to $d_{disc} = (1-q)o_1 + qd_{prior}$, where $d_{prior} = \sum_{i=2}^n o_i q^{i-1}$ is the discounted frequentist distribution before learning o_1. This is exactly the rule above with $q = 0.1$.

player's own toughness as well as that of the opponent. A utility maximiser needs to estimate therefore how likely an opponent is to play the various toughness levels. She does so with the learning mechanism outlined above. Starting with uniform priors, i.e. no information at all, she gradually learns about the behaviour of others and updates her distribution accordingly. This strategy type embodies the optimal strategy for a rational, risk-neutral player. Furthermore, according to the law of large numbers, this strategy should be expected to perform best in terms of long term accumulation of wealth.

- **Maximin**: Agents of this type always play a toughness of 0, no matter what. That is, they give in immediately, thus ensuring that a maximal amount of surplus is produced. The rationale behind the Maximin's strategy is that this player type is 'infinitely risk averse'. The Maximin takes a guaranteed pay-off of 1 per round rather than risking any incompatibility. As the name suggests, this strategy embodies the classic *Maximin* principle.
- **Maximax**: This type never gives in, but is prepared to outwait her opponent at any cost. That is, the Maximax's toughness is always set to the maximal possible level. The rationale for this strategy is the opposite of "Maximin". Maximax agent are willing to accept any number of incompatible rounds for the chance to obtain the maximal possible pay-off of 3. This strategy follows an iterated *Maximax* principle.
- **Experimenter**: Experimenter is a mixture between the types of MaxEU and Maximax. Before each game, this type chooses between two possible behaviour types. With a probability of 90 %, an Experimenter adopts the MaxEU strategy. However, with the remaining probability of 10 %, she adopts the Maximax strategy. The reason for doing so is different though. Whilst the original Maximax strategy is simply prepared to sacrifice everything for the prospect of high per-round-income, the main motivation of experimenters is to gain information about the opponent's behaviour, in order to make better choices while playing the MaxEU-strategy. The information gained is highest when outwaiting the opponent, since it is only then that an agent receives an unambiguous signal about the opponent's toughness. Hence, the choice of Maximax is the only strategy guaranteed not to give in first.
- **Increase-decrease**: This player-type follows a simplistic way of updating her strategy: Whenever such a player loses a game, she reduces her toughness by one. Vice versa, whenever she wins a game, she increases her toughness by one. The reasoning of this heuristically-rational strategy can be explained as follows: When the player loses, she holds that she is not able to maintain her demand long enough to secure a high pay-off. She thus reasons that she must give in a little earlier, in order to receive at least the lower pay-off over a greater number of compatible rounds. After a victorious round, however, the agent learns that being tough paid of in the end. She is thus encouraged to be even more tough in the next encounters. Increase-decrease describes a decision-making heuristic that is clearly

suboptimal, yet may characterise the behaviour of some real life agents appropriately, as argued by Mishra et al. (2015). Increase-decrease players start with a random toughness in the very first game.

Simulation Experiments

We simulate the described model and analyse its output on the basis of three major experiments. The first series constitutes the baseline model, in which we are interested in the average long term success of the different strategies only. The second experiment introduces evolutionary mechanisms, while the third studies whether evolutionary pressure is structurally different for rich and poor agents. To do so, we add an additional model parameter, *cost of living*. While players receive pay-offs in the games they play, they also incur a certain cost c for maintaining their lives through each round. In order to stay alive, each player must spend c from her accumulated wealth in every round. At the beginning of a simulation run, each agent is equipped with a small initial endowment. In the third experiment we introduce two classes of agents, poor and rich. These differ in their initial wealth endowments.

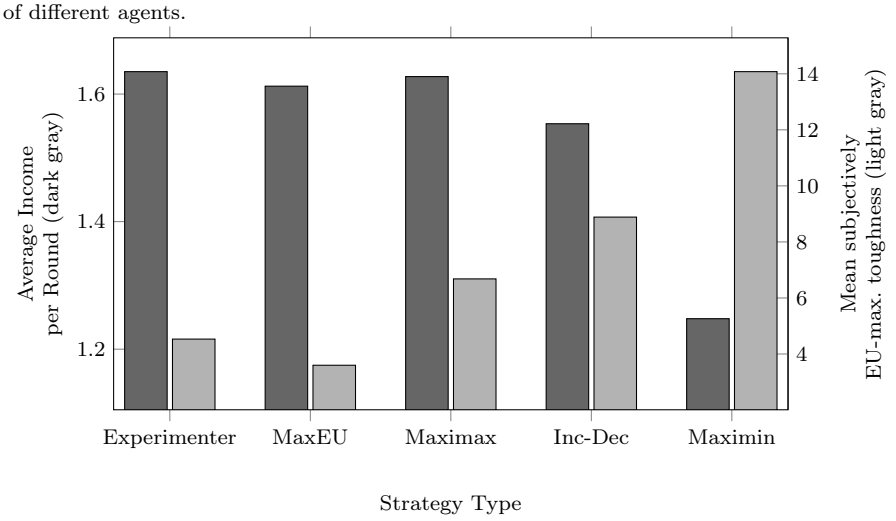
We use three different output measures for analysing the model. First, we are interested in the wealth accumulated by agents of different strategy types. We take wealth as an indicator for the long term bargaining success of the different strategy types. Second, in those experiments with evolutionary pressure, we measure the proportions of player types at the ends of simulation runs. These proportions of player types are used as a measure for the bargaining success of different strategies. Under evolutionary pressure, well performing strategies will survive and reproduce, while inefficient strategies are more likely to die out. Thus, successful strategies will be played by many at the end of a run, while unsuccessful strategies will barely be present. Third, we are interested in the content and quality of information collected by the different strategies. Each agent collects new information while being engaged in bargaining situations. Since an agent's toughness impacts the quantity and quality of information gained, the various agent types might differ structurally in the content of beliefs they hold at the end of a simulation run. We compare the *accuracy* of beliefs held by the different agents by means of a proxy measure. Each simulation run in these three experiments starts with a total of 100 agents, 20 of each from the five types described above.[7]

3 The Poor Performance of EU Maximizers

This section presents the results of the three described experiments. All data is based on the range of parameter values described earlier, with 100 iterations for each parameter combination. One simulation run lasts for 1.000 interaction

[7] To check robustness, we ran various simulations with other distributions of the five agent types and for different numbers of bargaining rounds. For a broad variety of input parameters the results are similar to those reported here.

Fig. 2 Average income per round (dark gray) and EU maximizing toughness (light gray) of different agents.



rounds, before final measurements are taken. For the second and third set of simulations, we aborted the simulation prematurely if no agent had to exit due to negative funds for 100 rounds. If a group of agents died out completely, its mean wealth was set to zero.

3.1 The Baseline Models

The baseline models works without evolutionary pressure and no cost of living. All agents start with the same wealth of zero, which makes the final wealth at the end of a simulation a direct measure of different strategies' bargaining success. In this experiment, the types Maximax, MaxEU and Experimenter all perform well, while Maximin performs by far the worst in terms of accumulated wealth, see Figure 2. These outcomes demonstrate that, in basic terms, the maximin approach comes at a price. Each time a player makes a high demand, she risks a round of failed coordination and hence no production, without being guaranteed to gain more later in the interaction. The Maximin strategy avoids such risks. It settles for a secure pay-off of one unit per round, with the slight chance of receiving two, should both players give in at the first round.

We should, however, emphasize that these effects are in some way dependent on the distribution of strategy types. In a small class of distributions which are somehow degenerate, Maximin might perform extremely well. This can be the case, for instance, when a few agents of this type enter a society almost exclusively consisting of Maximax types. For the more regular distributions of agent types, however, the results are similar to those presented here.

The second set of results from this experiment concerns the quantity and quality of information collected by the different strategies. Recall that only the 'winning' player is given exact information about her opponent's toughness, while the player giving in first merely receives the imprecise information that the opponent's toughness is above hers. Hence, strategy types that tend to give in first (i.e. Maximin) will collect very little information. On the other end of the spectrum, strategies that never give in, such as Maximax, will collect the maximal amount of information possible. However, we are not so much concerned about the quantity of information than its content. We will measure this by a simple proxy. Notably, the beliefs of an agent about the toughness of others impact her assignments of (subjective) expected utility to the various toughness levels available to her. In particular, her beliefs impact which choice of toughness *maximises* expected utility. We use this (EU-) optimal toughness level as a rough proxy to assess the content of an agent's beliefs or at least the beliefs an agent could have formed in light of the available information. We should emphasise here that most agents do not make use of the full information they collect. Only the type MaxEU and, in 90% of games, the Experimenter-types actually employ the information collected for calculating expected utilities. All other agents do not employ the information collected. In fact, for any interpretation of the model, we do not need to assume that those agents keep track of their incoming information at all. This may be important when discussing how cognitively demanding the different strategies are. The current analysis can thus be described as comparing the different types' available information by discussing which beliefs they could have formed, had they processed the information available to them.

In our experiment, Maximin identifies the highest value of toughness on average for maximising expected utility, see light gray bars in Figure 2. While this may seem counterintuitive at first, note that Maximin gains no information about any toughness higher than zero, and therefore never updates the higher parts of her probability distribution. Her optimal strategy is completely determined by her initial beliefs and thus by the uniform priors we chose. Had we started with a different initial distribution, Maximin would have displayed a different identified optimal strategy. Further, we note that all other agents obtain at least some information about the toughness of others.

However, the relation between the amount of information gained and the optimal toughness value identified is not monotonic. Indeed, the types collecting least and most information, Maximin and Maximax respectively, are relatively close to each other in their assessments of optimal toughness, while both MaxEU and Experimenters locate the ideal toughness far below, see Figure 2. Since Maximax agents never give in first and hence always learn the true value of each opponent's toughness, these results might be taken to indicate that more cautious agents systematically underestimate the potential of high toughness strategies. This points to an intricate relationship. Not only do agents employing moderate levels of toughness collect less detailed information, but also the content of this information turns out to be skewed. This skewdness might lead them to significantly underestimate the benefits of in-

sisting on high demands. This interpretation is further supported by analyzing the performance of the different strategies. By the law of large numbers, we *should* expect MaxEU to accrue the highest wealth in the long run. It does not. Rather, Maximax outperforms MaxEU significantly, see Figure the dark gray bars in 2. One possible explanation for this shortcoming is that the collected evidence of 'EU' agents is so strongly biased, that the subjectively optimal move differs widely from the objectively income maximizing toughness.[8]

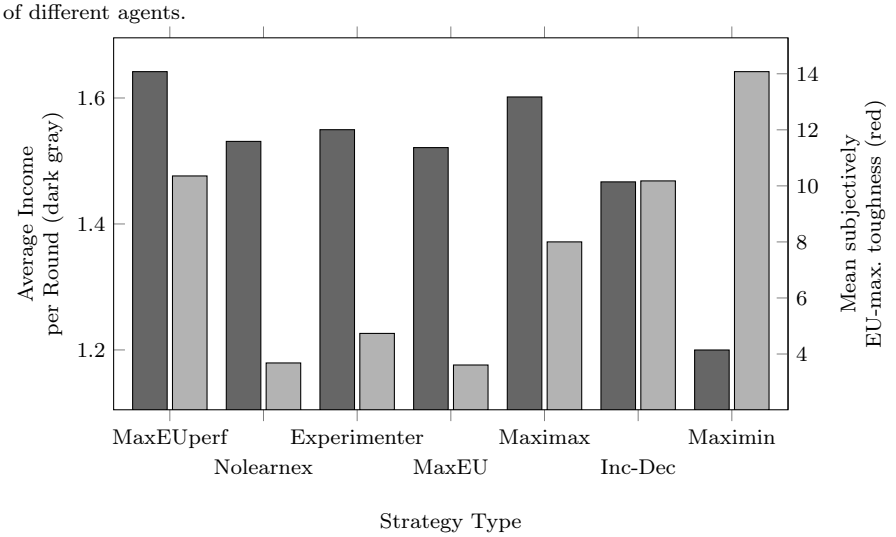
This leads us to a follow-up question, combining both findings so far. As depicted in Figure 2, Experimenters are apparently able to benefit from their strategy to be Maximax in 10% of the cases. But how so? There are at least two possible explanations. Either, an Experimenter benefits from the increased information that she gains in the 10% of cases where she plays Maximax (1), or an Experimenter merely benefits from the fact that Maximax perform better than MaxEU agents in some cases (2).

To decide between these two conclusively, we perform a further experiment with a new strategy type called *NoLearnex*. This type chooses her actions in the same way as Experimenter, i.e. a mixture of 90% MaxEU and 10% Maximax. On the epistemic side, however, the strategies differ. NoLearnex fails to incorporate any information acquired in the Maximax role. That is, NoLearnex combines the strategy choice of an Experimenter with the knowledge only from those games where she assumes the role of a MaxEU-agent. If this strategy fares as well as an Experimenter, we know that Experimenter's success is caused mainly by the higher benefits achievable as a Maximax. However, if Experimenter outperforms NoLearnex, we can infer that this difference must be caused by differences in the available information, i.e. the information collected as Maximax. As the dark gray part of Figure 3 shows, Experimenter outperforms NoLearnex significantly, thus supporting the first hypothesis. Or to put it differently: It is beneficial to be very tough and act as a Maximax every now and then, not only for its own expected utility, but also for the sake of collecting information that allows to act optimally in future interactions. For the sake of comparison, we also plot the income an EU maximising strategy could make were it based on fully accurate information about the current distribution of toughness. This information is listed as MaxEUperfect. The light gray part of Figure 3 shows which toughness values the different strategies identify as optimal. MaxEUperfect thereby constitutes the reference point what would have been the optimal toughness choice in such a situation.

We should highlight an intricate property of the strategy MaxEU that follows from this analysis: While this type maximises expected utility in light of the current beliefs, the information collected along the way leaves MaxEU not only poorly informed, but actively misinformed. It is exactly this misinforma-

[8] A further candidate explanation for MaxEU's lack of perfomance is, that the chosen priors are far off. To rule out this interpretation, we ran an extended simulation run over 2000 interaction rounds and measured the wealth accumulated only in the last 1000 rounds. This setup allowed MaxEU agents to adjust their beliefs well before pay-off collection began. The observed wealth distribution was similar to the one depicted in Figure 2, ruling out this explanation.

Fig. 3 Average income per round (dark gray) and EU maximizing toughness (light gray) of different agents.

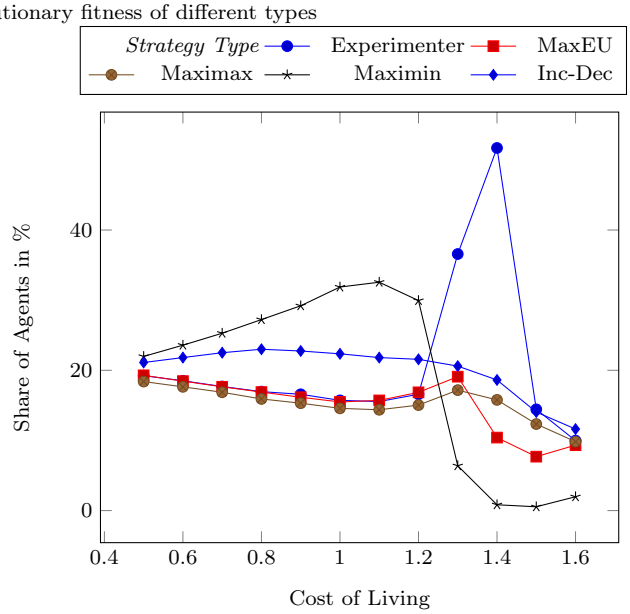


tion that allows MaxEU to be outperformed by Maximax, Experimenter and NoLearnex. In other words, MaxEU actively undermines the basis for its own success in the long run. The success of the Experimenter, conversely, shows how it can pay off sometimes to invest in examining the behaviour of others. Through its occasional adoptions of the Maximax behaviour, this strategy has more accurate beliefs at its disposal, allowing it to perform far better in the long run. Conversely, proponents of MaxEU-type strategies face a classic tradeoff between exploration and exploitation. They have to decide how often to engage in information searching and when to exploit the information gained. Or, to put it differently, such agents face a constant tradeoff between short term rational action (in terms of maximising expected utility) and long term performance due to rational collection of evidence.

3.2 Evolution in a World of Poverty

We now shift focus from long term maximisation to questions of survival. In a first extension of the baseline model, we introduce a slight evolutionary pressure on the individual agents. More specifically, a cost of living c is introduced that agents have to pay each round. This cost of living is constant within a simulation run and the same for every agent, yet varies between different runs. Once an agent's has negative wealth when leaving an interaction cycle, this agent is removed from the simulation. To fill the gap, one of the remaining agents is chosen at random and duplicated, thus keeping the number of agents

Fig. 4 Evolutionary fitness of different types



constantly at 100.[9] With this simulation, we address matters of survival and evolutionary success. If a strategy cannot guarantee survival, some players with this strategy might be forced out. Successful strategies, on the contrary, can increase their share through the replacing mechanism of evolution. At the beginning of each simulation run, there are 20 agents of every type, each with a starting wealth between 5 and 20 drawn from a uniform distribution. The number of agents per strategy at the end of a simulation run is thus a direct measure for its success.

When the cost of living is too low, below 0.3, there is no significant evolutionary pressure and all strategies maintain their initial share. For a moderate cost of living between 0.3 and about 1.1, the strategy type Maximin has a significantly higher chance of survival and hence a higher share of the population than all other agent types, see Figure 4. Notably, this performance does not translate into expected wealth: In line with the findings from the first experiment, Maximin still perform worse in terms of aggregate wealth than all other agent types. The explanation for this is as simple as instructive: Although Maximin does not perform well in terms of average income, none

[9] This framework is in line with a modest learning rule. A new agent arrives at the scene and, in lack of better information, decides to adopt the behaviour of a successful, i.e. surviving, agent of the incumbent population. Compared with most accounts in the literature on epistemic game theory, (Easley and Kleinberg, 2010, chapter 7) the evolutionary mechanism described here is rather mild. It is not the absolute economic success of different strategies that determines their evolutionary success, but merely whether or not they generate sufficient funds for survival.

of its adopters performs poorly enough to have her wealth fall below zero in which case evolutionary pressure would kick in. For a moderate cost of living, it is precisely the maximin approach that guarantees short- and medium-term survival. All other strategies may perform much better in terms of wealth, yet there are still some that do not survive due to unlucky circumstances, such as trying a high toughness strategy yet having to give in eventually.

Adopting the Maximin strategy maximizes the chances of survival, yet fares poorly in terms of long term expected gains. Hence, there is an inherent tradeoff between two sets of goals a rational agent might pursue: survival and income maximization. To put this in a language of rationality: Different strategy choices can be rationalized by weighing these objectives differently or by using one as a boundary condition for the other.

For a high cost of living, between 1.2 and about 1.4, the share of surviving Maximin decreases abruptly and dramatically. At high costs of living, the moderate yet guaranteed income of a maximin strategy cannot provide sufficient means for survival anymore. The Maximin strategy guarantees survival only as long as the costs of living are not too high. If the environment is more hostile, survival requires risk-taking and the risk-free Maximin strategy does not flourish. At the same time, all other strategies remain constant or perform slightly better than they do for a medium cost of living. We mainly attribute this to the poor performance of Maximin: As Maximin dies out, other strategies occupy a higher share of the population. There is, however, one notable outlier. The Experimenter-type exhibits a sharp increase in population size at a cost of living around 1.4. This can be explained by the previous findings: Experimenters are more successful on average than all but the Maximax. They are therefore more likely to survive and reproduce compared with these others. Maximax, conversely, as the only strategy earning a higher average income than Experimenters, fall prey to their risk taking behaviour. This type is willing to take arbitrarily high risks, hoping eventually to make the opponent concede. While this strategy might be successful on average, it produces a high variance with longer stretches of close to no income. If such stretches grow too long or too frequent while the cost of living is substantial, the agents' wealth might temporarily turn negative, at which point the evolution mechanism kicks in.

3.3 Third Series of Experiments: Evolution in a Heterogeneous World

The previous experiment revealed an inherent conflict between two dimensions of rationality: survival and maximisation. Any choice of strategy needs to make a trade-off between these two, as none of the strategies studied so far fared well with regard to both dimensions simultaneously. Within a third set of experiments, we introduce a further dimension that could be responsible for long-lasting inequality: the initial wealth endowment of agents. For this, we distinguish two types of agents, *rich* and *poor*. We then inquire whether both types face similar problems of trading off survival and maximisation, or

whether the strategic dimensions in strategy choice are significantly different for rich and poor types.

This experiment is based on the same scenario as the previous simulation. However, before starting the simulation, we randomly select a proportion of agents labelled as *rich*. These agents receive a bonus of 50 on their initial wealth. Across different simulation runs, we vary the proportion of rich agents between 10% and 90%.

In terms of survival rates, poor agents display a similar behaviour compared with the agents studied in the second series of experiments. This is far from surprising as these agents have exactly the same starting conditions as in the previous experiment. In particular, at a moderate cost of living, Maximin have a significantly higher chance of survival than other poor agents, see left side of Figure 5. The mechanism behind this finding is the same as identified above: While other strategies might be superior in terms of long term gains, they carry the risk of falling below the poverty threshold of zero.

Rich agents, on the other hand, display a completely different evolutionary pattern. At moderate costs of living, all richer agents perform similarly well in terms of survival rates. Thus, the mechanism identified above does not apply to rich agents. Their prior endowment helps them to survive even moderately long streaks of low income unharmed. Rich agents do not face the same trade-off between maximising expected gains and ensuring survival as poor agents at moderate costs of living. Rich agents are thus less constrained in their choice of strategies. They have access to a set of potentially high gain strategies that might be inadmissible for poorer agents due to their risk of non-survival.

In terms of rationalisation, this pattern is reversed. For poorer agents, many different strategies could qualify as rational, depending on the trade-off between maximisation and chances of survival. In the extreme case, even the Maximin strategy can be understood as rational, for it maximises the rate of survival. This is not the case for rich agents. For these, the Maximin strategy is always suboptimal, as it performs poorly in terms of expected gains and does not give an edge in terms of survival.

A second finding concerns the case of a competitive cost of living between 1.0 and 1.4. In this case, rich players are also highly privileged compared with their poorer counterparts. Moreover, as Maximin perform poorly in this region, there is no safe strategy for poorer agents anymore. These cannot escape the increased evolutionary pressure and die out. Once again, there is one notable exception. Within a certain, narrow cost of living margin, the Experimenter strategy fares extremely well, as already noted above, indeed well enough to enable even poor Experimenters to have a sufficient chance of survival. Notably, this strategy is the most cognitively demanding, as it requires agents to assess expected utilities and balance between exploration and exploitation. In other words, poor agents may be able to survive in a competitive market only if they command sufficient cognitive resources.

We should highlight two consequences from these results. Firstly within the third set of simulations, the rich-poor gap widens compared with inequal-

Fig. 5 Evolutionary fitness of different types starting as poor (left) and rich (right)

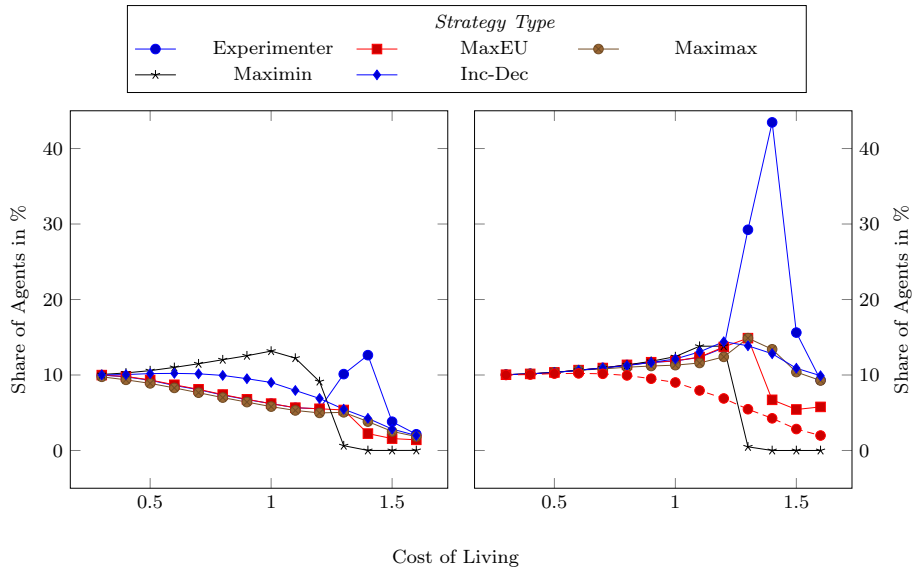
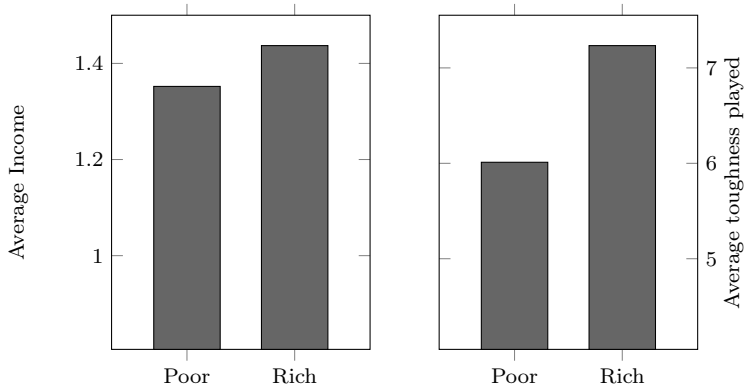


Fig. 6 Differences in wealth (left) and average toughness played (right) between rich and poor at the end of a simulation run



ity at the time of model's initialisation. While the average initial wealth gap between poor and rich is 50 points, this almost doubles to 95 at the end of the simulation - see the left side of Figure 6. As we have seen in the second experiment, evolutionary pressure among the poor agents affects everybody but the Maximin agents. Hence the latter will, in the long run, have a large share among the remaining poor agents. The first experiment, however, suggests that the Maximin strategy fares worst in terms of long term accumulated wealth. Hence, having a large share of Maximin is detrimental to the long term economic development of poor agents. No such reasoning applies to rich agents, as Maximin are not favored by evolutionary pressure there.

Secondly in the long run, rich agents develop a higher toughness on average than poor agents, see right hand side of Figure 6. With other words, richer agents have higher bargaining power and they are more likely to settle distributional quarrels in their favour. Notably, this development is fully endogeneous to the simulation, as both classes start with the same average toughness. The driving mechanism, again, is the different compositions of poor and rich classes in terms of strategies at the end of a simulation run. Having a larger share of Minimax agents with a toughness of 0 causes the class of poor agents to have a lower toughness on average. Taking the evolutionary mechanism as representing the learning process of rational agents, our simulation demonstrates how and why rich agents learn to negotiate tougher.

4 Rationality Revisited

There are three main findings from our analysis. First, contrary to the predictions of Rational Choice Theory (RCT), the MaxEU strategy is not optimal at maximising pay-offs in the long run. In short, the information collected while playing this strategy is structurally biased and hence utility calculations rely on false beliefs. Second, in some circumstances, the worst strategy in terms of expected utility might turn out to be the best for maximising chances of survival and vice versa. Third and finally, the considerations and boundary conditions for strategy choice might be structurally different for rich and poor agents. In this section we want to discuss some implications from these findings for the concept of rationality as it is used in RCT.

In economics, rationality is defined instrumentally as the ability to adopt the best actions to achieve given goals. Elster (1988), on the other hand, argues that this conception may be too narrow. He identifies three places where rationality plays a role in the concept of RCT:

1. The set of beliefs an agent holds has to be internally and externally consistent.[10]
2. The set of desires an agent holds has to be internally consistent.[11]
3. The desires and beliefs of an agent must cause an action in the right way and this action has to be the best choice given individual desires and beliefs.

Rational Choice Theory formalizes the latter property in an axiomatic way. From the set of all available alternatives, the agent chooses (or should choose) the strategy that maximizes expected utility. For decisions under risk, agents hold probabilistic beliefs about the possible consequences, allowing to calculate expected utilities. Agents compare the expected utility of the various available options in order to identify which action is optimal, given the agents' desires. Formally, the expected utility of an alternative a is defined as:

[10] For example, on the internal side, probabilities need to sum to one, impossible events need to receive a probability of zero and so on (Ramsey, 1931). On the external side, the agents beliefs have to respond to the available evidence in the right way.

[11] This means, for example, that the order of preferences has to be transitive, complete and connected Davidson et al. (1955).

$$EU(a) = \sum_j p_j \cdot u_j$$

where u_j stand for the utilities of various possible outcomes of action a, whilst p_j denotes their respective likelihoods. Agents can (or should) then compare different actions by their expected utility and choose the option that offers the highest expected utility. Classifying an agent as irrational could classically be interpreted as saying that she fails on one or more of the conditions (1), (2), and (3). Much of rational choice theory focusses on condition (3) exclusively, taking (1) and (2) as independent of (3) and given.

Additionally to condition (1), the amount of evidence available to an agent plays a role Spohn (2002). What beliefs are adequate for a given body of evidence might not only depend on the content of that evidence, but also on the quantity of evidence available. In some situations, rationality might require agents to accrue sufficient levels of evidence for their beliefs, if feasible. Collecting additional evidence, however, may come at a cost. Thus, the agent is faced with a strategic problem of how much evidence to acquire. A fortiori, the different facets of rationality may not be independent, but they can be intertwined in a complex manner Weisberg and Muldoon (2009). This shows in the present model, where the agents' information acquisition and their strategic choices are in a complex interdependency. Each move in a bargaining situation generates a new piece of evidence about how the opponent reacts. Thus, different bargaining strategies generate different flows of evidence.

Furthermore, the different strategies not only differ in the amount of information acquired, but also in its content. The agent's action indirectly influence her future beliefs and thus her future actions in a substantial and possibly non-neutral way. Hence, in choosing a strategy, an agent must not only make sure that she maximizes expected utility, given her current beliefs, but must also take care to ensure sufficient quality of her future beliefs. An agent may need to sacrifice short term income for creating an adequate evidential basis for future actions. As poorer agents may find themselves unable to make such short term sacrifices, this tradeoff may deepen existing inequalities. Only those who can afford to forego current income for collecting information can hope for long term optimality. Those who cannot may be structurally challenged in their quest for a better future.

We now discuss our main findings in light of Elster's thick concept of rationality. Our first result is that MaxEU fares structurally and significantly worse than other strategies. In focusing on short term maximization only, this strategy leads the agent to build up significantly false beliefs over time, thereby undermining the basis of its own success. The fact that MaxEU fares far from optimally suggests that a robust theory of rationality should encompass more than the single dimension of utility maximization. A second necessary dimension is constituted by rationality norms on the collection of evidence. Against the understanding of RCT that these dimension are independent of each other, our results gives evidence that information search and strategic behaviour are complexely intertwined. The choice of action impacts the amount and content of information collected, which in turn may influence future behaviour. In par-

ticular, agents might need to balance between maximising short term utility and collecting highly accurate information. A *thick* notion of rationality needs to take both these dimensions and their interplay into account.

The second finding sheds light on a further, often overlooked aspect of rationality. It invites the argument that long term maximisation goals may need to be balanced against or are constrained by short term needs created, for instance, by the accruing costs of daily life. It is no use having the highest expected utility in the long run if short term random fluctuations impede ever reaching that point. Or to put it in more mathematical language, it is not enough to consider the regularities from the law of large numbers; one must also consider the random irregularities of small numbers. Moreover, the current simulation suggests that such constraining factors might be unproportionally strong on the poor, who cannot afford higher risk levels or longer periods without income. By systematically constraining the poor in their choice of available long term strategies, side constraints may widen rather than close an existing wealth gap.

Finally, the third finding illustrates that strategies are not in themselves rational or irrational Galeazzi and Franke (2017). Rather, they depend on the context in which a player is situated, how other agents act and also the agent's endowment of information as well as resources. In the context of our simulation, certain strategies that are highly beneficial in the long run may be admissible for rich people, but carry too high a risk of failure for poorer agents. Conversely, what may be a defensible pick for poorer agents, sacrificing long term pay-offs in exchange for guaranteed short term income, might be completely unreasonable for richer agents not facing existential short term threats. Generally speaking, what is rational for one agent might not be rational for the next, even if both share the same goals and desires.

5 Conclusion

The findings of our simulation experiments contribute to two ongoing debates. The first concerns the nature of rationality as shown in section 4. By means of an agent-based model, we provide an example which illustrates that the standard conditions of rationality put forward by rational choice theory are too narrow for certain situations. In line with Elster (1988) we argue that a substantial account of rationality needs to address both principles of evidence collection as well as strategy and action choice. Both these dimensions are intricately connected.

Second, this paper contributes to a debate about the origins and structures of inequality. A recent body of work enquires into the various relations between inequality on the one hand, and rationality or cognitive resources and strategy choice on the other hand. With the present simulation we show that even within societies of rational agents, existing inequalities might persist and aggravate over time. The driving force here is that rich and poor agents may be subjected to different boundary conditions of rationality. These

findings directly relate to a variety of empirical results showing that poverty and inequality are correlated with a variety of distinct behavioural patterns concerning risk attitude Carvalho et al. (2016). Our results suggest that, at least prima facie, these differences need not always be driven by a failure to act and choose rationally. Rather, they might be indicative of rationality's context-sensitive requirements that affect the rich differently from the poor.

Finally, our results invite some general normative conclusions about inequality. Even in a society of equally endowed rational agents, spontaneous local variations can generate momentary inequalities. As we have shown, these inequalities tend to aggravate rather than diminish over time due to the interplay between information acquisition and strategy choice. Put simply, one can only succeed in bargaining over the long run if one dares to take risks from time to time. Moreover, one must be able to afford experiments and failure, as one would otherwise not have access to relevant information about the behaviour of other agents. Or, to quote a common saying: 'Those who never dare to achieve more will never know that they could have'.

6 Bibliography

Carvalho, L. S., Meier, S., and Wang, S. W. (2016). Poverty and economic decision-making: Evidence from changes in financial resources at payday. *The American economic review*, 106(2):260–284.

Dabla-Norris, M. E., Kochhar, M. K., Suphaphiphat, M. N., Ricka, M. F., and Tsounta, E. (2015). *Causes and consequences of income inequality: a global perspective*. International Monetary Fund.

Davidson, D., McKinsey, J. C. C., and Suppes, P. (1955). Outlines of a formal theory of value, i. *Philosophy of science*, 22(2):140–160.

Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Elster, J. (1988). *The Nature and Scope of Rational-Choice Explanation*, pages 51–65. Springer Netherlands, Dordrecht.

Galeazzi, P. and Franke, M. (2017). Smart representations: Rationality and evolution in a richer environment. *Philosophy of Science*, 84(3):544–573.

Gottschalk, P. and Smeeding, T. M. (2000). Empirical evidence on income inequality in industrialized countries. *Handbook of Income Distribution*, pages 261–307.

Hart, O. and Moore, J. (1999). Foundations of incomplete contracts. *The Review of Economic Studies*, 66(1):115–138.

Haughton, J. and Khandker, S. R. (2009). *Handbook on poverty + inequality*. World Bank Publications.

Lipman, B. L. (1986). Cooperation among egoists in prisoners' dilemma and chicken games. *Public Choice*, 51(3):315–331.

Mani, A., Mullainathan, S., Shafir, E., and Zhao, J. (2013). Poverty impedes cognitive function. *Science*, 341(6149):976–980.

McLeod, J., Lawler, E., and Schwalbe, M. (2014). *Handbook of the Social Psychology of Inequality*. Springer.

Mishra, S., Hing, L. S. S., and Lalumière, M. L. (2015). Inequality and risk-taking. *Evolutionary Psychology*, 13(3):1–11.

Nash, J. (1953). Two-person cooperative games. *Econometrica: Journal of the Econometric Society*, pages 128–140.

Neckerman, K. M. and Torche, F. (2007). Inequality: Causes and consequences. *Annu. Rev. Sociol.*, 33:335–357.

Piketty, T. (2014). Capital in the 21st century. *Cambridge: Harvard Uni*.

Ramsey, F. P. (1931). Truth and probability (1926). *The foundations of mathematics and other logical essays*, pages 156–198.

Sheehy-Skeffington, J. and Rea, J. (2017). *How poverty affects people's decision-making processes*. www.jrf.org.uk.

Skyrms, B. (2014). *Evolution of the social contract*. Cambridge University Press.

Spohn, W. (2002). *The many facets of the theory of rationality*. Bibliothek der Universität Konstanz.

Weisberg, M. and Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of Science*, 76(2):225–252.

Wood, A. (1995). *North-South trade, employment, and inequality: Changing fortunes in a skill-driven world*. Oxford University Press.

Paper 3

When do Groups get it right?

–

On the Epistemic Performance of Voting and Deliberation

This is article has been published in *Historical Social Research (HSR) – Special Issue: Agent-based Modelling across Social Science and Philosophy.* (ISSN: 0172-6404).

The suggested citation is:

Scheller, Simon (2018). When do Groups get it right? – On the Epistemic Performance of Voting and Deliberation. *Historical Social Research* 43(1): 89-109.

When do Groups Get it Right? – On the Epistemic Performance of Voting and Deliberation

Simon Scheller *

Abstract: This paper examines the claim that democratic decision making is epistemically valuable. Focussing on communication and voting, circumstances are identified under which groups are able to reliably identify the 'correct alternative'. Employing formal models from social epistemology, group performance under varying conditions in a simple epistemic task is scrutinized. Simulation results show that larger majority requirements can favour the veto power of closed-minded individuals, but can also increase precision in well-functioning groups. Reasonable scepticism against other people's opinions can provide a useful impediment to overly quick convergence onto a false consensus when independent information acquisition is possible.

Keywords: Deliberation, voting, agent-based modeling, group decision making, bounded confidence, social epistemology.

1. Introduction

Group decisions are a central element of social interaction. Be it in political assemblies, company boards or informal groups: A large number of decisions are not taken by single individuals, but by groups. It is thereby often claimed that such group decisions should be made *democratically*, which is mostly justified on the basis of two distinct arguments. The first argument concerns the *fairness* of democratic procedures: Democratic voting is claimed to be the only way to aggregate individual preferences fairly, since it gives equal weight to individual preferences. However, following Arrow (2012), social choice theory has extensively examined this claim and by and large found that all possible voting rules have substantial inherent flaws (see e.g. Riker 1982).

The second argument (which is also at the focus of this paper) concerns the *epistemic quality* of democratic decisions. Going back to Rousseau's idea of the 'general will' (Rousseau 1964) and following Habermas' deliberative ideal (Habermas 1996), deliberative democrats have frequently claimed that democratic decision making results in 'better' decisions: "The decision of majorities about which policies to pursue can provide good evidence about which policies

* simon.scheller@uni-bamberg.de.

are in fact best", (Cohen 1986, 34) – given that public deliberation is guided by the right principles.

There are two ways of substantiating this claim. On the one hand, *deliberation* facilitates the transmission of information: Individuals can improve their knowledge of an issue by talking to each other which, in turn, improves the group's decision (Elster 1998, 11). On the other hand, the epistemic value of democracy can be located in *voting*: As Condorcet's jury theorem illustrates, larger groups perform better under majority rule – as long as individual knowledge is sufficiently independent and individuals have a better than random-chance of being correct (de Condorcet 1785). Yet, both arguments are not without criticism. Effective information transmission is hindered by a variety of psychological biases, cascading phenomena or polarization. For Condorcet's jury theorem, the assumption of independence is usually not fulfilled, which renders the theorem broadly without real application.

The goal of this paper is to scrutinize those arguments by means of an agent-based model, in which a group of agents faces the simple, epistemic task of making a choice from a set of discrete options. The stylized process of democratic decision making incorporates communication under bounded confidence (Hegselmann and Krause 2002), and voting in the form of majority voting with varying majority thresholds. In an additional model version, agents can also acquire new evidence as an alternative to communication.

Insightful results are obtained: First, while large majority requirements may increase the chances of correct decisions when communication is functioning, it also increases the chances of gridlock when people are closed-minded about an issue. Second, when external sources of information are available, a certain degree of scepticism against other people's views impedes overly quick convergence onto a false consensus and can therefore improve the group's epistemic performance. This corroborates findings by Zollman (2010). On a methodological dimension, the paper makes a contribution by introducing models from social epistemology to the context of democratic decision making. Overall, the model supports advocating a multidimensional view on democracy, showing how deliberation and voting can be combined efficiently.

To lay out all those points, I contextualize the project by outlining the relevant theories and literature in *section 2*, summarizing the main arguments from democratic theory, previous theoretical and empirical findings on the impact of democratic mechanisms, as well as formal models from social epistemology. *Section 3* describes the basic model of democratic decision making as communication and voting. In a modified version of the model (*section 4*), agents also have the alternative of acquiring independent information. *Section 5* concludes by showing how these findings can lead to an augmented understanding of democratic decision making institutions, and how they support the epistemic democrat's claim that democratic procedures result in epistemically superior outcomes.

2. Theory and Literature

2.1 The Epistemic Value of Democracy

Talking about the epistemic quality of group decisions presupposes the existence of some sort of 'objective truth'. As List and Goodin argue,

> [t]he hallmark of the epistemic approach, in all its forms, is its fundamental premise that there exists some procedure-independent fact of the matter as to what the best or right outcome is. (List and Goodin 2001, 4)

While this supposition is frequently discussed, I continue under the assumption that matters with an objective truth exist, and others where the assumption is infeasible. Duggan and Martinelli (2001, 260) provide illustrative cases for such truth-issues: a jury deciding about the guilt or innocence of a defendant, or a group of doctors deciding about the best treatment for a patient. In both examples, it is clear that all participants share a common goal, and they discuss the best way of achieving said goal. Thus, the arguments of this paper apply to problems that can be reasonably seen as epistemic problems; and there exist enough relevant problems of this sort to render this discussion relevant.

According to Estlund (2009), epistemic quality of decisions must necessarily be guaranteed in order to legitimize democratic procedures, which is why it is essential to establish the epistemic potency of groups. In political philosophy however, discussions regarding the merits of democratic decisions is often solely a normative endeavour, which is, for example, expressed by List and Goodin: "[a] pure epistemic approach tells us that our social decision rules *ought be chosen* so as to track [the] truth." (List and Goodin 2001, 4, my emphasis)

Regarding deliberation, a range of scholars have provided criteria for optimal deliberative procedures: Habermas' ideal speech situation (Habermas 1996), Rawls' 'veil of ignorance' (Rawls 2009), or also Estlund's "imaginary model epistemic deliberation" (Estlund 2009, 175). While these approaches somewhat differ in primary focus and intention, they all prescribe an ideal deliberative procedure that produces desired outcomes if correctly employed.

The same is true for claims regarding the epistemic quality of voting. Condorcet's jury theorem (de Condorcet 1785) provides the key argument: Groups are more likely to make correct decisions under majority rule when the group becomes larger – assuming that individuals are more likely to be right than wrong, and that their individual judgments are probabilistically independent from each other (see e.g. Estlund 2009, 15 for a more detailed description and discussion). However, the theorem is largely dependent upon the fact that the assumptions which it requires are actually fulfilled in a certain situation. If, for example, individual guesses are not independent, the theorem does not say anything about the truth-tracking merits of majority voting.

On that basis, the rest of this paper scrutinizes *deliberation* as a tool of information transmission among groups, and *voting* as a way to aggregate information independently – with regards to their potential of finding better decisions. Doubtlessly, both procedures exhibit a large array of other merits. The work by Elster (1998) constitutes a good starting point for a more extensive overview. However, for the sake of conceptual focus, this stylized description of democratic procedures and the epistemic democrat's claim will constitute the basis of discussion for this paper.

At the same time, there are various alternative ways of exchanging information apart from direct communication. One crucial advantage of exchanging information directly via group communication, however, is that it combines an efficient way of connecting multiple actors as sources and receivers of messages simultaneously. Thus, having a public deliberative forum[1] facilitates the exchange of information in a uniquely efficient way. Apart from its efficiency, a public deliberative forum also provides the immediate opportunity of being subject to a potentially balanced input from all actors. These points motivate the special interest in this form of democratic information exchange.

2.2 Pitfalls and Problems of Deliberative Democracy

In contrast to these optimistic claims about the merits of democratic decision making, there is also a list of problems and pitfalls of deliberation and voting. Being aware of those is a necessary prerequisite to appropriately model democratic procedures on the basis of empirical findings. The following list gives a brief overview of some major issues with group processes.

- *Persuasion bias:* When group members exchange opinions[2], what weight should one give to other people's opinions, given that information is transferred between multiple individuals along complex channels. Certain pieces of information may therefore receive disproportional attention (Golub and Jackson 2010, 2). As a result, a person's position in a social network – and not just the quality of her information – determines her influence on the group's aggregate beliefs (DeMarzo et al. 2003).
- *The common knowledge effect:* Information that is available to a large number of people is more likely to be accepted, discussed and emphasized than less commonly available information (Gigone and Hastie 1993). This potentially leads to a homogenization of opinions, while uncommon opinions are more likely to be dismissed.

[1] In this context, this does not necessarily have to be in the form of public, personal group deliberation, but can also amount to other forms of exchange.

[2] The terms 'opinion' and 'beliefs' are used interchangeably in the context of this paper since it is explicitly stated that only epistemic matters are addressed here – and purely preferential questions are not, even if some of the described models are usually situated in preferential settings.

- *The social comparison effect:* People desire to be accepted and liked by other individuals in a social group, which some aim to achieve by taking similar opinions like one's peers. Similar to the previous heuristics, social comparison usually results in opinion convergence (Sunstein 2002, 179).
- *Homophily:* People tend to associate more often with people who are like themselves: Two people of the same ethnic background are more likely to get married; Republicans are more likely to exchange opinions with Republicans as opposed to Democrats. On aggregate, social groups homogenize (Lazarsfeld and Merton 1954). This may have a strong impact on the information one receives (McPherson et al. 2001) and even trigger a self-reinforcing process where people take their opinion itself as a selector for communication partners, and thus self-affirming their own positions.
- *The persuasive argument effect*: Due to her limited capacities, an individual's opinion is based on only a fraction of all available arguments. Since people holding similar sets of arguments can be expected to group around certain positions on the opinion spectrum (homophily), people will frequently be subject to arguments that are supportive of their current position, and much less to arguments that would run counter to their overall view. Hence, selective availability of information from like-minded people produces a confirmatory bias that makes people overconfident in their own opinion (Sunstein 2002, 179).

The described effects constitute a serious threat to the potential merits of deliberation: Contrary to what the arguments by theorists of deliberative democracy suggest, irrational biases and rationally justifiable pitfalls can hinder the efficient aggregation of information. Especially opinion polarization can severely undermine the finding of a rational consensus. One undermining factor is the occurrence of 'enclave deliberation' (Sunstein 2006, 186): Within homogeneous, isolated subgroups of people commonly held persuasive arguments increase the members' convictions. Between groups, less interaction takes place. This further polarizes opinions and also the information and arguments people hold and thereby hinders the efficient use of information for group decisions.

Central for the later model description, the essence of the described effects and biases is captured by the bounded confidence model (Hegselmann and Krause 2002, 2006). Their formal model of opinion dynamics is also able to reproduce the described effects of polarization and fragmentation of opinions in groups and requires only very sparse assumptions to do so.

2.3 Literature Review

Looking at the overall picture, positive claims about deliberation and voting are contrasted with serious shortfalls and problems. Assessing the functionality of these tools as democratic mechanisms, scholars have employed formal models

and empirical analyses. Economists have studied the interplay between communication and voting, yet with a focus on problems of preference aggregation and information revelation, and hence not for purely epistemic problems (see e.g. Austen-Smith and Feddersen 2006; Doraszelski et al. 2003). Gerardi and Yariv (2007), studying an epistemic group decision problem, find that prevoting deliberation renders a large class of voting rules equivalent, as long as they are veto-free. While those game theoretic models are able to identify equilibria for a variety of conditions, they capture interaction as a strategic process between fully rational actors. Realising the empirical inadequacy of this supposition, the model in this paper departs from the assumption of perfect rationality.

Goeree and Yariv (2011) find experimental evidence that most voting rules are rendered equivalent when communication takes place. They observe that deliberation uniformly improves efficiency and also diminishes the impact of institutional rules significantly. Interestingly, Guarnaschelli et al. (2000) find evidence for strategic voting under unanimity rule. More generally, empirical evidence on the presence and impact of deliberation is difficult to collect as strategic communication (bargaining) and non-strategic communication (arguing) are in practice hard to distinguish, specifically since they usually occur together (Bächtiger and Wyss 2013, 159). In their survey on empirical deliberation studies, Bächtiger and Wyss (2013) identify a variety of contributions that aim at measuring the occurrence of deliberative behaviour. Yet, when looking for effects of deliberation on the epistemic quality of outcomes in group deliberation, the authors report a general lack of studies (Bächtiger and Wyss 2013, 175). Similarly, Bozbay et al. (2014) report a lack of theoretical works on epistemic problems:

> So far, however, [… the literature] has paid only little attention to a different 'epistemic' approach of aiming to track the truth, i.e., reach true group judgments. The theory does not model the private information underlying voters' judgments, thereby preventing itself from studying questions of efficient information aggregation. Yet such an epistemic perspective seems particularly natural in the context of aggregating judgments (rather than preferences). (Bozbay et al. 2014, 2)

As this short (and far from all-encompassing) literature review suggests, a large potential exists for studies on judgment aggregation and truth-tracking in the context of democratic theory. Interestingly enough, such epistemic questions have been taken up by philosophers under the label of 'social epistemology': Scholars of this field envisage science as a process of social knowledge creation, circling around the question of "whether we ought to let our opinions be guided, even if only partly, by those of others." (Douven and Riegler 2009, 326). This question is equally central for considerations in the context of democratic theory.

Scholars of social epistemology frequently employ *formal models* to study truth-tracking capacities of different network structures, information exchange procedures or group constellations. Those formal, often agent-based simulation models have the advantage of providing tools to study concrete procedures and mechanisms in a stylized fashion. When it comes to evaluating democratic procedures, previous literature on democratic theory can merely hope for the merits of deliberation to overcome the shortfalls of social influence. Formal models, in contrast, can help to actually identify circumstances under which groups function well in epistemic tasks, and how certain arguments counterbalance each other.

Muldoon (2013) summarizes the main streams in the literature on social epistemology, of which some will be briefly touched upon here. Weisberg and Muldoon (2009) model science's search for truth as a search for the 'highest points' on an 'epistemic landscape'. Scientists can explore unknown territory or follow others in order to reach states of 'higher' knowledge, while they are driven by a self-interest to be credited for good results. This choice between exploring or following could be translated into a choice between researching or communicating, which informs the extended model in this paper (section 4).

Others have focused on the role and emergence of consensus in science and society. A baseline model of consensus formation has been presented by Lehrer and Wagner (1981). Although not initially intended as such, their model was later interpreted as depicting communication among scientists as an iterated, weighted updating procedure of individual opinions. Each individual takes into account every other individual's opinion with a certain non-zero weight. Simultaneous and repeated updating can be shown to result in convergence to consensus. French Jr (1956) and DeGroot (1974) describe similar dynamic models.

Zollman (2010) introduces network structures for modelling communication in scientific communities, where agents update their beliefs in a Bayesian fashion. Surprisingly, he finds that fewer connections between agents can be beneficial for reaching the correct solution as this prevents overly quick convergence onto a potentially false consensus, which in turn allows for broader diversity in exploring alternative possibilities.

Hegselmann and Krause (2002) describe a model of opinion dynamics, in which agents communicate by averaging their opinion and those of others. However, they average only with those others that have sufficiently similar opinions than themselves. The psychological effects from above provide an empirical foundation for the model formalization that was chosen: The fact that people only talk to like-minded others can be seen as a clear instance of homophilious behaviour or resulting from social comparison. As information is exchanged more frequently among like-minded people, they will be subject to the common knowledge effect, persuasion bias and the persuasive argument effect alike. Further, agents update opinions by simple myopic averaging over all opinions they consider. Thus, modelling behaviour in this way incorporates

the empirically found biases from above, especially when it comes to the agent's limited processing capacities and the described affective, non-rational behavioural patterns.

In a model extension, some agents are 'attracted' by the true value of the parameter they seek to identify, while regular agents still solely 'follow' others in their opinion (Hegselmann and Krause 2006). Notably, a relatively small number of truth-seekers is sufficient for the convergence of the group onto the correct consensus. A feature similar to this extension will also be part of the second model version to be presented below.

While those models have been employed for answering questions about convergence on true scientific knowledge in society, my goal in this paper is to apply such model structures to questions of democratic theory and of group decision making. This lays the ground for studying the efficiency and effectiveness of different decision making rules and communication schemes in producing correct outcomes.

3. The Baseline Model

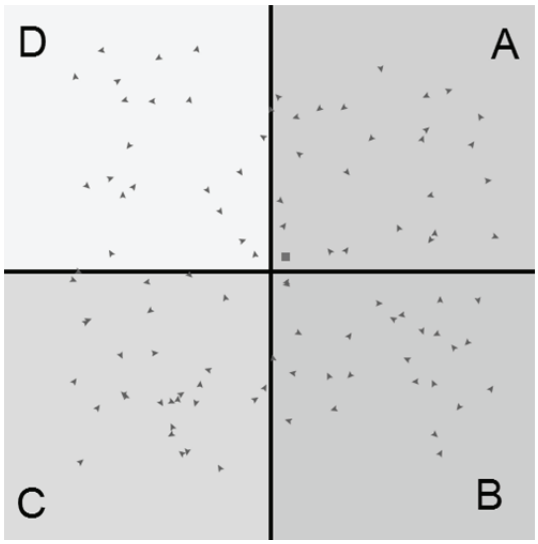
3.1 Model Description

The model at hand is based on a simplistic epistemic problem: A group of agents faces a choice between four options *A*, *B*, *C* and *D*. These options are graphically represented by the four quadrants of a coordinate system. Further, an optimal point *O* exists with the coordinates (x_o, y_o), which lies in one of the four quadrants. The task of the group is to find out in which of the quadrants the optimal point lies.

Each agent receives an independent signal for the coordinates of *O*. The signal for the x-coordinate is given by a random draw from the uniform distribution from the interval $[x_o - 50; x_o + 50]$. Analogously, the signal for the y-coordinate is given by a random draw from the uniform distribution from the interval $[y_o - 50; y_o + 50]$. Thus, a position on the two-dimensional plain with the four quadrants can be ascribed to each agent, enabling a neat graphical representation of the epistemic problem. The agent's position thereby represents her best guess for the true value of *O*.

For the subsequent simulations, the optimal point is fixed at (*3, 3*) in order to keep the epistemic problem's difficulty and each individual's primary chance of making a correct judgment constant. The model screenshot in fig. 1 depicts an exemplary random distribution of 100 agents on the opinion space. The optimal point *O* at (*3, 3*) is marked by the grey square in the bottom left corner of the top right quadrant *A*.

Figure 1: Graphic Illustration of the Underlying Epistemic Problem



An intuitive interpretation of such a problem would be a choice between policies, the utility of which is determined by two separate utility dimensions with linearly decreasing marginal utility. For example, imagine a city council that needs to evaluate different project proposals for a railway extension with regards to the dimensions cost and environmental impact, and needs to choose one of those discrete options. Each option must be evaluated regarding both dimensions, and the people need to judge what combination provides an optimal solution.

At the beginning of each round, all agents take a vote on the four discrete options. Each agent votes for the option she considers best, i.e. the quadrant she is located in. An agent's position, in turn, is based on all the signals she has previously had access to, and agents are assumed to vote solely based on that opinion without any strategic or other considerations.[3] As soon as one of the options reaches the necessary quorum of votes, this option is considered the group's choice.[4]

[3] Austen-Smith and Banks (1996) challenge the assumption that each agent simply votes for her preferred option as irrational under certain circumstances. This shall, however, be of no further concern in this paper, and a myopic and heuristic-based decision behaviour is prescribed for the agents, as for instance in Zollman (2010, 2013), and most importantly because it is considered the more realistic behavioural assumption for epistemic questions.

[4] If more than two options reach the quorum at the same time (which is only possible if no absolute majority is required), the option with most votes is chosen. If two or more quorum-reaching alternatives tie, the winner is picked randomly between these options. These special cases are very and have no impact on the model analysis.

If no qualified majority is found in a vote, agents have the possibility to communicate. They do this by 'proclaiming' their current position to all other agents. However, communication works under a setting of bounded confidence (Hegselmann and Krause 2002): Agents listen only to other agents which have beliefs that are similar enough to their own beliefs. More specifically, an agent listens to another agent if the distance between them is smaller than ε. Technically, the parameter ε describes a circle-shaped area with radius ε around an agent's position. All agents within this area are considered for opinion updating, all agents outside of this circle are ignored. A small ε-value therefore means that an agent listens only to others with very similar opinions, while a large ε describes an agent that is much more open for contrasting or contradictory information.

Updating occurs by simple averaging over all considered agents, including the agent itself. All communication and updating happens simultaneously. Each communication round is followed by another vote. This iterated process continues until the majority threshold is reached, or is terminated if no more communication occurs, i.e. if agents either have the identical positions or do not listen to each other. Failure to reach the necessary quorum when communication breaks down is counted as an incorrect decision. As soon as a decision is made, new random initial signals are allocated and the process starts anew. Thus, all individual decisions are independent from each other. This completes the description of the baseline model.

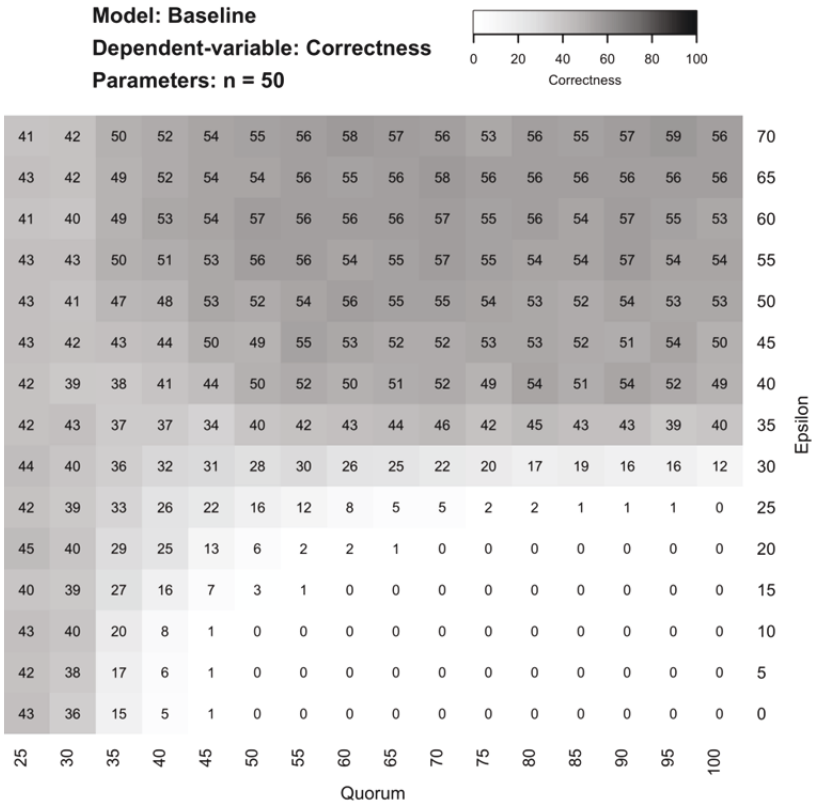
3.2 Results

This first analysis incorporates two parameters which supposedly influence the epistemic performance of groups. These are the *quorum*, i.e. the majority threshold that is required for a decision to be made, and ε, which describes how open individual agents are for divergent opinions. These parameters constitute the independent variables of the experiment. For each parameter constellation, 1.000 individual decision problems were simulated, each decision problem with a randomly assigned distribution of initial signals. As argued, the optimal point O remains the same for all problems in order to guarantee comparability.

Epistemic group performance is measured by the probability of making a correct decision (i.e. to choose option A) under a given parametrisation, labelled *correctness*. To evaluate efficiency of the procedures, the variable *time* measures the average number of communication rounds until a decision is made – regardless of whether the decision is right or wrong. *Correctness* and time constitute the dependent variables of the experiment. The interest of the analysis therefore lies on how ε and the *quorum* influence *correctness* and *time*. In figure 2, the *quorum* is depicted on the x-axis; ε is located on the y-axis.

Correctness is displayed in the resulting grid and illustrated by color-coding. Figure 2 shows results for a group of 50 agents.[5]

Figure 2: Correctness Results in the Baseline Model

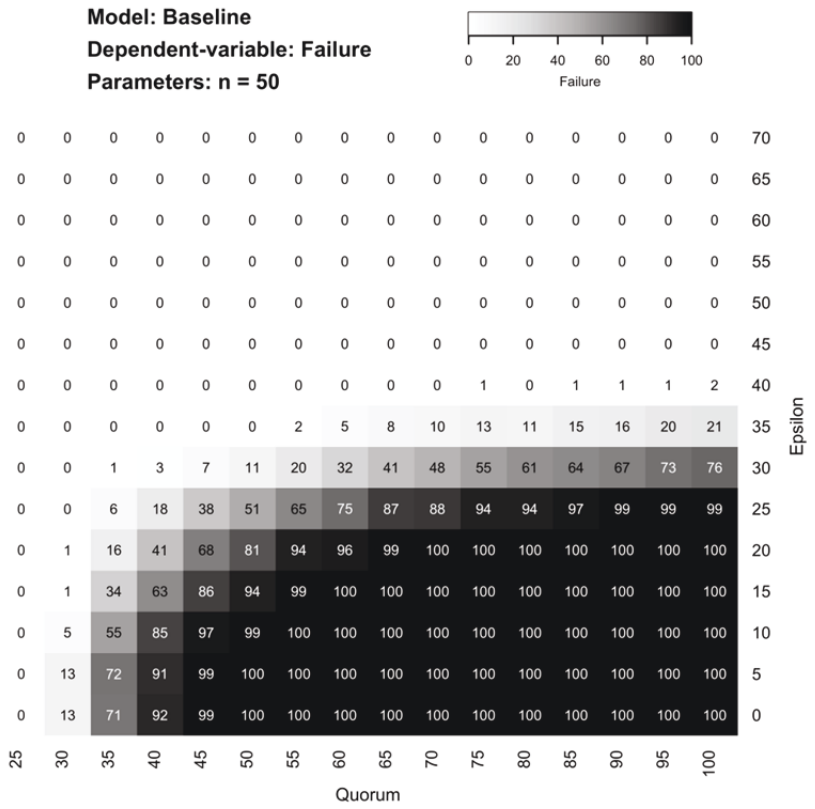


The most striking feature of figure 2 is twofold: First, no correct decisions are reached for small ε and a large enough *quorum* (bottom right corner). As it turns out, this stems from failures to reach a decision (see fig. 3): For small ε, it is very likely that agents do not communicate with each other since they are

[5] In a series of robustness checks, increasing the number of agents significantly increased correctness, and decreased correctness for smaller groups. This corroborates classic jury-theorem findings, since more agents offer a larger pool of information, individual errors are more likely to cancel out, and thus more people increase the likelihood of a correct group decision. For the analysis of this paper, detailed results regarding the impact of group size were omitted for the sake of focus and simplicity.

outside each other's confidence bounds. Thus, if the necessary quorum is not reached already at the outset, the lack of communication prevents the formation of the necessary majority. The larger the required majority, the more likely this scenario becomes.

Figure 3: Failure Results in the Baseline Model

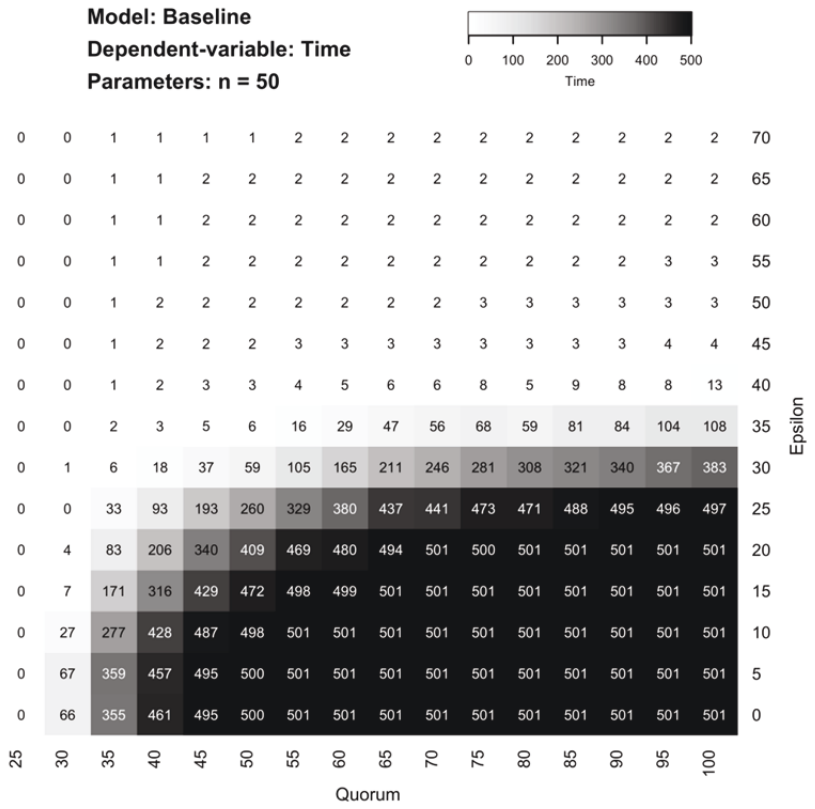


Second, apart from the parameter area where the group fails to make a decision, ε appears to have a slight positive effect on *correctness*. Unbiased communication, in this scenario, seems to be at least not worse than being sceptical against alternative opinions.

The impact of *quorum*-size is similarly small, yet a clear and strong positive impact occurs when the quorum is increased from 30% to 35%. Here, the quorum enables communication to happen in the first place, since for a 30%-threshold, a decision is usually already found in the very first vote and hence without any communication taking place. A large enough *quorum* thus forces people to communicate. Yet, this can work only if people are also willing to

communicate (e.g. if ε is large enough). For small ε-values, the larger quorum causes failure to reach a majority. Even larger *quora* do not seem to make any further difference for *correctness*, and only increase the chances of gridlock.

Figure 4: Time Results in the Baseline Model



Decisions are found very quickly for large and very small ε-values (see figure 4): For large ε, discussion is very efficient and does not take much time. For very small ε, no communication occurs anyway. Thus, either a decision is found relatively quickly – or not at all. Only for intermediate values of ε, decisions take up to 11 communication rounds. This behaviour stems from the basic mechanics of bounded-confidence updating process: Opinion convergence is slowest when ε-intervals are large enough to connect some agents, yet small enough not to connect too many agents at the same time. Then, as also described by Hegselmann and Krause (2002), opinions converge over multiple steps. The more they converge, the more likely it becomes that the necessary

quorum is reached, and, of course, the larger the *quorum*, the more convergence is required.

Unfortunately, longer discussions do not promote better decisions in this model. For those intermediate ε-parametrizations, no decision is reached even after this longer time. The intermediate values of ε lead to polarization between only a few points of attraction, which explains why communication takes longer. It also explains why there is no success eventually: When a small number of clusters forms, one of them needs to be much larger than the rest so as to attract enough agents for the quorum to be reached. The larger the quorum, the less likely this becomes. Thus, looking at the interplay of *time* and *correctness* implies that there is not really a trade-off between quicker yet less precise or slower yet more precise decision mechanisms in this version of the model. This is due to the specific dynamics of the bounded confidence process, which either leads to culmination in one point rather quickly – or not at all. Interestingly, this replicates findings by Golub and Jackson (2010). Their model is based on network structures that are subject to naive opinion updating on the basis of DeGroot (1974). The authors equally report that convergence time and whether or not the group converges onto the correct solution are usually independent.

4. The Research-Talk Model

4.1 Model Description and Background

In the baseline model, agents change their views only through communication. As has become obvious in the analysis, the time dimension could not be reasonably interpreted. One major reason for this is that communication has no opportunity cost, since time cannot be spent any other way. This is clearly not a realistic depiction of decision making in reality. A logical alternative choice of action is suggested once again by the theory of social epistemology, which models the scientific process as communication and *individual information acquisition by independent research* (see e.g. Hegselmann and Krause 2006; Weisberg and Muldoon 2009; Zollman 2010). The idea translates easily to the context of democratic decision making: Alternative to receiving information from others, people also have the possibility to collect information independently. This setting allows comparing whether time should be better spent researching or communicating. In practice, independent information acquisition can refer to collecting data, visiting the location of an infrastructure project, or simply researching a subject on the internet or by other sources. For the formal model, it is deliberately left open what specific actions 'researching' adheres to.

Formally, whether an agent communicates or researches is decided probabilistically before each round. The probability that the agent chooses to research is

given by the parameter *pr*, which is externally set and equal for each agent. If, for example, $pr = 0.3$, an agent will choose the option research with a probability of 30%, and communicate with a probability of 70%. If $pr = 0$, all agents will communicate all the time, which is equivalent to the baseline model. If $pr = 1$, all agents will perform research all the time and never communicate. The random choices of individual agents are independent from each other.

When an agent researches, the distance between her current 'best guess' and the true optimal point O is reduced by one percent[6]. In other words, researching moves the agent's position one percent closer to the true value. This is in close analogy to Hegselmann and Krause (2006), who capture researching by the parameter α.

The communication process itself remains unchanged. However, note that researching agents are excluded from the communication process for the round in which they perform research. Other agents do not receive signals about their positions and can, hence, not be considered by communicating agents.[7]

Further, a new criterion for decision-failure must be provided, since infinite researching would potentially stretch the decision process tremendously. This not only exhausts computational resources quickly, but is also unrealistic in the context of democratic decision making: After a certain time, groups can be expected to terminate a decision process when no agreement is found. A similar argument for process termination is employed by Zollman (2010, 31). I choose 500 rounds as the limit after which a decision process counts as failed.[8]

4.2 Results

Figures 5 and 6 display results for the extended model in the same way as before: The x-axis displays the *quorum*, the y-axis ε. To study the impact of the new parameter *pr*, a series of these plots for various values of *pr* is presented. For the analysis, I focus on the most interesting similarities and differences to the reference model without research.

---

[6] Robustness checks have shown that manipulating this one-percent value changes merely the relative strength of certain effects and for what parameter regions they occur, but it does not qualitatively alter any of the core results.

[7] A similar probabilistic setup was proposed by Douven and Riegler (2009) as an extension to the bounded-confidence model. Here, the model diverges from the specification by Hegselmann and Krause (2006), in which agents take a weighted average between researched and communicated information. Arguably, it is more realistic when an agent has to pick how to spend her time and therefore makes a choice between two discrete options. This may potentially also have interesting effects on the outcome since certain agents are temporarily excluded from the communication process.

[8] The impact of the decision time limit plays a major role for the results, and has been extensively analysed as part of the robustness analysis. While severe quantitative shifts in the results do occur, the quality of the results remains the same as long as they are not overruled by predictable effects of a significantly shorter or longer time limit.

For small probabilities of research and high enough quora, the parameter area where the group fails to find agreement remains. However, failure can be eliminated by increasing *pr*, as shown in fig. 5: Already for *pr* = 0.6, gridlock hardly occurs. This, however, comes with a cost: As fig. 6 shows, time strongly increases with larger *pr* and when ε is small. Had the maximum *decision-time* been lower, failures would occur more often.

Thus, a high research probability is beneficial if ε is small – a result that makes intuitive sense: Collecting one's own evidence is superior when communication does not work. Put the other way round: When ε is too small for agents to be connected, individual research can help to bridge the gap.

Figure 5: Correctness Results in the Research-Talk Model for Varying *pr*

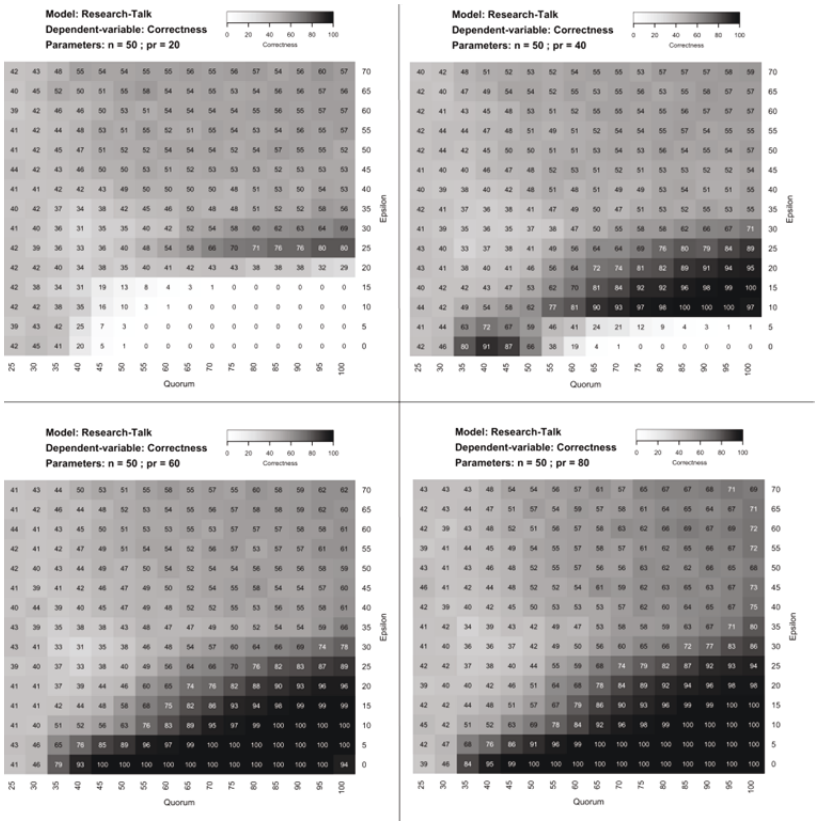
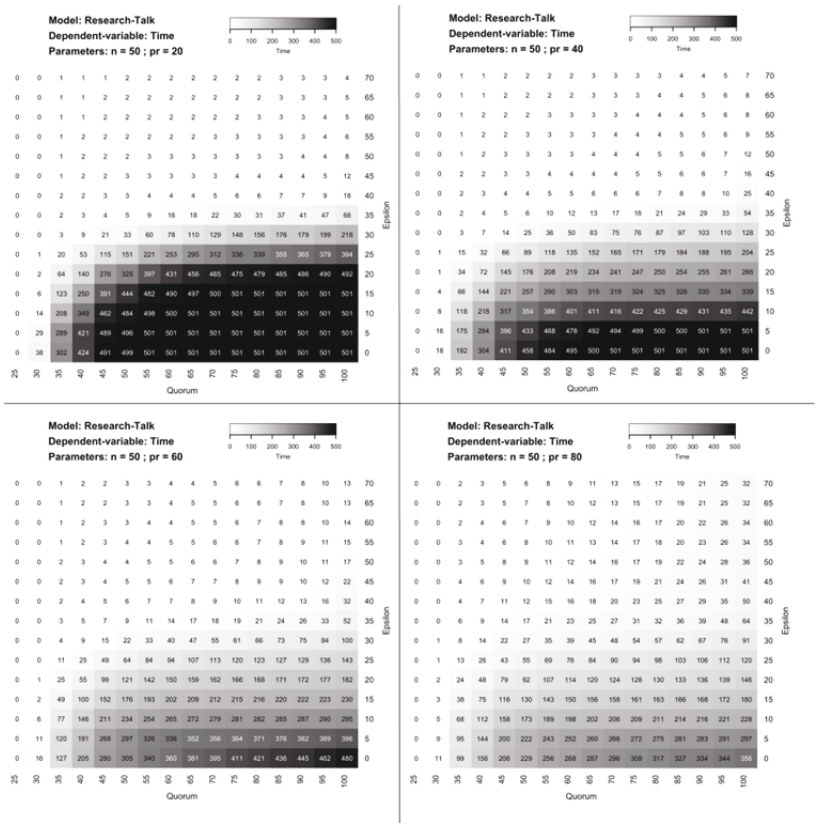


Figure 6: Decision-Time in the Research-Talk Model for $pr = 40$



The simulation also shows that the group performs better for moderate ε-values compared to both small and large values of ε if only some research is possible ($pr < 0.4$). How is this possible? I argue that this replicates findings by Zollman (2010) that less densely connected epistemic communities exhibit a better performance since they do not converge onto a 'false consensus' too quickly but give the group enough time to gather independent information. For larger ε, a rash decision would be made before agents had enough time to carry out sufficient independent research. For smaller ε, time becomes a strong determinant factor, in some cases leading to failure to find agreement, and generally rendering the process much less efficient. Too much independent research also slows down the process indirectly: Since fewer agents enter the communication arena, and the larger gaps between fewer agents reduces the amount of communication that takes place in a given round. In conclusion, a reasonable degree of scepticism against socially gathered information impedes herding ef-

fects and information cascades, while still being able to benefit from other people's input.

While the previous considerations are mostly applicable to large quora, more research is a feasible tool for smaller majority requirements. For example, pure research already performs really well when a 40% threshold is required. A focus on communication is therefore more efficient if large quora must be met. For lower requirements, more researching can result in a better performance.

In summary, the extended model enables studying the trade-off between correctness and time systematically and implies the following conclusions: Well-functioning and unbiased *communication* provides an efficient epistemic tool. Yet, it can be prone to convergence onto a false consensus if too little independent information is fed into the process. *Independent researching*, in contrast, makes the epistemic process more precise at the cost of slowing down the decision making process. Stricter majority requirements can make a process more reliable, but also slower and more failure-prone. When a large threshold-level is externally given, communication provides an efficient means for meeting such a demanding requirement. In a larger context, these findings justify a perspective on democracy that argues for a multi-faceted view on democratic procedures: While combining the beneficial effects of voting and deliberation as described by the Condorcet Jury Theorem and deliberative democrats is generally possible, the model analysis illustrates that voting and talking cannot be combined arbitrarily.

5. Conclusion

Deliberation and voting play a central role in arguing for the epistemic quality of democratic decision making processes. Yet, empirical findings suggest broad potential for various pitfalls in the processes. The described agent-based model provides a vehicle that allows for a more balanced view on the merits and problems of voting and deliberation, and to disentangle interaction effect between the two. This paper thereby provides a multi-faceted perspective on democratic decision making, combining seminal works from political philosophy, psychology and social epistemology.

By studying the interplay between voting rules and communication structures, insightful perspectives on democratic decision making are obtained. Consider for instance the impact of majority requirements: Unanimity is often argued for on the basis that nobody's opinion can be overruled. This might grant it the attribute of inclusiveness. Quite to the contrary, the model analysis highlights that large majority requirements lend strong veto powers to closed-minded individuals with extreme views. As long as people are open to communicating with each other, large quora can augment a group's epistemic capacities. However, if people's opinions polarize as a result of a lack of open-

ness, strict majority requirements can also lead to gridlock and standstill by granting a veto power to closed-minded people.

Does this mean that scepticism against other opinions is generally bad? As has been shown, moderate openness to communication can be superior to too much or too little openness by preventing the group from converging onto a false consensus too quickly. A balanced mix between independent information collection and dissemination can thus impede the occurrence of group think phenomena, information cascades and other herding tendencies. This corroborates findings by Zollman (2010), who finds similar results with regards to less densely connected communication networks. For this to work, however, reliable external sources of information must be available.

Taking such insights into account when choosing problem specific decision making rules makes for better institutional designs. This paper's analysis can thus inform structural decisions in politics and elsewhere. Additionally, such models can provide conceptual understanding for real world phenomena, such as the functioning or failure of expert groups, political committees or other decision making bodies.

From a methodological perspective, the model illustrates how agent-based simulations can contribute structure and substance to arguments that are hard to underpin empirically. Certainly, no purely theoretical model can appropriately substitute empirical data, and the model's high degree of abstraction implies that applications to real world cases cannot be made easily. Yet, when there is no well-founded basis on which to make a claim (in the present case: the epistemic quality of group decisions), it is better to make an argument on the basis of a clearly outlined, hypothetical scenario that captures a broad range of possibilities, rather than to make a claim on no basis at all. At the same time, there is an independent quality in identifying and understanding the causal structures underlying a certain process. In an abstract model, central causal effects can be focussed on while ignoring 'empirical noise'. Analysing such a model can help to highlight certain causal mechanisms and how they influence outcomes in a sense of studying the fundamental building blocks in a complexly interacting system. In doing so, potential causal explanations for certain real world phenomena can be presented.

In a larger framework, this paper also feeds into a justification of democratic decision making for epistemic groups. By employing formal models from social epistemology, it is shown that communication and voting structures can produce epistemically superior outcomes. Cohen's initially quoted suggestion that "the decision of majorities about which policies to pursue can provide good evidence about which policies are in fact best", (Cohen 1986, 34) can therefore not only be confirmed, it can also be made concrete and substantiated by the model. The findings from the model go beyond the classic jury-theorem-results by also considering the impact of communication style (especially openness for other opinions) and different decision rules. Additionally, time is

introduced as a dimension of analysis. This allows for the evaluation of questions regarding the efficiency of decision making schemes. Thus, one can not only say how groups can get it right, but also what an efficient decisions procedure for a certain group should look like, and how certain pitfalls can be overcome.

References

Arrow, K. J. 2012. Social choice and individual values, vol. 12. Yale: Yale University Press.

Austen-Smith, D., and Feddersen, T. J. 2006. Deliberation, preference uncertainty, and voting rules. *American political science review* 100 (2): 209-17.

Bächtiger, A., and Wyss, D. 2013. Empirische Deliberationsforschung – eine systematische uÜbersicht. *Zeitschrift für vergleichende Politikwissenschaft* 7 (2): 155-81.

Bozbay, I., Dietrich, F., and Peters, H. 2014. Judgment aggregation in search for the truth. *Games and Economic Behavior* 87: 571-90.

Cohen, J. 1986. An epistemic conception of democracy. *Ethics* 97 (1): 26-38.

de Condorcet, N. C. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris.

DeGroot, M. H. 1974. Reaching a consensus. *Journal of the American Statistical Association* 69 (345):118-21.

DeMarzo, P. M., Vayanos, D., and Zwiebel, J. 2003. Persuasion Bias, Social Influence, and Unidimensional Opinions. *The Quarterly Journal of Economics* 118: 909-68.

Doraszelski, U., Gerardi, D., and Squintani, F. 2003. Communication and voting with double-sided information. *Contributions in Theoretical Economics* 3(1).

Douven, I., and Riegler, A. 2009. Extending the Hegselmann-Krause model I. *Logic Journal of IGPL* 18 (2): 323-35.

Duggan, J., and Martinelli, C. 2001. A Bayesian model of voting in juries. *Games and Economic Behavior* 37 (2): 259-94.

Elster, J. 1998. *Deliberative democracy*, vol. 1. Cambridge: Cambridge University Press.

Estlund, D. M. 2009. *Democratic authority: A philosophical framework*. Princeton: Princeton University Press.

French Jr, J. R. 1956. A formal theory of social power. *Psychological review* 63 (3): 181.

Gerardi, D., and Yariv, L. 2007. Deliberative voting. *Journal of Economic Theory* 134 (1): 317-38.

Gigone, D., and Hastie, R. 1993. The common knowledge effect: Information sharing and group judgment. *Journal of Personality and social Psychology* 65 (5): 959-74.

Goeree, J. K., and Yariv, L. 2011. An experimental study of collective deliberation. *Econometrica* 79 (3): 893-921.

Golub, B., and Jackson, M. O. 2010. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics* 2 (1): 112-49.
Guarnaschelli, S., McKelvey, R. D., and Palfrey, T. R. 2000. An experimental study of jury decision rules. *American Political Science Review* 94 (2): 407-23.
Habermas, J. 1996. *Die Einbeziehung des Anderen. Studien zur Politischen Theorie*. Frankfurt a. M.: Suhrkamp.
Hegselmann, R., and Krause, U. 2002. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation* 5 (3).
Hegselmann, R., and Krause, U. 2006. Truth and cognitive division of labor: First steps towards a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation* 9 (3):10.
Lazarsfeld, P. F., and Merton, R. K. 1954. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society* 18 (1): 18-66.
Lehrer, K., and Wagner, C. 1981. *Rational Consensus in Science and Society: A Philosophical and Mathematical Study*. Dordrecht: Reidel.
List, C., and Goodin, R. E. 2001. Epistemic democracy: generalizing the condorcet jury theorem. *Journal of Political Philosophy* 9 (3): 277-306.
McPherson, M., Smith-Lovin, L., and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27 (1): 415-44.
Muldoon, R. 2013. Diversity and the division of cognitive labor. *Philosophy Compass* 8 (2): 117-25.
Rawls, J. 2009. *A theory of justice*. Harvard: Harvard University Press.
Riker, W. H. 1982. *Liberalism Against Populism*. San Francisco: W.H. Freeman.
Rousseau, J.-J. 1964 [1762]. *Du contrat social ou principes du droit politique.* Œuvres complètes, 3.
Sunstein, C. R. 2002. The law of group polarization. *Journal of political philosophy* 10 (2): 175-95.
Sunstein, C. R. 2006. *Infotopia: How many minds produce knowledge*. Oxford: Oxford University Press.
Weisberg, M., and Muldoon, R. 2009. Epistemic landscapes and the division of cognitive labor. *Philosophy of science* 76 (2): 225-52.
Zollman, K. J. 2010. The epistemic benefit of transient diversity. *Erkenntnis* 72 (1): 17-35.
Zollman, K. J. 2013. Network epistemology: Communication in epistemic communities. *Philosophy Compass* 8 (1): 15-27.

Paper 4

Fear Appeals as a Political Strategy

–

A Theoretical Exploration

This is an unpublished manuscript.

The suggested citation is:

Scheller, Simon (2017). *Fear Appeals as a Political Strategy – A Theoretical Exploration*. Bamberg: University of Bamberg (mimeo).

Noname manuscript No.
(will be inserted by the editor)

Fear Appeals as a Political Strategy – A Theoretical Exploration

Simon Scheller

Received: date / Accepted: date

Abstract Fear appeals constitute a popular populist strategy, especially among extreme parties. This paper sets out to examine the psychological mechanisms and the strategic rationale behind this strategy, and how such attempts can be counteracted.

Findings from Affective Intelligence Theory suggest that people can be more or less influenced by others depending on their associated emotional state. By influencing people's emotions, political actors can thus influence the dynamic process of public opinion formation in order to increase their electoral support. These aspects are cast into an agent-based model.

It is found that fear appeals are an effective yet dangerous tool. They increase a party's reach for new potential supporters, yet also increase the risk of loosing former core supporters. Extreme parties therefore need to clearly differentiate themselves from others. Moderate parties can counter such attemps by moving towards the extreme themselves and targeting the voters in between.

Keywords Political Psychology · Bounded Confidence · Emotional Appeals · Opinion Dynamics · Populism · Party Competition

1 Introduction

Populist rhetoric experiences a recent upsurge in a variety of western liberal democracies. But not only has populism become more prominent, it has also

University of Bamberg
Feldkirchenstr. 21, 96052 Bamberg
Tel.: +49 (0)951 863-2818
Fax: +123-45-678910
E-mail: simon.scheller@uni-bamberg.de

become more successful – a development of which Donald Trump and the Brexit-campaign constitute the most prominent examples. These developments emphasize the necessity to understand populist strategies, how they work and what can be done to counterbalance them.

While the concept of populism circumscribes a variety of facets, emotional appeals – specifically frequent and consistent addresses to people's fears – represent one central element of populist strategies, especially among extreme[1] parties (Wodak, 2015; Heinisch, 2003; Chevigny, 2003; Cincu, 2017; Pfau, 2007). This paper sets out to explain the impact of emotional appeals, how they can be employed strategically in politics, and what implications they carry for public opinion formation. While the developed framework can be applied more generally, this paper focusses mainly on the use of fear appeals by extreme parties.

Emotional appeals influence how voters perceive and process information. Findings from *affective intelligence theory* (AIT) suggest that fear cues in political messages decrease people's reliance on partisan habits and increase their openness to new information. Appeals to enthusiasm, in contrast strengthen people's reliance on previous affiliations, thereby diminishing reliance on new sources of information (Marcus et al, 2000; Brader, 2006). While these findings are well established on the individual level, little is known about how they play out in a social context where political opinions are formed in exchange with one's peers. Theorizing this transfer from micro to macro level constitutes a core contribution of this paper. I thereby take a strategic perspective by asking how especially politicians with non-centrist views can employ these insights to maximize electoral support.

To answer this question, I construct an agent-based model of opinion dynamics where people are subject to the influence of emotional appeals by political parties. Borrowing from Hegselmann and Krause (2002), the model assumes a setting of *bounded confidence*: People are more or less open to alternative opinions, depending on their confidence bound ϵ. Capturing AIT's findings, appeals to fear widen the ϵ-interval, appeals to enthusiasm narrow ϵ. This model framework allows a systematic analysis of various parametrization in order to assess the strategic potential of emotional appeals.

The analysis finds that extreme parties can potentially increase support by appealing to fear, but only under risky conditions: Weaker party affiliation and more openness to alternative information creates a larger basin of attraction for more distant voters. Yet at the same time, former core voters can also be attracted by others. A party, therefore, needs to consider where it stands relative to the voters it wants to attract: Either, it clearly distances itself from potential competitors or takes a very similar position. Notably, this provides a micro-level causal explanation for a strategy of "product differentiation", which is frequently attributed to extreme parties in Europe (see e.g. Kitschelt and McGann 1997).

[1] The term 'extreme' is used solely in a spatial sense as the opposite of 'centrist' and without any value judgment implied.

The paper proceeds as follows: Section 2 lays the groundwork by describing empirical findings from AIT regarding information processing in political communication, with a special focus on the effects of emotions and their deliberate evocation. Section 3 presents the agent-based model to be used for the subsequent analysis. Section 4 thoroughly analyses the model and presents its most intriguing findings. Section 5 contextualises these results and evaluates their scope and relevance. I conclude by arguing that such models constitute a useful tool to connect individual level findings with macro level phenomena, thereby generating nuanced hypotheses to guide further empirical inquiries into the realm of opinion dynamics in politics.

2 Emotional Appeals in Political Communication

Determinants of political opinion formation

Following Downs (1957), political actors are assumed to aim at maximizing a party's vote share in an election.[2] How can political actors influence voters' opinions in order to achieve this goal? In order to address new segments of voters, one crucial hurdle to be overcome is people's strong predispositions and party affiliations (Lazarsfeld et al, 1948; Campbell, 1960). Frequently, individuals merely reinforce their already held opinion, as they expose themselves more strongly to the campaign of the candidate they perefer anyways. Partisanship, therefore, constitutes a centrally important decision heuristic for voters, as it provides easily understandable cues for complex decisions (Kam, 2005).

One central way to reach voters despite those obstacles is by political communication, which McNair (2011, p. 4) defines as any "purposeful communication about politics", thus highlighting its intentional aspect.[3] In their messages, politicians can appeal to the electorate *rationally* or *emotionally*. A *rational* appeal conveys relevant information and argumentation in order to elicit an informed and reasoned decision (Rosselli et al, 1995, p. 165). It "presents facts in order to persuade viewers that the evidence (statistics, logical arguments, examples, etc.) favors a particular position" (Kaid and Johnston, 1991, p. 56). In contrast, *emotional* appeals circumscribe "any communication that is intended to elicit an emotional response from some or all who receive it" (Brader, 2006, p. 68f).[4] In practice, this dichotomy is by far mutually exclusive, nor can it always be clearly drawn since most political statements

[2] More detailed differentiations, such as Strom (1990) between vote-seeking, office-seeking and policy seeking parties will be neglected in the context of this paper for the sake of simplicity.

[3] In order to narrow analytical focus, only one-directional acts of communication from political actors towards the electorate are considered, which encompasses still a rather broad range of communication channels, e.g. speeches, TV and radio advertisements or media interviews. For a broad overview of modern campaign strategies see Strömbäck and Kiousis (2014).

[4] Referring to the work of Payne and Baukus (1988), one could further distinguish between content and stylistic techniques of appealing to emotions. Instead of discussing the

contain aspects of both. However, for the sake of this paper, suffice it to say that for most political communication, emotions constitute a significant and identifiable part of political communication.

While one might wish political debate to be dominated by rational appeals, various findings attest their relative ineffectiveness: On the one hand, people's political knowledge is severely limited (e.g. Galston 2001; Fishkin and Luskin 1999; Page and Shapiro 2010). On the other hand, even if knowledge was significantly improved, it is found to have little if any impact on attitudes and behaviour (Fishkin, 1991, 1997). It is thus little surprising that according to conventional wisdom among political campaigners, emotional appeals are by far more effective in swaying voters than rational arguments (Brader, 2006, p. 22ff).

Examining the strategic use of such emotional appeals in political communication requires a thorough understanding of the underlying cognitive and affective processes. For this purpose, the following sections explain (1) how emotions in general impact political awareness and perceptions, (2) that people's emotions can be deliberately evoked, and (3) what these findings imply for the strategic use of emotional appeals.

Emotions: Definition, Scope and Effects

According to Affective Intelligence Theory (AIT), human behaviour is controlled by both affective and cognitive processes. *Cognitive* processes are considered conscious and purposeful actions, e.g. "thoughts, beliefs, inferences, and application of rules" (Brader, 2006, p. 55ff), while *affects* are automated, sub- and pre-conscious reactions to outside stimuli. Affective responses engage in a detective and directive function: They guide one's senses in discovering the environment's relevant features (such as threats) and to allocate cognitive resources accordingly (Brader, 2006, p. 56) (Marcus et al, 2000, p. 28f).

Emotions are "reactions provoked by a particular person or event" (Houghton, 2014, p. 133) and are hence *object-specific* (unlike a person's 'mood'). They can also be purely *subconscious* (unlike actively experienced 'feelings', see e.g. Damasio 2000). Finally, emotions can also be triggered by *abstract* stimuli (such as verbal messages, metaphors or pictures) which are removed from the original source of an emotion (Brader, 2006, p. 119). Crucially, these three characteristics imply that specific emotional responses can be strongly associated with, say, political messages. Emotions can be evoked in and closely tied to political communication – even if such instances are rare and infrequent. As a consequence, emotions in political communication can be modeled as an isolated system, allowing to neglect a person's full emotional state or all of her communicative actions in a model of political opinion fomation.[5]

effectiveness or efficiency of different tools or styles in evoking emotional responses, this paper is concerned solely with the strategic dimension of when and which emotions to evoke, and deliberately ignores practical matters.

[5] This does not question the potential impact of general emotions or moods for making political decisions. Yet, such general non-cognitive aspects are rather treated as empirical

Emotional responses are triggered by two distinct affective systems. The *dispositional system* monitors the success of the currently performed task and produces a feeling of enthusiasm if everything is going well, and triggers frustration or depression if things are not going as planned (Marcus et al, 2000, p. 46ff). The *surveillance system*, on the other hand, monitors the environment for external threats. If an unusual stimulus is perceived, the system breaks the individual out of her routine by causing a feeling of anxiety, which enables the body to immediately focus all cognitive resources on the imminent threat (Marcus et al, 2000, p. 53ff).

As a result, Marcus et al (2000) find that people's political thinking and decision making is strongly influenced by emotional responses. While emotions generally lead people to engage more in the political process, the type of discrete emotion evoked makes a crucial difference as to how this increased engagement plays out: Enthusiasm strengthens people's habitual choices and thinking. Anxiety leads people to abandon political affiliations, engages them to seek out new information and to rationally reconsider their current opinion or position (Marcus et al, 2000, p. 65ff). These reactions can be traced back to the previous distinction between affective-systems: Positive stimulation of the dispositional system leads to a continuation of previous habitual behavior, since the experience of enthusiasm suggests no need for change. In contrast, activation of the surveillance system by appealing to fear suggests increased attention to the environment is required – which, in the political context, translates into increased attention to political information and a weaker reliance on partisan habits.

The claim that emotions influence political thinking and decision making is supported by the evidence from a series of empirical cases, including support and opposition to NAFTA, attitudes towards the first Gulf War and evaluations of candidates in American presidential elections. In all cases, voters who were anxious about the political situation were more likely to abandon their partisanship and instead support the candidate that matched their individual preferences on the issue more closely. Those enthusiastic about an issue were more likely to stick with the candidate in accordance with their ideological identification (Marcus et al, 2000, p. 96ff). As the authors themselves put it most eloquently: "Citizens who feel calm about presidential candidates are more likely to act on partisan habits, but those who feel anxious are more likely to attend to new information, defect from partisanship, and vote on the basis of issue and trait assessments." (Marcus et al, 2000, p. 111).

Deliberate Evocation of Emotional Responses

Further, emotions can be deliberately *evoked* in order to generate the previously described effects. Brader (2006) experimentally studies the effectiveness and impact of emotional appeals in political advertisments: Participants were

noise in the present study, as there is no reason to assume that they would exhibit any consistent and systematic pattern.

subjected to a campaign ad embedded in a stream of a regular television broadcast. The advertisement's identical verbal message was underfed with emotionally laden music and pictures in the treatment group, and with emotionally neutral music and imagery in the control group. Prior and after the exposure to the stimulus, subjects were asked about their political behaviour and preferences. Deriving its hypotheses from AIT, the study focussed on the deliberate evocation of fear and enthusiasm, and the effects thereof.

As predicted, fear cues diminished the impact of partisan habits: Initially opposed or indifferent individuals were more likely to evaluate a candidate more positively and to actually vote for the candidate when the ad was presented in a fear evoking way (Brader, 2006, p. 115). Subjects also expressed less certainty in their choices after having been exposed to fear-evoking ads (Brader, 2006, p. 119). For those subjects that were exposed to enthusiasm-evoking ads, the opposite effect was found: The likelihood of relying on previous affiliations increased, while the openness for outside alternatives diminished significantly. This was, again, found for both the subjective assessment of a candidate's appeal as well as the likelihood of voting for the candidate: People who were subjected to an ad that enthusiastically portrayed their already preferred candidate were convinced even more after the stimulus, and subjects also expressed significantly more certainty in their decision (Brader, 2006, p. 131).

Another dimension that illustrates the manipulability of emotions in accordance with AIT: Fear cues doubled the recall of subsequent news stories (in the TV news stream people were subjected to), while enthusiasm even distracted viewers. Participants who were subject to fear evoking ads also stated more often that they might contact the campaign for further information (Brader, 2006, p. 135ff). In a similar vein, Brader et al (2008) report on a series of experiments indicating that anxiety increases information searching behaviour qualitatively as well as quantitatively, while enthusiasm has a reducing effect.

There exists further evidence for the proposition that emotional responses in people can be deliberately evoked: Kühne et al (2011) find that emotionally laden news segments influence the emotions elicited by the viewers. This also translates into influencing people's political choices and opinion formation in the context of concrete policy decisions. A series of experiments by Gross (2008), in which emotionalized framing influenced people's attitudes towards minimum sentencing regulations, provides further corroborative evidence.

In summary, emotions can be deliberately evoked by specifically tailored messages in order to strengthen or diminish partisan habits. This enables political actors to reach out to an "otherwise inattentive audience" (Brader, 2006, p. 126).

Strategic use of Emotional appeals

A limited number of authors have discussed the use of emotional appeals from a strategic perspective. Schnur (2007) argues that one major determinant of

choice of emotional strategy is whether a candidate wants to sustain her position in the race, or attack the status quo. A candidate lagging behind in a race, for example, has an interest in appealing to a broader audience in order to gain majority support. Such reasoning intuitively suggests fear-evoking messages as the appropriate choice, as there is not much to lose for someone lagging behind in a winner-take-all election. A leading candidate, in contrast, has an incentive to reinforce and strengthen current levels of support by appealing to enthusiasm. Ridout and Searles (2011) formulate this more generally, stating that "[l]eading candidates should be more likely than trailing candidates to use emotions associated with the disposition system (anger, enthusiasm and pride) to maintain existing public support [, while t]railing candidates should be more likely than leading candidates to use emotions associated with the surveillance system (fear) to encourage political learning and upset existing public support." (Ridout and Searles, 2011, p. 444). Brader (2006) finds supporting evidence for these claims: "Consistent with expectations, political challengers are more likely to produce fear appeals than incumbents, and front-runners are more likely to produce enthusiasm appeals. Moreover, in state elections, candidates are more likely to appeal to enthusiasm when their party has the advantage among voters statewide, and are more likely to appeal to fear when the opposing party has the advantage." (Brader, 2006, p. 15f). When looking at election races over time, fear appeals by a trailing candidate are more likely to occur towards the end of a race in order to break partisan habits so as to encourage last-minute political learing (Ridout and Searles, 2011, p. 445).

Beyond the works just mentioned, I am (to the best of my knowledge) not aware of any further theoretical or empirical inquiries into the strategic use of emotional appeals. While the suggestions from above comply with common sense, a variety of aspects potentially complicates seemingly straight-forward strategic imperatives. Consider for instance the situation in a multiparty political system where vote shares, and not just a binary candidate choice are at stake. A candidate or party with only low or moderate support might decide – in accordance with the above argument – to follow a strategy of launching fear appeals to potentially widen its scope of influence. Yet, while there is a chance of drawing more supporters, there is also the risk of diminishing them even further – after all, fear appeals should make individuals also more open for arguments from the other side. This paper provides a theoretical framework which allows scrutinizing the strategic dimension of employing emotional appeals in such proportional electoral systems with a specific focus on fear appeals by non-centrist parties to enlarge the group of potential supporters. For that purpose, the subsequent section formally conceptualizes the process of opinion formation under the influence of emotional appeals – based on the described findings on the effects of emotions on voter's political behaviour.

3 The Model

The Bounded Confidence Framework

As a modeling baseline, I recur to the *Bounded Confidence* (BC) model by Hegselmann and Krause (2002). In the BC-model, each of $n \in N$ *agents* holds an opinion which is represented by a real number from the interval $[0, 1]$. The opinion of agent i at time t is given by $x_i(t)$, and let $x(t)$ be the profile of all opinions at time t. Further, each agent is characterized by her *confidence bound* ϵ_i. An agent's ϵ_i determines which other agents she considers for opinion updating: An agent considers all other agents with a distance smaller or equal to ϵ_i from her own opinion; all those agents farther away than ϵ_i are disregarded. Formally, agent i's *influencing set* I_i is given by: $I_i(\epsilon_i, x(t)) = \{j| \ |x_i(t) - x_j(t)| \leq \epsilon_i\}$. Logically, I_i always includes agent i herself.

Agents update their opinions simultaneously in discrete time steps by averaging the opinions of all agents from I_i with equal weights, which is expressed by the following formula:

$$x_i(t+1) = \frac{\sum_{j \in I_i} x_j(t)}{\#I_i},$$

$\#I_i$ is the number of agents in the influencing set I_i. This formal description of the BC model encompasses a variety of appealing characteristics in light of this paper's purpose:

- As is most apt for a political context, opinions are displayed on a continuous one-dimensional spectrum. This way of formalizing opinions is therefore sufficiently rich to include key factors, such as comparability of ideological distances between individuals, while still being maximally simplistic.
- Opinion formation is described as a social process among peers. The model can be easily extended to capture communication under the influence of political actors, as described below on the basis of Hegselmann and Krause (2015). While, certainly, a broad spectrum of influencing factors can be identified, peers and political actors constitute the central influences on an agent's opinion.
- As the discussion in section 2 illustrates, selective attention to and selective acceptance of certain opinions or individuals are key features of political communication. The variable ϵ embodies exactly this mechanism: When updating opinions, agents listen to only those others they have a close enough affiliation with. Again, the model describes the phenomenon in the simplest possible fashion. It captures the essence of the described selective attention effect while reducing the model's complexity to its barest minimum.
- As a result of the two previous points, the model provides a suitable structure to consider the role of emotional appeals in political communication. For this, only minor modifications and a slight reinterpretation of ϵ are required. These points are described in more detail subsequently.

– Finally, the BC model is possible to replicate and explain fragmentation, polarization and consensus formation in public opinion. These processes are crucial in explaining how party support comes about. As they are endogeneous to the model, it is able to discuss and explain strategic aspects in direct connection with these phenomena.

As argued in these points, employing the BC-model constitutes a suitable choice for the research question of this paper. Without doubt, certain features could have been designed or chosen differently. For example, Deffuant et al (2002) also use a BC-framework, yet with a random pairwise updating mechanism. Yet, as for other agent-based models, discussing such alternatives can only be fruitful to a certain extent. Beyond the previously given justifications for the model specifications, one should not underestimate the fact that Hegselmann and Krause's BC framework is among the most cited and most broadly accepted theoretical frameworks in the literature on opinion dynamcis and hence provides an ideal frame of reference for the present as well as further projects.

Affective Opinion Formation in the Formal Model

As section 2 outlines, an agent's emotional state plays a central role in determining how she deals with incoming signals. More specifically, individuals are more or less open to other people's views depending on their associated emotional state when receiving a message. The parameter ϵ directly incorporates this openness to outside sources: The larger an agent's confidence bound, the more open she is for listening to other agents.

In accordance with the findings from section 2, being in a state of enthusiasm reduces an agent's openness and strengthens reliance on partisan habits. In the model, this is reflected as agents having a comparatively smaller ϵ. On the other hand, if an agent experiences anxiety, she would be ascribed a comparatively smaller ϵ. This, in a nutshell, is how the empirical findings from section 2 and the formal model come together.

Emotions in their entirety are of course far too rich a concept to be ever reasonably captured by one single parameter such as ϵ. The model strongly simplifies the complexity of emotions, which owould be overstretched if one were to say that 'an anxious citizen exhibits a larger ϵ'. Emotional cues have a variety of other effects aside from influencing political perceptions. Yet, the purpose of this paper is not to model emotions and actors in its fullest detail, but to narrow the focus of attention on the impact of affective states on information processing. While further effects of emotions doubtlessly occur, they have only limited relevance for the analysis of opinion dynamics in this context.

At the same time, emotions are by far not the only potential factor of influence on ϵ. To mention only a few others, political sophistication, character traits or the situational context doubtlessly play their part in information processing as well. Again, I argue that this stark simplification is justified in

the abstract model because I do not claim to provide an all-encompassing explanation of information processing, but to address only the limited area of how emotions can be used to affect opinion processing. In summary, emotions are more than ϵ, and ϵ is more than emotions. Still, the connection between information processing and emotions is a central piece of the model because it constitutes a central factor in real world political communication.

Similarly, the model claims to sufficiently capture political communication, and not communication in general. Political communication amouts to only a minor share of a person's communication and usually occurs infrequent and seldom. However, the distinct effect of emotional cues in political messages is not lost by this acknowledgment. As AIT asserts, it suffices that emotional cues trigger the respective emotional responses whenever political information processing occurs, regardless of the fact that long periods with completely unrelated affective states may occur in between. Due to the object-oriented nature of emotional appeals, they trigger affective responses in the very moments a political message is presented. As a result, politically enduced emotions do not fade or become watered down by other instances of communication but become closely associated with the very context of politics. It is therefore justified to envisage a model that is solely focussed on those instances where an agent is subject to political statements. This is what the model in this paper represents.

Modelling Political Parties in the bounded confidence model

Having discussed the core of the relationship between model parameters and their empirical underpinnings, i.e. emotional states and their representation in the BC model, I now turn to the formalization of political actors and how they cast emotional appeals. Conveniently, Hegselmann and Krause (2015) already describe something very similar in the BC-framework when talking about radical groups and charismatic leaders. A radical group has an opinion on the one-dimensional opinion space just like regular agents, and they influence the opinions of regular citizens just like regular agents influence each other. However, the radical group itself does not change her position at all. The model with radical parties produces remarkable results. For example, a group that has a very extreme opinion does not necessarily attract more agents when its strength increases.

Political parties can be more influential than average citizens for various reasons. They may explicitly be considered direct sources of political information, they might be regarded as experts on political questions or, in the case of parties, they may speak for large and influential groups or institutions. To incorporate this, the parameter φ for *party-strength* is introduced. It expresses the strength of a party by the number of how many agents it represents. If a party has the strength $\varphi = 1$, it has the same impact as one normal citizen; if $\varphi = 20$, then voters consider the party as if the party were 20 individuals, all with the exact same opinion.

Trivially, party k's updating rule for her position Φ is: $\Phi_k(t+1) = \Phi_k(t) = \Phi_k$. The influencing-set for regular voters now also includes parties:

$$I_i(\epsilon_i, x(t)) = \{j| \, |x_i(t) - x_j(t)| \leq \epsilon_i\} \cup \{k| \, |x_i(t) - \Phi_k| \leq \epsilon_i\},$$

with $j \in N$ and $k \in K$. The updating-rule for agent-opinions changes to:

$$x_i(t+1) = \frac{\sum_{j \in I_i} x_j(t) + \sum_{k \in I_i} \Phi_k \times \varphi_k}{\#I_i + \sum_{k \in I_i} \varphi_k}$$

Note that party positions are deliberately assumed to be fixed in this setting, even though many other studies focus solely on the aspect of strategic party positioning, most famously Downs (1957), or more recently Laver and Sergenti (2011) and Hegselmann et al (2014). Compared to these seminal works, the perspective and focus in this paper differs in one key regard: While strategic positioning may be a feasible strategy for parties in the long run, fixed positions are a reasonable assumption for shorter time frames. Parties must provide potential supporters with a consistent world view so as not to be regarded as opportunistic or fickle. In short time frames such as the period of one election campaign, positional changes would therefore be practically unreasonable for political parties. For the model, they can therefore be reasonably considered as fixed. Especially in the context of emotional appeals, such a shorter time-frame are more relevant. Further, even for cases where could use both tools, the model's idealisations would still allow for a fruitful analysis.

Modeling Emotinal Appeals

As a final step, emotional appeals by political actors are introduced. Since emotional states are, as argued above, captured in the agent's confidence bound ϵ_i, an emotionl appeal must influence a voter's ϵ_i. On a technical level, influence on ϵ works the same way as regular influence on opinions: An agent averages over all ϵ-values of the agents and parties within her confidence bound. However, the interpretations strongly diverge between regular voters and political actors here:

When regular agents influence the ϵ_i of their peers and vice versa, this amounts to emotional cues being transmitted in voter-to-voter communication. Empirical justifications for this are manifold. One can see this as a matter of emotional contagion that happens via facial recognitions (Hatfield et al, 1993), via adopting and reiteration of political frames and language, or through explicit references to emotional cues towards certain issues.

In contrast, when a party influences a voter's ϵ_i, this is interpreted as a deliberate emotional appeal by the party. For better distinguishability, label a party's ϵ_k-equivalent parameter Π_k. Crucially, parties influence voter's confidence bounds, but not vice versa – just as with opinions. Hence, Π_k is constant for each party and the updating rule is once again trivial. For regular voters,

the updating rule is analogous to opinion updating, only that instead of opinions, confidence bounds are averaged among all influential actors (i.e. peers as well as parties):

$$\epsilon_i(t+1) = \frac{\sum_{j \in I_i} \epsilon_j(t) + \sum_{k \in I_i} \Pi_k \times \varphi_k}{\#I_i + \sum_{k \in I_i} \varphi_k}$$

The parameter Π_k is where a party's emotional appeals are to be located. In the model logic, Π_k is assumed to be strategically chosen by party k in order to maximize voter support. For example, if a party chooses a large value for Π_k, this is equivalent to saying that the party employs appeals to fear. All agents listening to that party would be subject to the fear appeal and, as a result, their ϵ_i increases. The influenced agents experience anxiety and, according to AIT, become more open to stranger's opinions and less driven by partisan habits.

Consequently, a party's emotional appeals can carry over to more remote individuals via interjacent agents. In practice, this may happen when an extreme party shapes the language of a discourse in a certain direction, and while not everyone may be prone to listening to the party itself, certain features of their position and rhetoric may carry over to ideologically more distant individuals.

As with emotions, a common critique might address the model's overly simplistic nature when depicting political parties and their influence. Without doubt, an empirically appropriate description of parties would need to include many more functions, internal processes, aims and many other aspects. Yet, again, the justification of this simplification is one of modelling purpose and focus: The goal is not to create a detailed depiction of political parties, but to capture one highly relevant aspect for the context of how political parties influence public opinion by means of emotional appeals, and to dismiss all aspects which are not centrally relevant for opinion formation and information processing.

From a technical point of view, there are no restrictions as to the available strategies for different actors. Such limitations, however, clearly exist in practice. For example, it is not easy to say whether or not centrist and non-centrist parties can employ appeals to fear equally easily. In order to avoid complication from that side, I deliberately limit and focus the analysis on the potential use of fear appeals by extreme parties. The reasons for this specific focus, as has already been explained, lie in the empirical occurrence of such appeals and the pre-existing academic interest in the topic. A broader application of the framework is possible and most certainly constitutes a pathway for further fruitful research, but would require contextual analysis up to a level that is unsuited for the nature and scope of this paper.

With the formal model at hand, I now turn to analysing the impact of emotional appeals in the described framework.

4 Results

Parameters and Output Variables

The overall goal of this enquiry is to analyze the strategic possibilities of emotional appeals, focussing on the role of fear appeals for extreme parties. The explicit goal of a party is defined as maximizing electoral support. In the model context, electoral support is defined as the number of regular voters that end up approximately[6] at the party's position at the end of a model run. For further reference, this group of agents is referred to as *followers*.

While the paper's core interest lies in Π's influence on these output measures, there is a variety of other model parameters that need to be considered. These parameters constitute the conditions for a party's Π-strategy to be successful or not. In analyogy to previous works studying the BC model (e.g. Hegselmann and Krause 2015), some simplifications ensure manageability of analyzing the paramerter space.

First, this study starts out with scrutinizing the strategies for *one party only*, and interaction effects between two or more parties are not considered. While this doubtlessly constitutes a limitation, its insights constitute a valid and reasonable starting point for analysing situations with multiple parties.

Second, all voters start out with *homogeneous confidence bounds*. This is a standard assumption in the BC-framework.[7] Since heterogeneity is introduced through the party's emotional appeals, the initial homogeneity assumption is an important control and manageability requirement.

Third, the initial opinion profile $x(0)$ is chosen according to the *expected value distribution*: 100 agents are placed on the [0,1] interval with equal distance to each other, so that the expected value of the distribution is represented directly. Previous studies (Hegselmann and Krause, 2002, 2015), also making use of the expected-value distribution, have shown that minor changes in ϵ can have a strong, non-monotonic impact on outcomes. In contrast to random start-distributions, fixing $x(0)$ simplifies comparability between different parametrisations, as tipping-point effects could otherwise be hidden through averaging over large numbers randomly initialized runs.

In summary, this leaves the following parameters for the model analysis:

- $\epsilon(0)$: The confidence bound all agents start out with. For the major part of the analysis below, $\epsilon(0)$ is set to 0.1. This can be equated with a moderate to relatively low openness for other opinions, and hence a state of relative emotional neutrality, which can be increased through party's fear

[6] In most cases, voters under the opinion-influence of a party approach the party's position in the limit and, hence, theoretically after an infinite amount of time. As this is obviously mathematically untractable, model runs end as soon as agent movements are all smaller than 0.001 per round, and consequently, every voter who is within 0.1 of the party's position is considered a follower of the party.

[7] Hegselmann and Krause (2002, 2006, 2015) and many others work under that assumption. Others, like Lorenz (2010) study the impact of heterogeneous bounds of confidence explicitly.

propaganda over the course of the dynamic process. Without parties, this parametrisation results in a moderate opinion dispersion into four clusters, and thus constitutes an intermediate case in terms of polarization or fragmentation. A robustness check in the appendix varies ϵ and identifies resulting changes.
- φ: The strength of the party. Below, φ is first set to 10, which can again be seen as an intermediate case. The appendix provides rubustness checks.
- Φ: The position of the party. As this parameter is expected to directly interact with the strategic influence of emotional appeals, it is considered immediately within a range of 0 to 0.5 in steps of 0.01. Hence, all possible positions on the opinion space are considered.[8]

Results Overview

Figure 1 reports the number of followers for various values of Π and Φ, and for fixed $\epsilon(0) = 0.1$ and $\varphi = 10$. Call this plot the *follower landscape* for further reference.

The *horizontal* dimension represents the intensity and kind of the party's emotional appeal Π: For the given value of $\epsilon(0) = 0.1$, the vertical line at $\Pi = 0.1$ constitutes those cases where the party does not alter people's emotional states. Moving further to the right intensifies the party's fear appeals. Moving to the left displays cases where the party appeals to enthusiasm.

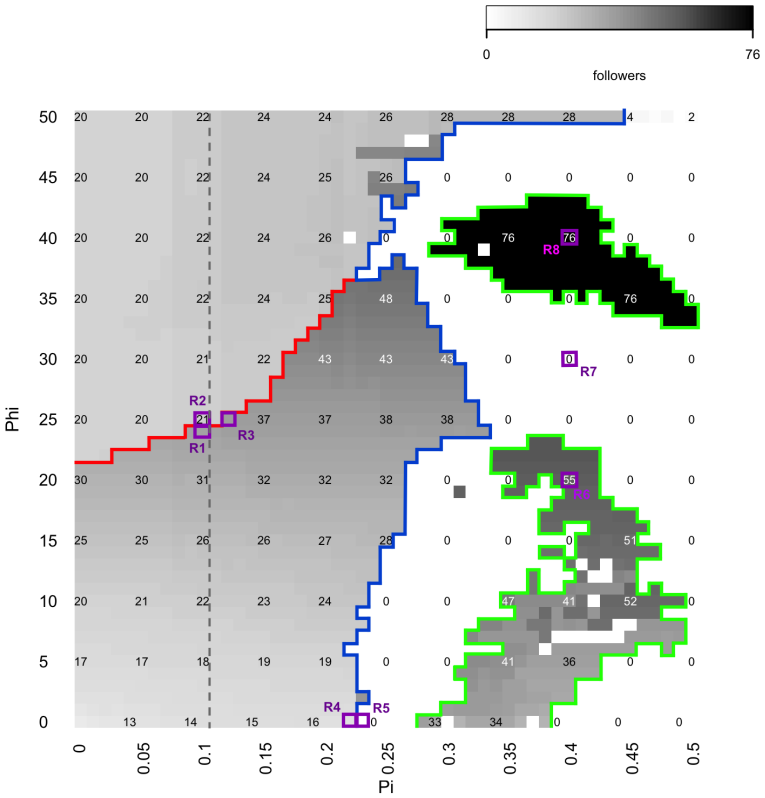
The *vertical* dimension represents the party's position Φ: At the bottom of figure 1 lie the cases where the party takes up the most extreme positions on the opinion spectrum. The further up one goes, the more moderate the party's position becomes.

Figure 1 displays a complex pattern with sudden jumps from high follower numbers to low follower numbers, with both an impact of Π and Φ on followership. Hegselmann and Krause (2015) detect patterns of a similar nature when analyzing the influence of radical groups, which means that the occurrence of such patterns itself is not unusual. The key issue for analyzing and interpreting the model is to understand how the specific distribution of high and low follower numbers comes about. Once the generating mechanisms for the patterns are understood, conclusions about the conditions for successful use of emotional appeals can be drawn. Overall, there are three phenomena that stand out:

- *The red line*: On the left hand side of figure 1, follower numbers below and above the red line differ significantly. The red line itself has an upward slope.
- *The blue line*: For small enough Π, an increase in Π mildly increases follower numbers. At a varying threshold level of Π however, follower numbers experience a sudden drop to zero.

[8] Choosing an even more fine grained step size would be possible but not necessary to identify crucial effects in the dynamics. All cases with values between 0.5 and 1 are symmetric to $[0.5; 1]$ and therefore require no special consideration.

Fig. 1: Number of Followers for $\epsilon(0) = 0.1$ and $\varphi = 10$



– *The green islands*: On the right hand side of the blue line, there occur two more or less clearly outlined clusters of very high follower numbers, surrounded by parameter areas where the party attracts no followers at all.

For the analysis of these phenomena, it is neccessary to look at the opinion dynamics of the *individual model runs*, as these depict the process of how the number of followers in a given case is produced. Looking at individual model runs is equivalent to 'zooming in' into a single point in figure 1. Subsequently, exemplary runplots at decisive points are depicted to illustrate the mechanisms that generate the overall patterns. These individual runs are referenced as R_i in figure 1.

The red line: A more extremist cluster

The first phenomenon to be analysed is the sudden and significant decrease of follower numbers when crossing the red line from bottom to top on the left

side of figure 1. Equivalently, one could say that extreme parties at one point lose followers when becoming more moderate. However, the red line's sloping trajectory implies that this loss of followers can sometimes be compensated by employing stronger fear appeals – yet only up to the point where the blue line is crossed.

To understand what happens when the red line is crossed, consider as an example the two adjoining boarderline cases R_1 and R_2 in figure 3. There is a clear difference between R_1 and R_2: With a party position of $\Phi = 0.24$, all voters with a more extreme position than the party end up supporting the party. When the party is positioned one step further to the middle at $\Phi = 0.25$, a small cluster of voters remains outside the party's sphere of influence at the bottom end of the opinion space.

Further, it becomes apparent why a slight increase in Π can prevent the formation of that bottom cluster, and thereby prevent the loss of followers: Through the stronger fear policy Π, the party enlargens the confidence bounds of its closest supporters. These, in turn, pass on the increased ϵ_i to their immediate neighbors. When this effect is large enough and reaches enough of the extremist voters, their ϵ_i becomes large enough so that they finally end up listening to the party. As there is no force of influence on the the extreme end, they are drawn towards the party. This is illustrated in figure 3 as well, comparing R_1 and R_3.

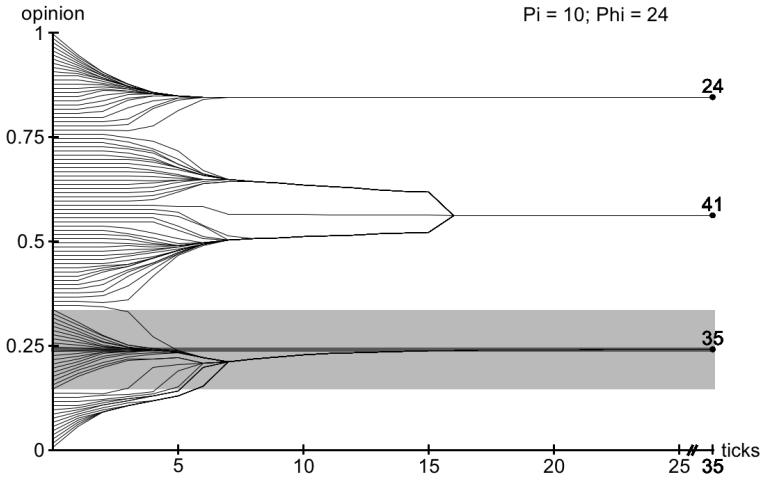
This finding provides a direct interpretation: When extremist parties aim at increasing their support by taking more moderate positions, they benefit from attracting more moderate voters. Yet when moving too far, the party runs the risk of leaving voters with very extreme opinions behind. To compensate for this, emotional appeals to fear can be a partial solution. In doing so, they (indirectly) increase the extreme voter's openness so that they take into account the party's opinion even if it is further away. However, why this cannot constitute a strategy to be used too excessively will become apparent through the further analysis.

The blue line: Appeasement by the Moderates

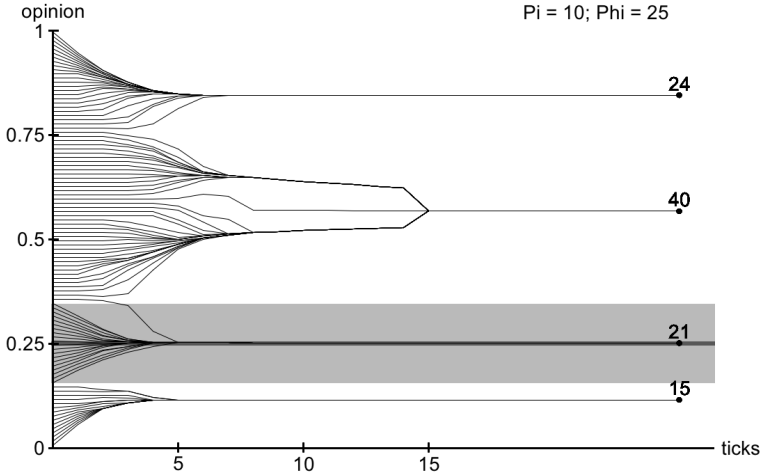
For the area on the left side of figure 1, moderately strong appeals to fear can partly prevent the formation of a more extreme cluster. As long as the blue line is not crossed, they also increase followership in general, as evidenced by the slight increase of followers from left to right in this area. Thus, moderate appeals to fear are beneficial for a party – no matter if it is extremist or moderate in its position. However, there is a clear maximum to the described strategy rationale: When increasing the intensity of fear appeals, there exists a certain threshold level for each Φ where followership suddenly drops to zero. These varying thresholds are marked by the blue line.

Fig. 3: Exemplary runplots R_1, R_2, and R_3 along the red line

(a) Runplot for R_1: $\Pi = 0.1$, $\Phi = 0.24$



(b) Runplot for R_2: $\Pi = 0.1$, $\Phi = 0.25$



(c) Runplot for R_1: $\Pi = 0.12$, $\Phi = 0.25$

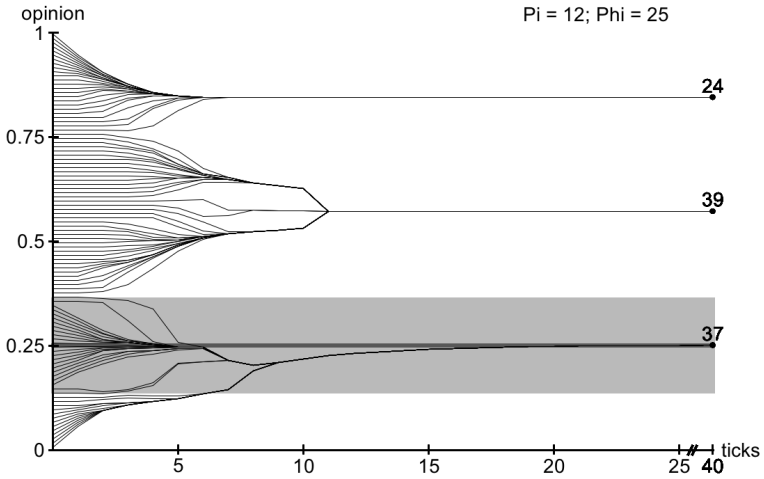
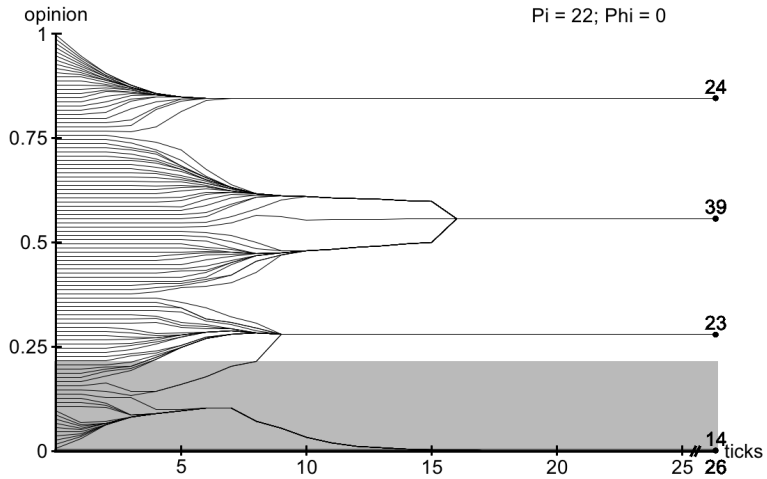
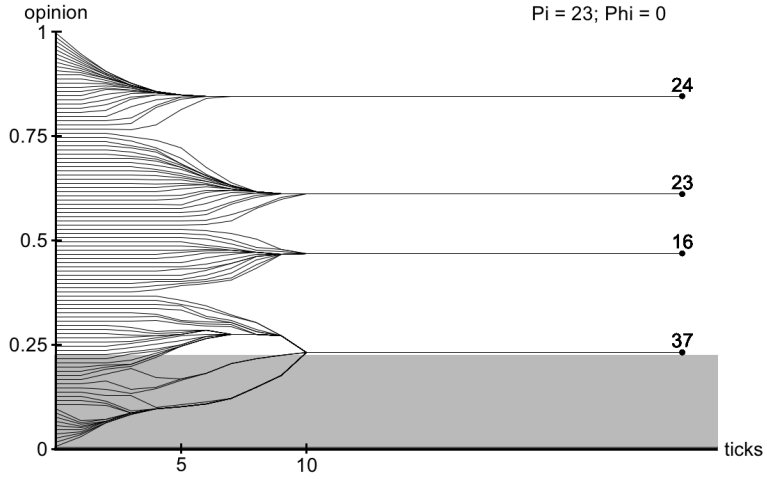


Fig. 2: Exemplary runplots R_4 and R_5 along the blue line

(a) Runplot for R_4: $\Pi = 0$, $\Phi = 0.22$



(b) Runplot for R_5: $\Pi = 0$, $\Phi = 0.23$



Runs R_4 and R_5 serve as explanatory examples for this phenomenon and are displayed in figure 2. While the party in the examples holds the most extreme position of $\Phi = 0$, the logic of the causal mechanism remains the same all along the blue line. The crucial difference between R_4 and R_5 consists in the fact that in one case, the voter cluster closest to the party splits away from the rest of the voter population and is pulled towards the party. In the other case, these voters remain connected to the rest of the population and are eventually pulled towards the adjoining moderate cluster of voters and therefore out of the party's reach.

What is the role of the party's Π-strategy in this process? Obviously, increasing the voter's ϵ_i does not only make those close to the party more prone to being influenced by the party. Since those voters become more open for influence by others in general, they are also influenced by a larger number of voters on the party's opposite side. Furthermore, the moderate voters are less fearful because they are not under the direct influence of the party's fear appeal; hence they exhibit lower ϵ_i-values. On the one hand, this explains why these moderate voters are not pulled towards the extreme end of the spectrum. On the other hand, this also explains why the extreme voters leave the party's influence eventually: Once they are pulled away far enough from the party, their ϵ_i decreases again through the moderating influence of the other agents. The moderates calm the fears that had been stirred by the extreme party.

To evoke fear for strategic purposes is thus a two-edged sword: On the one hand, more people are potentially attracted by the party when their openness is increased. On the other hand, these people are also attracted more strongly by opposing forces when openness becomes even larger. This can be seen as a rediscovery of a mechanism described by Lorenz (2010) with regards to heterogeneous bounds of confidence: If there are two groups of agents – one with a larger confidence bounds than the other – the final position of the combined cluster will be shifted in favour of those with the smaller ϵ_i, since they are not influnced as early as the group with the larger ϵ.

The green islands: The Buridan's Donkey effect

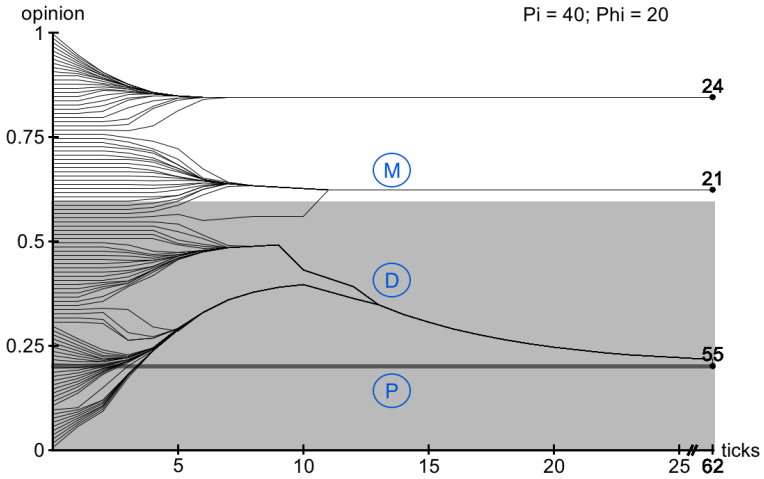
The analysis thus far has focused on the left side of the follower landscape depicted in figure 1. What is still left to be explained is the occurrence of the two green islands of strong followership in the midst of large spaces of no support at all.

Looking at the the exemplary runplots R_6, R_7 and R_8 presented in figure 4, the main cause for these abrupt changes appears to be a large cluster of voters (marked "D") located in between the party (marked "P") at the extreme end of the opinion space, and a large cluster of more moderate voters (marked "M") on its opposite side. In analogy to a parabola by the French philosopher Jean Buridan, call this in-between cluster *Buridan's Donkeys*. In Buridan's story, a Donkey faces a choice between two haystacks which are equally far away and equally appealing to her. Due to her indifference between the two haystacks, she is unable to decide which one to go to and, as a result, ends up starving. Being caught between two equally strong forces of attraction, the behavior of the voter cluster is very similar to the Donkey in the story. This pattern constitutes a crucial mechanisms in explaining the model's outcome under the envisaged parameter settings.

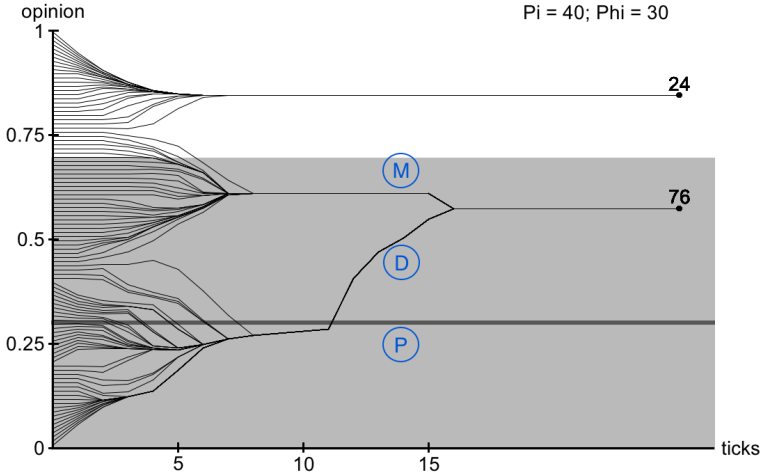
On the extreme end, the party (P) constitutes a fixed attractor by definition, as the party does not change its position. On the other side, the adjoining moderate cluster (M) constitutes a (temporarily) fixed attractor as well since their ϵ is too small for this cluster to be influenced by the donkeys in turn. As

Fig. 4: Exemplary runplots R_6, R_7, and R_8 for the green islands

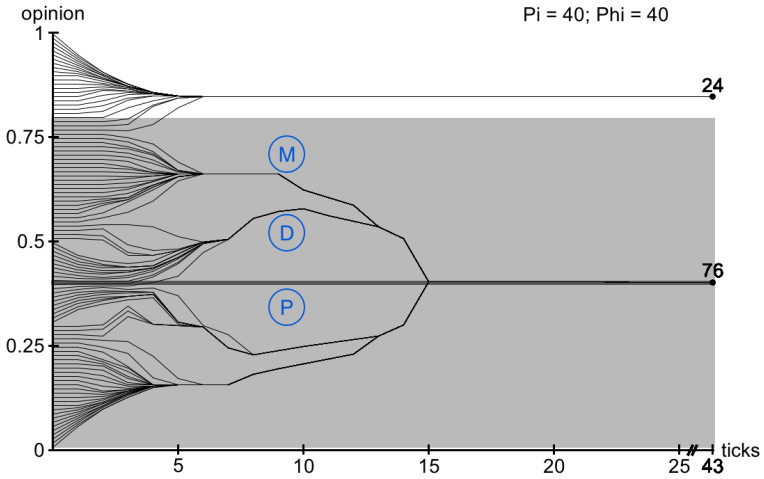
(a) Runplot for R_6: $\Pi = 0.40$, $\Phi = 0.20$



(b) Runplot for R_7: $\Pi = 0.40$, $\Phi = 0.30$



(c) Runplot for R_8: $\Pi = 0.40$, $\Phi = 0.40$



long as Buridan's Donkeys (D) are caught between the described poles, the situation constitutes an instable equilibrium for two reasons:

- *First*, the donkey-cluster can leave the sphere of attraction of one of the poles. As soon as this happens, the influence becomes one-sided. If the party remains the sole influence on the Donkeys (for example R_6 in figure 4), they approach the party's position hyperbolically. If the moderate cluster becomes the sole influence over the Donkeys, the two clusters usually merge together after very few timesteps (as in R_7 in figure 4).[9]
- *Second*, the moderate cluster can end up being influenced by they Donkeys itself. This triggers a dynamic process in which the moderates approach the donkeys quickly, while the Donkeys themselves are held in balance by the party. As soon as they merge together into one big cluster, both the donkeys and the former moderates are pulled towards the party's extreme position rather quickly. This causes the large numbers of followers in the parameter regions encircled by the green lines: Not only Buridan's donkeys are pulled towards the party, also the closest adjoining cluster is sucked in. An example for this is R_8 in figure 4.

Small changes in distances and ϵ_i influence in which direction the dynamic goes. This leads to either very strong support for the party, or none at all. In most cases, the equilibrium eventually tips into one of the two directions and the donkey cluster either merges with the party or the moderates. In some rare cases, the equilibrium remains intact, and the donkey cluster remains under the influence of both poles.

From a *strategic perspective*, the Buridan's Donkey effect carries central importance for the strategic use of fear appeals. As illustrated by figure 1, the interplay between Π and Φ determines whether or not the party is able to attract a large group of followers or not.

The *intensity of the fear appeals* Π needs to be large enough so as to initialize bridging the gap between the party and moderate voter groups. Yet, when Π becomes too large, the party's impact will always be overruled since agents with extremely large confidence bounds will be under the impact of more other voters. This development can be presented more clearly for the series of individual runs found under **this link**[10], which displays all individual runs for fixed position ($\Phi = 0$) with increasing Π (hence 'walking east on the x-axis of the follower-landscape'). The argument is essentially equivalent to the explanation behind the blue threshold line: Spreading fear is a two-edged sword since fearful voters are also attracted by a larger group of other voters.

With regards to the *party's position* Φ, strong fear appeals work when Φ is in $[0; 0.23]$ or $[0.34; 0.44]$, and do not work for $\Phi \in [0.23; 0.33] \cup [0.44; 0.5]$.

[9] In cases where the two groups have very unequal confidence bounds, it is also possible that the donkey's ϵ_i is reduced so quickly that in the process of approaching the other cluster, they get closer to the moderate cluster yet their ϵ_i shrinks more quickly. Then, the convergence is terminated and the two remain separate clusters outside each other's confidence bound. Although this effect is interesting in itself, it has no effect on party followers and will therefore not be given further attention in the subsequent analysis.

[10] See attached file "Phi0.gif".

Hence, when moving the party from the very extreme towards the center, strong fear appeals alternate in their effectiveness.[11] This is displayed in the sequence of individual runs for fixed $\Pi = 0.4$ and increasing Φ under **this link**[12].

The crucial determinant for the success of strong fear appeals appears to be the distance between the party and the moderate cluster. This distance has a major influence on how the Buridan's Donkeys situation above is resolved:

- If the distance is *large*, it is more likely that a significant group of voters breaks away from the rest of the voters and ends up at the party's extreme position.
- If the distance is *intermediate*, the moderate cluster is close enough to influence the Donkey-cluster, yet far enough not to be drawn into the party's sphere of influence itself. This explains the area between the two green islands where no party support occurs.
- If the distance is *small*, the moderate cluster is first attracted by the Donkeys, and then attracted by the party itself, leading to an even larger support for the (now not so extreme) party.

Generalizing this finding, if a party wants to use strong fear appeals to attract supporters, its position needs to be either clearly different compared to other groups, or very similar. Formulated as a strategic imperative: Either distance yourself from potential competitors, or approach them directly. Notably, this conforms with arguments brought forward by for example Kitschelt and McGann (1997) and Wagner (2012), who argue that small parties have an incentive to differentiate themselves from mainstream parties by taking more extreme positions and findings as a matter of political product differentiation. The logic discovered by the model analysis thereby also provides further insight into the incentives of extreme parties beyond the classic median voter argument by Downs (1957) and its manyfold successors.

Summary of findings

In summary, the model analysis invites several strategic considerations and recommendations. *First*, parties can successfully employ moderate fear appeals to reach out to more distant voters. Especially extreme parties can benefit from employing moderate fear appeals as they diminish the risk of leaving behind an even more extreme cluster when taking more moderate positions. Yet, as increased anxiety also augments openness to opinions from other actors, supporters may be drawn away from rather than towards the party.

Second, especially when strong fear appeals are at play, a crucial role falls to voters who stand in between – under the influence of the party on one side and a cluster of voters on the opposite side. In those situations, minor changes

[11] Neither of the described areas show a homogeneous and clearly definable pattern, so all of these interpretations apply to roughly defined parameter regions only.

[12] See attached file "Pi40.gif".

decide whether these undecided voters end up at the moderate's position or approach the party – sometimes even taking more moderate voter groups with them. In those cases, the party should aim at splitting this group away from the rest of the population's influence at one point. For this, it needs to take a position that is clearly distinct from other groups. Alternatively, the party can also move very close to the moderate cluster so that they easily become influenced by the party themselves. Never should a party carry out either of those strategies half-heartedly, as this is more likely to lead to core voters being led away from the party without gaining influence over any other groups.

Third and more generally, emotional appeals to fear always as a strategic device constitute a two-edged sword. While it offers high benefits through attracting further distant groups of voters, it also comes with the risk of losing one's core followers to the incrased influence of moderate groups.[13]

5 Conclusion

This paper has set out to examine the role of emotional appeals in political party's strategies, focussing on appeals to fear as a central building block of populist rhetoric by extreme parties. On the basis of findings from Affective Intelligence Theory, I have argued that one central effect of emotions is that they influence how we deal with other people's views: Fear decreases reliance on partisan habits and increases the relevance of newly incoming input; Enthusiasm narrows our perceptual focus. These emotional responses can be deliberately evoked by political actors.

While these effects are well established on the level of individual persons, it is uncertain how they play out under the influence of aggregation dynamics. To theorize this transition, this paper has proposed an agent-based model of opinion dynamics, in which a political party tries to influence voters by casting emotional appeals in order to increase its followership. With a specific focus on appeals to fear, the model analysis shows that fear appeals can be an appealing tool to reach broader voter audiences. However, parties also run the risk of losing core supporters through the potentially increased influence of competing groups. For fear appeals to be successful, the party's position plays a decisive role: A party should aim at increasing its reach through strong fear appeals only if it is able to clearly distance itself from ideological competitors, or when it can closely approximate the core audience's position. This provides a micro-level narrative for why clear political differentiation can be a successful strategy (e.g. Kitschelt and McGann 1997; Wagner 2012).

Doubtlessly, this project deals with complex subject matters in a simplifying way. As argued, however, this manifests one of the approach's strengths rather than a weakness. None of the model's simplifications claim descriptive accuracy (e.g. of human nature or of social processes). Instead, it offers a way

[13] All results presented in this section are qualitatively robust against variation in the parameters $\epsilon(0)$ and φ and only become overlain by other effects when the parameters take exteme values.

to put certain key variables at the center of attention, and to isolate characteristic causal mechanisms. Having isolated the effects of emotions as described by AIT, the simplification and agent-based representation allows transferring them to the aggregate societal level. This is something especially agent-based models are able to provide most effectively, and describing this translation constitutes the major methodological contribution of this paper.

Of course, the present study is also subject to a variety of limitations. For example, only single-party scenarios are analised, and different distributions of voters on the opinion space are disregarded. For reasons of conciseness, such endeavours lie outside the scope of this paper, yet provide potentially fruitful avenues for further research. Further, the obtained results are purely theoretical in nature even though they are substantially based on empirical observations. More empirical research is needed to complement the theoretical considerations of this paper.

Nonetheless, this paper contributes to the existing literature in two crucial ways. First, from a practical point of view, the analysis provides an inituitive guideline for understanding a variety of real world phenomena. Many right wing parties in Europe, for example, utilize instances of crime and terror to spread fear among the population. This boosts attention to their positions on their core ideological issues which would, in other times, simply be ignored. While many centrist parties have tried to clearly distance themselves from more extremist parties, it seems that at least to some extent, they aim to reduce ideological distances. One might consider this ideological weakness, yet according to the model's logic, this may also be a strategic move that builds bridges for otherwise undecided middle-ground voters.

Second, considering possibilities of further research, this paper provides theoretical guidance in an area of empirical uncertainty. There is a variety of hypotheses that result from the arguments in this paper: For example, it is suggested that when extremist political movements are rising to prominence on a wave of exploiting people's fears, they may experience a crucial point in time where they either manage to establish an isolated and fully convinced base of followers (such as the german PEGIDA-movement), or where their prominence fades in light of a moderation of discourse in favour of more moderate forces (which may or may not happen with the AfD-movement). To understand and further analyse such situations, the model's theoretical insights provide a sound basis.

References

Brader T (2006) Campaigning for hearts and minds: How emotional appeals in political ads work. University of Chicago Press

Brader T, Valentino NA, Suhay E (2008) What triggers public opposition to immigration? anxiety, group cues, and immigration threat. American Journal of Political Science 52(4):959–978

Campbell A (1960) Surge and decline: A study of electoral change. Public Opinion Quarterly 24(3):397–418
Chevigny P (2003) The populism of fear: Politics of crime in the americas. Punishment & Society 5(1):77–96
Cincu AE (2017) Far right populist challenge in europe. alternative for germany and the national front. Europolity 11(1)
Damasio AR (2000) A second chance for emotion. Cognitive neuroscience of emotion pp 12–23
Deffuant G, Amblard F, Weisbuch G, Faure T (2002) How can extremism prevail? a study based on the relative agreement interaction model. Journal of artificial societies and social simulation 5(4)
Downs A (1957) An economic theory of political action in a democracy. Journal of Political Economy 65(2):135–150
Fishkin JS (1991) Democracy and deliberation: New directions for democratic reform. Yale University Press
Fishkin JS (1997) The voice of the people: Public opinion and democracy. Yale University Press
Fishkin JS, Luskin RC (1999) Bringing deliberation to the democratic dialogue. In: The poll with a human face: The National Issues Convention experiment in political communication, pp 3–38
Galston WA (2001) Political knowledge, political engagement, and civic education. Annual review of political science 4(1):217–234
Gross K (2008) Framing persuasive appeals: Episodic and thematic framing, emotional response, and policy opinion. Political Psychology 29(2):169–192
Hatfield E, Cacioppo JT, Rapson RL (1993) Emotional contagion. Current directions in psychological science 2(3):96–100
Hegselmann R, Krause U (2002) Opinion dynamics and bounded confidence models, analysis, and simulation. Journal of artificial societies and social simulation 5(3)
Hegselmann R, Krause U (2006) Truth and cognitive division of labor: First steps towards a computer aided social epistemology. Journal of Artificial Societies and Social Simulation 9(3):10
Hegselmann R, Krause U (2015) Opinion dynamics under the influence of radical groups, charismatic leaders, and other constant signals: A simple unifying model. NHM 10(3):477–509
Hegselmann R, König S, Kurz S, Niemann C, Rambau J (2014) Optimal opinion control: The campaign problem
Heinisch R (2003) Success in opposition–failure in government: explaining the performance of right-wing populist parties in public office. West European Politics 26(3):91–130
Houghton DP (2014) Political psychology: situations, individuals, and cases. Routledge
Kaid LL, Johnston A (1991) Negative versus positive television advertising in us presidential campaigns, 1960–1988. Journal of communication 41(3):53–064

Kam CD (2005) Who toes the party line? cues, values, and individual differences. Political Behavior 27(2):163–182

Kitschelt H, McGann AJ (1997) The radical right in Western Europe: A comparative analysis. University of Michigan Press

Kühne R, Schemer C, Matthes J, Wirth W (2011) Affective priming in political campaigns: How campaign-induced emotions prime political opinions. International Journal of Public Opinion Research 23(4):485–507

Laver M, Sergenti E (2011) Party competition: An agent-based model. Princeton University Press

Lazarsfeld PF, Berelson B, Gaudet H (1948) The peoples choice: how the voter makes up his mind in a presidential campaign. New York Columbia University Press 1948.

Lorenz J (2010) Heterogeneous bounds of confidence: meet, discuss and find consensus! Complexity 15(4):43–52

Marcus GE, Neuman WR, MacKuen M (2000) Affective intelligence and political judgment. University of Chicago Press

McNair B (2011) An introduction to political communication. Taylor & Francis

Page BI, Shapiro RY (2010) The rational public: Fifty years of trends in Americans' policy preferences. University of Chicago Press

Payne JG, Baukus RA (1988) Trend analysis of the 1984 gop senatorial spots. Political Communication 5(3):161–177

Pfau MW (2007) Who's afraid of fear appeals? contingency, courage, and deliberation in rhetorical theory and practice. Philosophy & rhetoric 40(2):216–237

Ridout TN, Searles K (2011) It's my campaign i'll cry if i want to: How and when campaigns use emotional appeals. Political Psychology 32(3):439–458

Rosselli F, Skelly JJ, Mackie DM (1995) Processing rational and emotional messages: The cognitive and affective mediation of persuasion. Journal of Experimental Social Psychology 31(2):163–190

Schnur D (2007) The affect effect in the very real world of political campaigns. The affect effect: Dynamics of emotion in political thinking and behavior pp 357–374

Strom K (1990) A behavioral theory of competitive political parties. American journal of political science pp 565–598

Strömbäck J, Kiousis S (2014) Strategic political communication in election campaigns. Political communication Berlin: Walter de Gruyter pp 109–128

Wagner M (2012) When do parties emphasise extreme positions? how strategic incentives for policy differentiation influence issue importance. European Journal of Political Research 51(1):64–88, DOI 10.1111/j.1475-6765.2011.01989.x, URL http://dx.doi.org/10.1111/j.1475-6765.2011.01989.x

Wodak R (2015) The politics of fear: What right-wing populist discourses mean. Sage

Appendix

Robustness analysis

In the main party, all results were obtained for $\varphi = 10$ and $\epsilon(0) = 0.1$. It has been argued why this parameter constellation embodies cases with central relevance in practice. Further, the analysis below shows that the results remain qualitatively stable for a larger parameter region, and when the parameters are brought to their extremes, predictable limit cases occur.

First, consider *variation in* φ while keeping $\epsilon(0) = 0.1$ constant. This change produces a different follower landscape is depicted under **this link**[14]. Interpreting these plots, one can draw the following conclusions:

- In the vicinity around $\varphi = 10$, the results remain qualitatively the same, although the regions where the described effects occur shift around. This shows that the effects are robust for a relevant parameter region.
- With significantly smaller φ the formerly green islands of large support disappear. Strong appeals to fear do not work when the party is too weak, which can be explained simply by the fact that the party is not influential enough to compete with clusters consisting of more voters. A weak party should thus not aim at capturing too many voters, as the danger for loosing even close followers becomes increasingly strong.
  Nonetheless, moderate fear appeals are still a feasible strategy to increase supporters even for weaker parties. Surprisingly, weaker parties attract nearly as much followers as parties with a stronger inherent force of attraction. One potential explanation is that the increased ϵ_i is amplified enough through the initially close voters so that the party's sphere of influence can be enlarged significantly through moderate fear appeals. In a larger context, this is in accordance with a finding from Hegselmann and Krause (2006) that even weak forces of attraction can lead large groups towards a certain point of attraction as long as the contact between the direct and the indirect followers is upheld.
- For more influential parties, emotional appeals become less important, and the party's position becomes the major determinant for followership. As in Hegselmann and Krause (2015), this clearly shows that Φ has a non-monotonic impact on the party's success. Such a party's strong force of attraction reaches those in the party's close vicinity anyways. Since these voters split away fairly quickly from the rest of the population, they fail to provide indirect bridges from the party to more distant voters.

In conclusion, even relatively weak partys can employ moderate fear appeals successfully, even though the risk of loosing supporters is specially present. Very strong parties have no way of using fear appeals to increase their support, but in most cases, there is no need for them to do so anyways.

Now consider *variation in the parameter* $\epsilon(0)$, while $\varphi = 10$ is fixed. Speaking in terms of the model interpretation, $\epsilon(0)$ describes the emotional state the

[14] see the attached file "varystrength.gif"

society starts out with. The case $\epsilon(0) = 0.1$ from above results – in the absence of any party intervention – in the formation of four separate opinion clusters, which can be seen as both an interesting and a realistic case. The series of figures under **this link**[15] depicts the follower-landscape for varying values of $\epsilon(0)$.

Interpreting these plots, one can draw the following conclusions:

- In the close vicinity around $\epsilon(0) = 0.1$, the identified effects are robust on a qualitative level. Variation in effect strength and parameter regions where effects occur is to be expected and does not fundamentally alter the nature of the results.
- When $\epsilon(0)$ is smaller, all effects slowly fade. This is not surprising since for smaller $\epsilon(0)$, less and less communication in general takes place, the opinion space becomes fragmented into a large number of small clusters and there is not really any dynamic interaction between opinions. Hence, these cases are not worth studying as these parameter regions constitute an unrealistic and thus unimportant extreme case.
- For larger $\epsilon(0)$, the system would result in fewer yet larger clusters in the absence of party influence. This observation, which is one of the first observations that has been made about the standard BC-model, drives the shifts and changes in the identified effects also here. For example, the larger $\epsilon(0)$ is, the more the party's position gains in importance relative to Π: The more centrist a party, the more likely it is to attract large groups of voters. As there are fewer clusters in general, there are also fewer sudden jumps in followership that could arise from one more cluster being influenced by the party or not. Thus, the strategic imperatives from above remain valid in principle also for cases of higher initial fear levels and thus stronger initial polarization, even though fear strategies become less utilizable.

In summary, while the impact of $\epsilon(0)$ on the follower landscape appears dramatic at first sight, the core mechanisms identified remain valid even if they shift around and vary in their impact. This is also true for the rubustness analysis in general. Thus, the strategic imperatives from the exemplary case in the main part can be transferred to a larger class of relevant cases. Only when parameter values are brought to the extremes, the results fade in light of other effects overriding those caused through emotional appeals.

[15] see the attached file "varyepsilon.gif"