Collective Irrationality
An Agent-Based Approach

Christoph J. Merdes

München 2018

Collective Irrationality
An Agent-Based Approach

Christoph J. Merdes

Dissertation
an der Fakultät für Philosophie, Wissenschaftstheorie und
Religionswissenschaft
der Ludwig–Maximilians–Universität
München

vorgelegt von
Christoph J. Merdes
aus Fürth

München, den 20.03.2018

Erstgutachter: Prof. Dr. Stephan Hartmann
Zweitgutachter: Prof. Kevin Zollman, PhD
Tag der mündlichen Prüfung: 18.07.2018

Contents

List of Figures

List of Tables

Zusammenfassung

Kollektive Irrationalität durchdringt alle Bereiche des sozialen Lebens. Von der Herdenbildung an Börsen zur politischen Polarisierung und sogar bis in die epistemische Gemeinschaft der Wissenschaft misslingt es Gruppen, die optimalen Mittel zu wählen, um ihre Ziele zu erreichen. Der Begriff der kollektiven Rationalität lässt sich präzise fassen durch eine Analyse der Argumente, die seine Zuschreibung rechtfertigen. Dazu ist es erforderlich, die Entscheidungsumgebung, die ablaufenden sozialen Prozesse und die relevanten normativen Standards zu erfassen, wozu sich insbesondere agentenbasierte Modellierung und Simulationen eignen.

Diese Methode wird auf zwei Fallstudien angewendet, jeweils eines aus dem Bereich der theoretischen und eines aus dem der praktischen Rationalität. Die erste Fallstudie beschreibt die Entstehung sogenannter *unpopular social norms*, also solcher sozialer Normen, die dem Interesse der überwiegenden Mehrheit der Gruppe zuwiderlaufen. Die Analyse eines entsprechenden Modells, in dem begrenzt rationale Akteure mit eingeschränkter Information versuchen, eine optimale Norm zu wählen, zeigt, dass der zugrundeliegende Prozess zwar ineffiziente Normen generieren kann, häufig jedoch effiziente Ergebnisse liefert. Das unmittelbare Urteil der kollektiven Irrationalität muss daher zurückgewiesen werden. In der zweiten Fallstudie wird der Einfluss strategischen Verhaltens in einer Gruppe von Agenten, die Informationen aggregieren, untersucht. Simulationen unterstützen die These, dass es keine universell optimale epistemische Strategie gibt: Agenten, die unter idealen Bedingungen erfolgreich sind, unterliegen im Falle strategischer Einflüsse konkurrierenden Agenten, die wiederum unter idealen Bedingungen suboptimale Ergebnisse erzielen. Eine evolutionäre Analyse des Modells belegt darüber hinaus, dass nichtepistmisch motiviertes Verhalten unter einem wengistens teilweise auf epistemischen Werten basierenden Belohnungssystem zurückgedrängt wird, ohne jedoch vollständig zu verschwinden.

Die Verwendung agentenbasierter Modellierung und Simulation lässt sich insbesondere durch die epistemischen Eigenschaften der Methode rechtfertigen. Anders als Computersimulationen im Allgemeinen sind Modelle epistemisch transparent, das heißt nicht opak. Sie können außerdem unter den richtigen Bedingungen auch Bestätigung liefern. Gepaart mit der großen Flexibilität in der Modellierung handelt es sich daher ABM um ein wertvolles Werkzeug für ingenieursmäßig betriebene Philosophie. Dieser Ansatz strebt konkrete Lösungen für spezifische philosophische Probleme an, ohne vorher die Rechtfertigung einer fundamentalen Theorie zu fordern. Philosophische Probleme sind real und allgegenwärtig und verlangen daher Lösungen; die obigen Untersuchungen schlagen solche Lösungen vor.

Chapter 1

Introduction

1.1 Arguments on Collective Rationality

In one of the most famous passages in the historiographic writings of Livy, a large portion of the Plebs leave the city of Rome and gather outside on Mons Sacer. The Roman aristocracy, needing them both as soldiers and peasants to maintain the state, send Agrippa Menenius to negotiate the return of the Plebs to the city. What was the conflict about in the first place? The plebeians perceived themselves as doing all the work, while the aristocracy was just enjoying their leisure, living off the fruits of the labor of others, though still wielding all the power in the city.[1]

Agrippa Menenius realized that he would neither be able to coerce or threaten the Plebs back, nor could he reasonably rely on any moral argument. Therefore, he decided to approach the Plebs by reference to the interest of the state as a whole and in turn their own interest as a class within this state. Put otherwise, he appealed to their rationality as part of a larger collective.

To this end, he narrated the well-known allegorical story of the belly and the members. In brief paraphrase, this is its content: Once, the members started rebelling, complaining that the stomach was just sitting in the middle of the body, being fed by them, but not contributing anything itself. Therefore, they decided to stop feeding it: the arms no longer put food into the mouth, which the mouth wouldn't have chewed if they did. Deprived of food, the stomach could no longer provide the necessary nutrition and the whole body got ill.

What is this story besides a glaring piece of aristocratic propaganda? It is an interesting case study in what is difficult about collective rationality. Let me pick out just three central points. First, the members do not actually understand how the system they are a part of works. As a consequence, they fail to acknowledge the contribution certain parts of the system make to its maintenance, resulting in an overall deterioration of system performance *against their own best interest*. Thus, lesson one for the student of collective rationality is to understand the processes and interrelations in their object of analysis, be it society as

[1] My interpretation is based on Giebel (2015), book 2 paragraph 32.

a whole, a school class or the legislative body of a democracy.

Second, there are multiple levels of collective rationality involved. On the highest level, the maintenance and development of the state as a whole is at stake and figures as an evaluative standard. But to gain full argumentative force towards the Plebs, the argument needs to establish a connection between their interest as a class and the state as a whole – a connection that is presumed by the analogy with a body, though not clearly mapped onto Roman society. At any rate, the Plebs's class interest itself is still a collective end, as opposed to an individual one. Thus, the second lesson is to clearly determine the relevant evaluative standard, for otherwise the argument will fail to convince its target audience.

Third, the whole story of the members and the belly seems to assume a very limited set of options for the Plebs. Why shouldn't they go somewhere else and create a whole new state, where the leadership is chosen by lot from their ranks, and there is no inherited aristocracy at all? The answer is, at least as far as the historiographical depiction goes, that none of the people involved considered such an option realizable. So the third lesson is that any argument for or against the rationality of a pattern of behavior has to be presented against the backdrop of certain restrictions on what can and cannot be done, sometimes in an apparently arbitrary fashion.

Finally, it is worth pointing out that collective rationality is conceptually disconnected from morality. There is nothing in the narrative suggesting the immorality of the behavior of the plebeians; at worst, it is foolish. On the other hand, the Roman state itself might have been an immoral agent without taking away one bit from the argument.

The purpose of this essay is to contribute to a more thorough understanding of what is necessary to argue for – or against – an attribution of collective irrationality. Explaining and understanding social processes, explicating and applying a variety of evaluative standards for the collective rationality of behavior; and taking into account the situational restrictions on implementable options. As a result, we should be able to not only give better arguments than the one Menenius famously gave, but also protect the audience of such arguments against falling for invalid lines of reasoning. Ideally, even suggestions for interventions on collectively irrational behavior can be supported by the suggested analyses.

This chapter proceeds in the following way: In the next step, I shall establish the real-world relevance of collective irrationality in more recent times. Then I proceed to establish the methodology of agent-based modeling and simulation as a useful tool for social philosophy and the study of collective rationality (Section 1.3), as it will be employed in the central chapters of this essay. Finally, Section 1.4 gives the larger agenda for the rest of my investigation.

1.2 The Real World of Collective Irrationality

One of the most important questions to any abstract intellectual enterprise certainly is whether it serves any practical purpose. What are the real-world problems the concepts, theories and models are supposed to apply to, and why should we care? The analysis of

collective (ir)rationality finds application in all areas of human social life, and a better understanding of the phenomena, the underlying processes and the evaluative standards could greatly improve our ability to organize everything from markets over democratic government to cooperative scientific inquiry and the social norms of everyday life.

The social philosopher therefore can pick and choose their examples, and I will sketch just a few: some for their significance, others for the exceptional clarity in exemplifying a phenomenon.

The political polarization of large portions of the citizenry of democratic states in Europe and North America might be the most significant case of a threat to collective rationality in recent history.[2] Political polarization, i.e. the concentration of opinion in (most often two) opposing camps, is often perceived as both irrational and a threat to democratic institutions. It is viewed as a symptom of irrationality since most citizens have access to largely the same sources of both empirical evidence and arguments built on it. So how can they disagree on matters of fact, such as anthropogenic climate change, unless irrationality plays a part?[3]

But not only the origins of polarization are worrisome: its potential consequences pose a threat to the functioning of democratic polities and effective collective action. The debate on climate change provides a frightening example: failure to resolve polarization towards the most rational shared assessment of the matter may irreversibly deteriorate living conditions on earth. More immediately, polarization makes it more and more difficult to arrive at political compromises, which seem to be a necessary precondition for democratic rule. Therefore, it is important to analyze the underlying processes and come up with solutions to these problems if possible.

A quite different field of application stems from research in social psychology and sociology. To the initial puzzlement of observers, certain groups uphold social norms which are against the interest of all, or almost all its subjects. Bicchieri and Fukui (1999) discuss a wide variety of instances of such norms, from footbinding over alcohol overconsumption to gang violence. The difficulties with such unpopular social norms start with correctly identifying and explaining them. How does a population end up with such a norm, if close to nobody wishes to adhere to it? To be more precise and to increase the explanatory difficulty, the agents generally also do not prefer others to obey the social norm.

With respect to their normative status, unpopular social norms provide a relatively simple case, at least prima facie: their collective irrationality derives from the aggregation of its universal individual irrationality. As in a simple coordination problem, the agents

[2]Opinion polarization has been a topic in political science for a long time (McCright and Dunlap, 2011; Fiorina and Abrams, 2008; Evans, 2003; Poole and Rosenthal, 1984), but the theory of collective opinion formation leading up to polarization is still developing (Hegselmann and Krause, 2015; O'Connor and Weatherall, 2018).

[3]There are numerous routes to explain the phenomenon, some which are not reliant on anything particularly irrational happening on the collective level; maybe everything social is adequate, but the individuals engaged in the process are not sufficiently rational. However, if either rational or irrational individuals could end up with poor outcomes due to a poor process, the resulting explanation is robust across various actual and counterfactual populations, and may also open up interventions that do not require changing the cognitive capacities of hundreds of millions of individual agents.

are just worse off by miscoordinating, and as a consequence, the result seems collectively irrational. In practice, the problem is more complex, since universal indifference or dispreference for the norm-guided behavior do not prevail across all members of the group of norm abiders; often a small subpopulation actually prefers the behavior suggested by the unpopular social norm. This insight may not necessarily threaten its descriptive status as an unpopular norm, since unless the relevant subgroup is able to coerce the group into norm-following, we are still facing an explanatory puzzle.

But to still claim that something normative is going wrong, it becomes necessary to introduce additional assumptions and stronger normative standards. This potentially creates a need for controversial argumentative moves such as the assumption of intercomparability of utilities. All told, unpopular social norms provide a prime example for the prevalence of collective irrationality and the difficulties of its analysis.

Turning to the field of economic behavior, so-called herding (Banerjee, 1992) threatens the efficiency of markets. Herding occurs when economic agents follow the lead of particularly visible agents, thereby causing an information cascade: once a few traders started to herd, it becomes more and more compelling to believe they are correct in their judgments, and therefore to follow them. However, as Bannerjee shows, such cascades can be caused by arbitrary initial decisions under the right – or more accurately, wrong – circumstances, even though the agents are assumed to be rational Bayesians. Once again, the devil is in the details of the information-gathering process and interactions of traders. A proper analysis of the overall rationality of herding cannot simply point out the fairly uncontroversial inefficiencies it results in, but also explain how they realistically could be avoided.

A final remark on these real-world examples: it should already emerge from the sketches provided here that the actual collective irrationality of behavioral patterns is itself controversial. Ambiguity can arise from competing evaluative standards, disagreement about the process at work or the implementability of its alternatives, and potentially a number of other factors. One of the most demanding – but also most important – requirements for a philosophical analyst of collectively irrational phenomena is therefore to be precise about the limitations of any judgment of collective (ir)rationality.

1.3 An Agent-Based Approach

There are at least three major reasons for a social philosopher to employ agent-based modeling and simulation: (1) Arguments in social epistemology, political philosophy and related disciplines often have to rely on plausible descriptions of social processes and their evolution in time. (2) Agent-based models (ABM) offer specific, precise reconstructions of philosophical problems, supporting an engineering approach to philosophy. (3) Formal models create a discursive bridge to the empirical sciences, since they are formulated in the shared language of mathematics. Let me argue these points in some more detail.

State-of-nature theory in political philosophy serves as a perfect example to the first point. From Hobbes (1651) to Nozick (1974), political philosophers relied on informal descriptions of the state of nature and offered speculative derivations of the evolving be-

havioral patterns of their agents.[4] A common weakness of the resulting arguments is their reliance on implicit assumptions about human behavior in counterfactual circumstances. This problem is compounded with further implicit assumptions on social processes and human interaction.

For example, Nozick handwavingly argues that his process will not yield rogue states that willingly transgress the basic rights he presumes to be followed in the state of nature (cf. Nozick, 1974, p. 17). But the argument is less than clear, largely due to the lack of a sufficiently precise description of the process he is imagining.[5]

Agent-based models solve the problem by requiring the modeler to specify precisely all the initial assumptions and transition rules of the modeled process. Given the model, the resulting outcome can be checked and replicated by any recipient of the argument. Note, that ambiguity can re-enter on the level of interpretation; but at that point, the argument already gained substantially in interpersonal transparency.

Naturally, this advantage comes at a certain cost, namely a loss in generality. The philosopher employing agent-based modeling is not suggesting a grand theory of human nature and statehood (or the human epistemic condition and epistemic communities in other cases), but restrains himself to a more specific problem. This leads immediately up to the second reason.

In the preface to their "Bayesian Epistemology", Bovens and Hartmann (cf. 2003, Preface) argue for what could be called the engineering approach to philosophy. In an argument resounding in Titelbaum (2017), they worry that the grand theories of philosophy tend to go past our most pressing present problems in epistemology. According to this argument, it is therefore worth to concentrate on problems as specific as possible, provide focused accounts and potential solutions to these concrete problems before aiming for the full picture of the heaven of ideas.

I suggest to supplant their claim with an argument from a grand philosophical theory, namely reflective equilibrium.[6] The engineering approach to philosophy consists in developing contained accounts of problems, revising them in the light of empirical evidence, analytical insight and general principles and iterating the whole procedure until an at least temporarily satisfying picture has been established. This methodology fits perfectly the requirements of reflective equilibrium, and it recovers the intuition of many proponents of the method that in non-ideal conditions, perfect convergence is not always possible.

Agent-based models in turn are a natural tool for the engineering approach to philosophy. They need to be concrete to be implementable; they are always revisable given the vast design space every ABM lives in; they are able to incorporate highly general principles, but they are often judged in part by their fit to our empirical knowledge about the process to be modeled. Thus, ABM are certainly a prime method of the engineering philosopher, although not the only one.

[4]Nozick briefly discusses a game-theoretic representation of his account, but his general argument relies on the informal account.

[5]Even though this is not an essay on political philosophy, the widespread knowledge of state-of-nature theory suggests it as an example. The argument translates to problems in social epistemology.

[6]cf. Rawls (1975) for the most famous statement of this method.

Finally, the capability of ABM to bridge the gap between philosophy and empirical science constitutes a third important reason. Over the millenia, the sciences emancipated themselves from the philosophical enterprise. For all the advances it enabled, this differentiation also deprived the resulting disciplines of certain intellectual resources. Specifically, odern empirical science is limited in its capacity to analyze normative questions. Therefore, it is crucial to acknowledge the shared interests of, in the case of this essay, social philosophers and sociologists, economists, political scientists and psychologists.

But not only do these disciplines differ in methodology and perspective, they often do not even share a common language. This is where agent-based models – among other formal models – can step in. A social epistemologist can read and understand a paper on opinion dynamics written by a sociologist, and vice versa, due to the shared methodological apparatus and language of ABM.[7]

Therefore, agent-based modeling and simulation is, despite its limited utilization in philosophy in general, a particularly reasonable choice in an investigation of collective irrationality. Any such investigation needs to take into account observations and insights on collective behavior from the empirical sciences, and should aim to feed back into those sciences in turn in the currency of theoretical explication, conceptual clarification and hopefully normative guidance.

1.4 Overview

The number of legitimate ways to approach the challenges posed by collective rationality and its failure is legion. Therefore, there is no choosing without some degree of arbitrariness. A great advantage of this situation is the creative freedom it provides the philosopher with; and this essay certainly takes some liberties to stray away from the more traditional philosophical outlook.

But as any thorough analysis, first some conceptual foundations have to be laid out. The second chapter deals with the general form of arguments to attribute collective rationality or irrationality. What reasons can support a claim to collective rationality? Where are the critical points to look for failures of collective rationality? How does the resulting understanding of group-level rationality relate to the more familiar concepts on the scale of individuals? Are norms of collective rationality fundamentally reducible to those applicable to individuals?

As I shall argue, there are standard components to any analysis of collective action and belief formation, namely the environment the group is situated in, the process governing group behavior – the actual object of evaluation – and the evaluative standards applicable to the given case. To clarify the resulting argument scheme, two case studies are introduced: unpopular norms and belief formation under incentives.

[7]This is not to dispute the fact that there are major inter- and intradisciplinary problems in communicating ABM in a canonical fashion. Efforts like Richiardi et al. (2006) and Grimm et al. (2010) stand witness both to the difficulties and the possibility of resolving them.

These two informal case studies then form the backdrop for chapters 3 and 4. Chapter 3 offers an agent-based model for the evolution of unpopular social norms. Social psychology frequently finds social norms being held up in a population to the interest of almost none of its members, posing not only a normative, but also an explanatory challenge. The model relies on theoretically grounded assumptions on informational limitations and the bounded rationality of its agents to recover such observations and to more clearly state the conditions of their emergence.

Besides providing a potential explanation, it showcases the possibility of apparently collectively irrational behavior to be the result of a generally effective and efficient process. By discussing minor interventions, which greatly influence outcomes, it also cautions the analyst of collective behavior: the (ir)rationality of a social process generally depends on the available and implementable alternatives, and often a comprehensive analysis, and therefore a deductively valid argument for either rationality or irrationality is practically impossible.

Chapter 4 moves from the problems of practical rationality, i.e. choosing social norms to govern one's actions and expectations towards the actions of others to the theoretical[8] problems facing *epistemic* agents in a social setting. Exemplar instances are to be found in the aforementioned problems of opinion polarization on subjects such as climate science, but also difficulties within mainstream science to aggregate and evaluate socially available information in an environment of mixed motivations and limited cognitive as well as material resources.

Once again, the arguments are building on an agent-based model and its analysis by simulation. The major result supported by the model is a trade-off between epistemic performance under ideal circumstances and robust performance under a wider variety of conditions.

Finally, no philosophical essay would be complete without a thorough reflection on its main methodology. As an engineer needs to be able to build her own tools, philosophers should understand the epistemology of their methods. Therefore, chapter 5 discusses the epistemic features of agent-based modeling and simulation, and in particular the kind of arguments stylized formal models support and their relationship to alternative methodologies. Chapter 6 summarizes the results and sketches important future directions for the enterprise of assessing collective rationality.

[8]The term "theoretical" is used here in the sense of the distinction between theoretical and practical rationality: it refers to the rationality of beliefs and other mental states as opposed to the rationality of actions.

Chapter 2

Conceptualizing Collective Rationality

2.1 Motivating Collective Rationality

There are many ways to approach a concept, but maybe the most genuinely philosophical way is by asking for the kind of argument one needs to supply to justifiably apply that concept. What does the proponent have to provide to attribute collective irrationality? How can a group defend itself against the charge of irrationality, or establish its collective rationality?

Taking the perspective of a disputant on a group's rationality has several distinct advantages, most importantly its direct connection to motivation. In the case of individual agents', arguments for the instrumental rationality of a particular behavior often provide immediate motivation to the agent. Though there are clearly exceptions, such as weakness of will, cognitive shortcomings or the insight that one's original ends supporting the argument are in need of revision, an agent realizing their irrationality would usually be motivated to react to the charge.

The same possibility is licensing the argumentative approach to collective rationality, though with some restrictions. In the case of a group, it is not always clear *who* exactly should be motivated to revise their behavior – or even just their belief. A group is only able to act through the actions of its members, and in some cases, no member might feel obliged to serve the rationality of the whole.

This limitation notwithstanding, constructing arguments to motivate the group is in practice a major objective of research on collective irrationality, as a number of applications reviewed in the course of this essay show. Even if the group itself is not motivated by such an argument, the social surroundings of the group might be the adequate recipient. To preview an example, the jury in a criminal trial potentially does not care at all about the accuracy of their verdicts on the individual level; but if there is a strong argument for their systematic incompetence, the legal system can be motivated to take action.

The argument-centric perspective has the further advantage of specifying a context of

justification. As I argue throughout this essay, little is to be said about collective rationality with perfect generality; it is extremely context-dependent. What alternative processes are available to a group and which ends can plausibly be attributed to it changes dramatically with context. Having a target audience for an argument in mind helps delineating the relevant context for a particular study.

Finally, an argument-centric view sidesteps problems associated with the traditional approach of providing necessary and sufficient conditions; arguments naturally provide pro tanto reasons, and hence are not made obsolete by arbitrarily chosen counterexamples. Furthermore, it becomes clear in the next sections that this approach combines particularly well with a modeling methodology, with various models being able to fill the roles required by a generic argument scheme.

This chapter proceeds in the following way: First, an argument scheme for collective (ir)rationality is constructed and its major components, environment, processes and ends are analyzed in some detail, with an eye on formally modeling group behavior. Thereafter, Section 2.3 applies the argument scheme informally to two stylized examples to clarify the details of both its structure and the individual premises. Finally, Section 2.4 addresses the fundamental problem of attributing collective ends, refuting a functionalist view and suggesting more appropriate alternatives.

2.2 An Argument Scheme

Constructing an argument for or against collective rationality requires three main premises: What is possible, what is realizable within the constraints of the possible, and by what norms or standards it is to be judged. This delivers the following structure:

P_1 Given an environment E (What is possible?)

P_2 group G performing a social process P on E (What is realized within the constraints of E?)

P_3 results in outcome $O = P(E)$ judged by norm N with $N(P(E)) \geq N(P'(E))$ for all known alternative implementable processes P'. (Which normative standard applies?)

Cc Then, performing P is collectively rational in E according to N.

As a corollary it follows that P is collectively *irrational* in E according to N if there is a process P' such that $N(P'(E)) > N(P(E))$. Of course, much of this argument depends on spelling out in more detail what is meant by environment, processes and norms or collective ends, a task to be approached momentarily.

With respect to the argument scheme itself, it is generally more difficult to argue for rationality than irrationality, since the former requires to justify a universally quantified premise, while the latter only needs an existential statement. The observation is accurate, but requires a qualification. In practice, it is not necessary to show that it is logically

impossible to find an alternative process yielding better results, but only that there is no *known* process that is also actually *implementable* given the circumstance.

Rationality cannot require the group to perform epistemically inaccessible options. There is some room there for discussion, since rationality might require the group to inquire into possible mechanisms it is currently unaware of, but these details can be put aside for the sake of simplicity. Before such intricacies are to be discussed, the main components of the argument scheme have to be explicated in more detail.

2.2.1 Environment

An environment is constituted by everything externally given under the chosen description of a scenario up for evaluation. As a matter of absolute fact, very few things are given strictly in this way in real-world targets; however, from a modeler's, or more general, observer's or intervening agent's perspective, and even more from from inside the social system under consideration, quite a few features that are not logically, metaphysically or nomologically impossible to change have to be taken as fixed.

Analyzing, for example, the phenomenon of schoolyard bullying, student-teacher ratios, the allocation of students to schools and classes, or the preferences of the children involved will often be taken as a given, despite the fact that all of these can be intervened on in principle. Similarly, the quality of empirical data, research funds or the cognitive capabilities of scientists are often beyond the reach of intervention in social epistemology and therefore have to be understood as features of the environment.

The environment determines the input for the process layer, and it is important to specify it precisely; it restricts what is assumed to be possible, and therefore defines a ceiling achievable outcomes relative to the pertinent evaluative standard. Since rationality demands only to make the best of what is available, e.g. to hold the most accurate beliefs given the evidence available, different environments justify the rationality of different processes. While this might seem obvious, it is rarely stressed sufficiently how much the result of an analysis depends on the details of what can and cannot be changed.[1]

There is no general set of rules to determine what should be considered part of the environment, even if a given framework, such as game theory, often provides the analyst with useful modeling patterns to determine the line between environment and process. Certain variables are more easily endogenized, while others are often left exogeneous for tractability considerations only. In such cases, the argument is obviously to be qualified with the conditional "if the tractability assumptions do not prevent a process superior with respect to N to be implemented".

In many cases, endogeneity varies across applications. Evolving networks provide a good example of a variable shifting from exogeneous to endogenous in different contexts: as Buskens et al. (2015) suggest, treating both networks and behavioral rules as evolving

[1]This is particularly true for formal models, where everything to be part of the scenario has to be described explicitly. An example for dealing with this problem can be found in the discussion of model limitations in Chapter 4.

increases the difficulty of analysis, and hence often requires the analyst to leave either component as exogenous as an idealization.

Besides interaction structure, behavioral rules, agent preferences, signal qualities, available resources, population size (or at least the limiting value) and initial mental states of agents are standard components of the environment. In different frameworks, further features of the scenario under consideration are identified as part of the environment. In the Lotka-Volterra model of predator-prey interactions (Volterra, 1926), coefficients for death, replication and predation as well as the assumption of random encounters all form parts of the environment.[2]

The pragmatics of environment modeling are quite intricate. When models are presented, they often create the impression that most or all decisions to treat a feature as part of the environment or the moving parts of the model were more or less rationally necessitated. Arguments for treating something as given are often derived either from a particular modeling framework or tradition such as game theory, tractability limits or features of the specific target to be modeled and analyzed. As a consequence, it is often necessary to go back and forth between modeling the process and deciding what has to be endogenous in the process model and what is deferred to the environment. Particular mechanisms and processes the modeler is interested in may suggest to exogenize certain features rather than others, once the process is better understood.

For instance, a study of voting procedures normally calls for binary propositional attitudes, whereas belief revision procedures akin to Bayesian learning require the agents to hold graded beliefs. One response to this can be to argue that graded beliefs are more fundamental and introduce additional behavioral assumptions to map credences to binary judgments for voting. Whether this approach makes sense or binary propositional attitudes should just be included in the environment immediately is a pragmatic choice.

As a consequence, the actual differentiation between environment and modeled processes is often an iterative procedure, where the modeler has to revise the details of the environment submodel or adapt the process submodel according to the needs of her investigation. To be clear, the iterative nature of the pragmatics of the procedure does not influence the structure of the final argument, which resides entirely in the context of justification.

Let me conclude on a remark about the formal structure of an environment description. In the argument scheme, little is specified about the internal structure of E. It needs to be able to be treated as the domain of the function or algorithm describing the social process, but nothing more can be required in general. In practice, the environment is likely stochastic, and the investigation is run by sampling from the environment distribution, e.g. a distribution of network structures, and to apply the social process to the sample, evaluating only the results generated from this sample. Therefore, the actual argument constructed is generally inductive, not deductive, since the generic scheme is realized for a sample of environments, not for the entire domain of possible environment configurations.

[2] If a feature as the coefficient for predation was variable instead, a substantially different model family would arise.

2.2.2 Process

The processes, procedures and mechanisms determining a group's behavior form the core of any thorough analysis of collective behavior. When it comes to (ir)rationality, these are the proper subjects of criticism or appraisal. They also delineate, by participation in the process, who is considered a proper group member for the adjudication of collective rationality. Claiming that collection A of agents is irrational means they employ a procedure, are engaged in a process or implement a social mechanism P that is worse than the implementable alternative P' relative to the group's ends represented by N.

Formally speaking, the process component maps a certain environmental input onto a certain outcome that can be evaluated by whatever standard of rationality is appropriate in the case under dispute. Processes need not be deterministic, as the equivocation with mechanisms may have suggested, and often aren't, at least for any observer equipped with limited knowledge.

The analysis of social processes can actually aim at two very different purposes. Much of the analysis (and formal modeling, by extension) in the social sciences tries to capture causally crucial features of *actual* processes to assess them. In that epistemic context, it is important to attain a sufficient match with reality, since an insufficient fit invalidates any inferences to the target system.

But particularly in areas such as the design of markets, voting rules or semiformal and formal knowledge aggregation procedures, the direction of fit changes. A process is analyzed, and if evaluated positively, suggested for implementation in reality. Such models generally still need to make certain assumptions about the behavior of actual agents within the process, because there is usually some part of it beyond the reach of the designers control. However, the core of the process, e.g. an auction system or voting procedure is not supposed to represent anything actual, and thus cannot fail in that respect.

Most commonly, to instantiate the argument scheme, an analyst has to take into account at least two processes: One that is criticized for being suboptimal, and another one that both proves that suboptimality and may be considered as an alternative to be implemented. The first of the two usually falls into the first category above, while the second one often will be a so far fictitious process.

Although this type of comparison might be the most common, there are also examples both for comparisons between two actually implemented procedures, like electoral systems across democratic countries, and multiple fictitious processes to determine which of them is the most advantageous. Such studies are best interpreted as partial instances of the argument scheme, since they explicitly make no claim about the universal quantifiability of their evaluative results.

Turning to the pragmatics of process modeling, it is a common strategy to split the process into rules for individual behavior, interactions and aggregation. Consider, for example, the following sketch of an analysis of jury deliberation: every single juror may update their beliefs by Bayesian (or approximately Bayesian) belief revision; they receive evidence during the trial individually. Later on, they engage in interactive deliberation, a process that could, at least as a first approximation, be described by a model such as the

Lehrer-Wagner-model, where agents assign each others' opinion weight and change their beliefs by calculating a weighted average over the set of beliefs (Lehrer and Wagner, 1981). Finally, they are subject to a unanimous voting procedure for aggregation.

The separation into subprocesses should be understood conceptually, rather than causally. A natural objection to the above model would be that the fact that aggregation takes place under unanimity rule likely impacts the way agents interact with each other before, and maybe even how they revise their beliefs in the light of evidence. But this argument supports a demand for a more precise description of interactions and individual behavior and does not imply that the three parts are conceptually intermixed.

Furthermore, any of the three subprocesses can be empty. In a standard analysis of voting procedures, there is no interaction; agents cast their vote (individual behavior) and the collection of votes is aggregated, e.g. by majority voting (aggregation). In other cases, there is no actual aggregation taking place, as in a decentralized market. Individuals engage in actions and are able to interact with each other, but there is no centralized aggregation in the end. In such cases, if an analyst wants to describe a separate subprocess formally, the algorithm modeling the respective empty subprocess just passes its input on, not posing any problem for the generality of the scheme.

Importantly, this implies that evaluative schemes constructed from the group's ends need to be adapted to the actual outcome's degree of aggregation. In some cases, the normative standard applies to a unique decision or single expressed aggregate belief, while in others, it needs to account for a whole distribution. This point becomes more important when discussing the logical independence of individual-level and group-level norms in Section 2.4.

2.2.3 Ends

A group's concrete ends can vary drastically: from the policy goals of an NGO over the educational purpose of a school class to the ends of a scientific community to grasp important features of the social or natural world, groups' ends differ as vastly as those of individuals. However, when cast in more abstract terms, large classes of ends can be collapsed into more generic evaluative standards. Epistemic groups strive for accuracy, political agents may strive for utility maximization or equality preservation, and so on.

The most difficult task when it comes to the analysis of collective ends is to attribute those ends in the first place. Which goals does a newly emerging political movement pursue? Can the evaluative standard applying to behavior in a school class be derived merely from the students' preferences? I defer the question of attributing collective ends, and focus on the role it plays within the argument scheme, assuming that an end or set of ends has been established for the group under consideration.

The task is to translate a concrete purpose, goal or end into a more abstract normative standard. For example, many epistemic goals, though not all, can be translated into a desire for accuracy, which can in turn be modeled in various ways. Casting the group's ends in terms of more generic, abstract norms is, however, not only useful for modelers: it allows the analyst to abstract from all the irrelevant variations in the way that a group's

goal can be expressed in everyday language.

A highly generic argument scheme enables the observer to capture not only norms that are typically investigated in formal frameworks, such as accuracy and utility. Nothing in the argument scheme prevents one from comparing outcomes to the prescriptions of a catalog of rights the group is committed to uphold. The only condition is that there is a formulation applicable to the outcomes generated by the social processes in question. As a formal sidenote, the argument scheme also assumes that the normative standard implies a partial order over outcomes. This should be understood as a minimal standard to allow for comparisons, which is a necessary precondition for meaningful normative judgment.

However, there is one critical assumption in the way the argument scheme is set up: a normative standard describing the group's ends is not an arbitrary product of the social process in operation. It persists entirely across various possible processes to enable an actual comparison according to the same standard.[3] But social processes in general impact a group's end structure. For example, a political party organized in different ways may end up with substantially different concrete political aims. Is the scheme thus not nearly as general as I claimed?

To answer this question, let me distinguish two classes of such moving targets. First, in many cases, when different ends emerge in concert with different processes, the relevant norm for the argument scheme can be recast as a more general norm. Different interaction structures may lead the members of a scientific collaboration to pursue quite different concrete ends, which may still be cast in terms of accuracy. There is an overarching goal, and the more concrete ends take the form of conditionals referring to the implemented process.

However, a second class of cases is not adequately captured by reference to a unifying, more general end. Imagine a commercial enterprise manufacturing drugs. The purpose of this enterprise is profit, which can, for example, be cast in terms of utility. But at some point, the owners just have enough money and decide to turn their operation into an entirely epistemic enterprise. They no longer aim at selling drugs (even though they might still engage in doing so!), but instead want to further mankind's medical knowledge.

Admittedly, it might still be technically possible to devise a sufficiently abstract measure of utility to incorporate this shift in goals. But a more adequate response is to deny the possibility of judging collective rationality across this shift in goals; a shift that may very well be a consequence of the internal processes within the group forming the enterprise. Another way of phrasing this claim is that the group became a different entity at some point, even though it is still constituted by the same individual agents.

This conclusion may seem unsatisfactory, but it is a consequence of limiting the discussion to instrumental rationality. As for an individual, shifting ends prohibit a unique answer to the question of optimal means. For the case of instrumental rationality, the argument scheme remains fully general. To fill the concepts used in the argument scheme with life, I will now turn to two case studies that will follow me throughout this essay.

[3] I thank Nicole J. Saam for pointing out this assumption.

2.3 Case Studies

2.3.1 Norms of Violence

Norms of violence are an interesting test case for any account of collective irrationality. As Elster (1990) argues, there are norms sanctioning the use of violence in a community to its overall detriment, in his case norms of revenge. Let me preface my discussion by repeating an important observation by Elster: if one allows to stipulate any combination of preference and belief structure, any analyst worth their salt can find a collectively (as well as individually, though maybe not both at the same time) rational reconstruction of any pattern of behavior. This is what makes the task demanding, since one does not only have to provide a coherent analysis, but also ascertain its utility as an account of the behavior in question.

A second important observation is that norms of violence often pose a moral problem in addition to the potential failure of collective rationality. This offers a strong, but also distracting motivation to the analyst: since rationality, unlike morality, is usually something the agent, collective or individual, would try to obey if only aware of its requirements; hence, to offer an argument confirming the irrationality of norms of violence is particularly useful if those norms are to be abolished. If only the group realized how detrimental their accepted norms are, they would be motivated, if maybe not immediately able, to change them. While this is both a morally compelling and rational motivation, it creates an incentive for wishful thinking the analyst has to resist: rational behavior is not by necessity morally right behavior.

With these preliminaries in mind, what is a norm of violence? A social norm[4] of violence refers to any stable pattern of violent action in a given type of situation within a population, together with the expectation to act according to the pattern.[5] That expectation is to be understood both descriptively and normatively. Not only do the norm-followers believe of each other to act violently in the right type of situation, but if they fail there might be sanction. The sanctioning agent is usually not the victims of a norm violation themselves but members of the larger population subject to the social norm. Even if there is no actual sanction, the norm-following agent operates under the assumption that others, if not all others, prefer them to follow the norm. A few examples should be illuminating.

(1) Longstanding international conflicts often take this form. Imagine three countries, Tomainia, Bacteria and Osterlych[6] They have a long-standing tradition of minor border violations returned more or less immediately in kind. Now assume Tomainia violates the border of Osterlych, with Osterlych not retaliating. This might surprise both Tomainia and Bacteria, and prompt Bacteria to worry about the potential weakness of Osterlych to threaten their balance of power. If Bacteria is to still follow the norm, they cannot

[4]For a more detailed discussion of social norms, cf. Chapter 3.

[5]This definitory sketch has been inspired by the definitions offered by Elster (2000) and Bicchieri (cf. 2005, p. 11).

[6]These names are from the famous Charlie Chaplin movie to avoid political disputation of the assumptions.

retaliate in place of Osterlych, and their only option is to try to motivate, by whichever means, Osterlych to eventually retaliate themselves.

Such a pattern can be upheld for a long time if nobody gains the upper hand, and it certainly creates a descriptive expectation. What makes this example interesting is that not everyone in the population has precisely the same normative expectation: Tomainia prefers Osterlych and Bacteria to retaliate against each other, but has no interest that either retaliates against them. Mutatis mutandis, the same is true for the other nations. Such cases should still be considered social norms in the sense introduced above. Normative expectations are not necessarily held uniformly across the population.

(2) Imagine a street gang, frequently using violence to signal strength, and thereby creating a norm of violence (cf. Bicchieri and Fukui (1999)). The pattern of action is imprudent violence against the outside world, e.g. in the form of vandalism.[7] While no individual might prefer to be violent in this imprudent fashion – it is imprudent after all – having built up both descriptive and normative expectations itself may keep up the repeated pattern. Even sanction, in such cases referred to as false enforcement (cf. Centola et al., 2005, for an account), can occur and be upheld as long as everyone sticks to their belief in each other's preferences.

In this case, the norm is a likely candidate for a collectively irrational behavioral pattern. However, it is also possible that the gang can only be held together by relying on violence as a signaling mechanism. This is not the place to provide a detailed analysis of gang violence, but the case exhibits a common thread: often there is more than one plausible mechanism (or social process) to be stipulated, and the (ir)rationality depends on subtle details regarding the implementability of alternative processes. If the gang requires violence as a signal and will disband without and the persistence of the gang is one of it's collective ends, the disadvantages of following the norm may not render it irrational on their own.

(3) Schoolyard bullying is a common problem, threatening both the personal development of children and the efficacy of learning institutions (Salmivalli et al., 1996). Not all bullying is necessarily the expression of norms of violence, physical, mental or social, but social norms constitute a legitimate candidate explanation. The social norm governing violence in a schoolyard regulates who does what kind of harm to whom on which occasion. Furthermore, it also includes subnorms of defense, support and retaliation.

One interesting feature of bullying not as obvious in the previous examples is the role differentiation of the norm. There are distinguishable bullies, victims, bystanders and defenders, and depending on the role assigned, descriptive and normative expectations vary substantially. More generally, several difficulties arise when applying the tripartite argument scheme to norms of schoolyard violence.

With respect to the environment, is there an implementable social norm prohibiting any violence, even if the majority of children did prefer such a norm? With respect to the social process of norm formation, what can be assumed about the children's capabilities

[7] There might be a substantial amount of within-gang violence, too, but potential norms governing those patterns are not the subject.

of judgment and their preferences?[8] Are there children who just *like* to be violent, be it by nature, nurture or any other influences beyond the reach of their schoolyard? Finally, which ends is the analyst justified to project onto the school population as a whole to argue the irrationality of norms furthering violence? Are the school's educational goals a legitimate standard to impose, or does one have to be constructed from the students' preferences themselves? If so, what can legitimately be assumed about these preferences?

A formal account of the process of norm formation is provided in Chapter 3, but an informal sketch of an argument for the irrationality of bullying as a social norm could run like this: children generally have the capability to establish norms of at least low violence, i.e. they actually choose in an environment where little to no violence is an available option in principle. This assumption rests on the combined premises that an average school does have few enough children who are unconditionally violent and enough potential defenders to keep those that exist in check.

However, the children have a limited understanding of the process of norm formation, and have to rely on their own observations largely limited to their friendship clique.[9] School children, like human beings in general, learn by social influence. They try to conform to what they perceive as common behavioral patterns and fulfilling descriptive and normative expectations. In the next step, it needs to be shown that this process frequently generates norms of excessive violence, i.e. bullying.

If so, we are warranted to call that process collectively irrational given that we are able to construct a standard that implies bullying to be a bad outcome. Social norms supporting violence hinder numerous potential collective goals. But it is actually fairly easy to construct a standard restricting violence from reasonably weak assumptions on preferences. If preferences are distributed uniformly along the possible levels of violence, everyone is interested in coordination on a social norm and otherwise self-regarding, equal weighing leads to the average of the distribution of preferred levels as a plausible yardstick for evaluating outcomes.

Thus, a medium level of violence would be acceptable under this standard. If instead the distribution of preferences is shifted to lower levels of violence, the resulting standard requires ceteris paribus a lower level of norm-supported violence to concede rationality to the process of norm formation.

This is how far I will take the informal analysis. However, it allows for a number of observations. First, the norm of violence itself is neither rational nor irrational under this description. It is understood as the product of a process subject to evaluation, not as the process itself. If instead the analyst took norm-following behavior as a social process creating another outcome, e.g. progress in learning, one could ask whether the norm itself would be rational. Either are legitimate questions, but it is important to understand them as distinct.

A second lesson to be drawn is that what looks like collectively irrational behavior in the end result may well be the outcome of a process itself rational in the sense that it

[8] For an analysis of the relevance of moral judgment capabilities in bullying, see (Doehne et al., 2018).

[9] cf. Doehne et al. (2018) for an analysis of clique conformity in bullying behavior

generates on average better outcomes than the alternatives.[10] As holding a false beliefs does not immediately render someone irrational, the fact that a group operates under a detrimental norm at any given point in time does not imply their collective irrationality. It needs to be shown that the agent *could* have followed a better procedure, e.g. one that would have left them with an accurate belief more likely than the procedure they employed.

Similarly, it needs to be shown that the group was not merely failed by their environment or the failure of an otherwise rational process. This is not to say that the group should stick with an outcome they are aware of being suboptimal and they could change. But in practice, they need to implement a better process to achieve, to stick to the example, lower levels of accepted violence. Changing social norms is in practice a difficult and involved process; more than understanding the suboptimality of an end result is required, namely the reform of an underlying process.[11]

2.3.2 Aggregating Knowledge

The establishment, maintenance and change of norms is generally considered a question of practical rationality, since it involves action, choice and preference. But what about the rationality of collective belief, the efficacy of aggregation procedures or organization of behavior in epistemic groups? The tripartite scheme of environment, process and ends applies to such questions quite analogously, only the ends are now epistemic ones.

Before getting into an illustrating example, I shall digress a little on an important difference between individual and collective theoretical rationality. While contested, on the individual level there is a plausible distinction between theoretical and practical rationality. As Lackey (2010) points out, what is practically most rational to accept as a belief may clash with what is best supported by our evidence and thereby seems to be the belief sanctioned by theoretical rationality.[12] But how is the state of belief in a group determined independently from the behavior of its agents?

Consider a standard procedure to determine a group's state of belief: majority voting[13]. The outcome of the vote, and thereby the group belief, cannot be directly determined from the beliefs of the individual members. They might choose to vote strategically, they may fail to cast their vote, or they may actually make a mistake; but the appropriate state of belief to attribute is the outcome of the actual vote. The alternative to this conceptualization of group belief in this example would be to assume that the actual collective belief is an aggregate created by an ideal omniscient observer, or less mythical by an idealized

[10] The relevance of the average across outcomes is an additional normative assumption, that could be replaced by, for example, maximin norm when the worst case outcome is of greater importance to the group's ends.

[11] Bicchieri and Mercier (2014) provides a good example for this: Their approach is to create awareness of the suboptimality of the norm but at the same time being very deliberate about the processes governing norm maintenance and change to actually enable progress.

[12] A classic example is the athlete who would increase their chances by believing in their ability to win, but have actually good evidence they won't, cf. Lackey (2010).

[13] cf. List and Pettit (2011) for a detailed philosophical discussion of the procedure.

process.[14]

But there are good reasons to dismiss this idea. The main argument is provided by List and Pettit (2011), who show that once multiple propositions are in play, there is no non-arbitrary way to construct consistent beliefs from an unrestricted domain of beliefs under anonymity. This also implies that *actual* votes may be inconsistent, but the actual vote is, unlike any suggested idealized procedure, unique.

Furthermore, an analogy to the case of an individual provides another compelling reason to reject an ideal aggregate as the "true" group belief; imagine an agent who holds the prior degree of belief $p(H)$ on hypothesis H, and the likelihood $p(E|H) = q$. Assume, for the sake of argument, that the uniquely rational way to respond to evidence E for that agent is to employ Bayesian updating. Does it follow that the agent then holds the credence $p'(H) = p(H|E) = \frac{p(E|H)*p(H)}{p(E)}$? That seems absurd. The agent could end up with any credence, it just might render him irrational. Similarly, the group as a whole may be considered holding any belief that is actually created by an aggregation procedure the group performs.

The main concern here is that, by applying various procedures, the group seems to end up with different and incompatible beliefs. If that is the case, and there is no reason to belief that something relevant changed in between applying the procedures, the group will count as irrational, since it might be attributed both the belief in a proposition and its negation by this standard. Therefore, in the case of groups, the analyst always has to rely on behavior and therefore at least individual-level practical considerations to determine collective theoretical rationality.

With these preliminaries in mind, let me turn to the case study. Imagine a community of researchers, studying human choice behavior in various simple situations, such as in the field of behavioral economics.[15] Given different theoretical background assumptions, methodological choices, and potentially non-epistemic individual-level goals, how should these researchers treat the results provided by their peers? What would be a collectively rational process to employ? Once again, I provide an informal analysis to be complemented with a formal account in Chapter 4.

The environment in modern day science is very different from more everyday epistemic contexts: researchers generally have to assume that socially available information on the research of others is an indispensable resource. Everyone seriously engaged in scientific research plausibly finds partially, if generally imperfectly, reliable results. Thus aggregating them is in principle desirable. At the same time, most of the time such results are only made available by publication, leaving significant room for the above mentioned factors such as different theoretical background assumptions to partially confound the results.

Around the field of behavioral economics and the psychology of choice behavior, an intricate debate has developed on the actual prevalence of individual irrational behaviors. While prominent figures such as Kahnemann, and scholars such as Thaler and Sunstein

[14]Under different circumstances, other options are open to ascribe collective belief; the group may have chosen an expert panel explicitly to determine their collective belief, to give just one common example.

[15]A paradigmatic piece of work in this field is Kahneman and Tversky (1979), more popular accounts are to be found in Ariely (2008) and Kahneman (2011).

(2008) building on them are convinced of widespread individual irrationality, authors such as Gigerenzer (2015) heavily questioned their conclusion. This is not the place to pass judgment on which account is more accurate, but it shows that contemporary research can lead into discussions where it becomes difficult to aggregate all socially available results due to their inconsistency.

Let me postpone the question of the precise process assumed to be in place, and turn to the deceptively simple question of norms in science. The obvious goal of authentic scientists is to produce (approximately) true insights about the world. As a standard for analysis, this can be rephrased in terms of accuracy, i.e. closeness to the truth. Accuracy can in turn be measured, for example, by employing proper scoring rules (Brier, 1950). However, for practical purposes, one often has to rely on what may be called *proxy* standards. It can take a scientific community years, if not decades, to judge the accuracy of a scientist's results – even if it is assumed, at least for the sake of argument, that such a thing is actually possible.

The community therefore has to rely on alternative standards to assign credit. These are not necessarily the norms the analyst applies in their judgment of rationality, but they are crucial to reconstruct the enterprises of epistemic communities. This leads to the postponed question of the processes in question.

The description of the exchange of research results in terms of opinion exchange suggests itself; however, there is significant design space to analyze and model opinion exchange. A particular interesting feature in the case of a largely decentralized epistemic project such as a scientific discipline, potential interventions are likely to be implemented at the level of incentives for individuals.[16] But there is a serious risk that agents present their opinions strategically. Hence, it becomes an important questions whether there is a behavioral strategy, and thus a social process, that is robustly rational across the relevant environment. Chapter 4 offers a formal argument supporting the claim that there is no such process across important instances of the epistemic environment of science.

But for the time being, it should have become clear enough how to apply the three components of the argument scheme to a given case. It should also be well-understood that there is substantial room to attribute features of a case study to different components, thereby changing the course of the analysis. This is not a weakness of the argument scheme, but an unavoidable consequence of the contextual nature of collective rationality. The available background assumptions to attribute norms, goals and standards or the distinction into an environment and a social process depend not only on the scenario, but also the analyst's epistemic context. With the conceptual framework set up by the argument scheme, it is time to turn to the problem of attributing collective ends; up to this point, I had to appeal to reasonable assumptions and common sense, which is generally insufficient for a compelling argument, and this profound problem needs to be addressed before focusing attentions more on modeling social processes.

[16]This is not to say that the interventions are not *targeted* at structural features such as communication networks, but the implementation often has to run through the pathway of individual motivation.

2.4 Attributing Collective Goals

The problem of end attribution to collections of agents already cropped up multiple times. More precisely, the question is which ends can legitimately figure in an argument on collective rationality for a given group as the ends-component.

To avoid confusion, the following discussion is not committed to the claim that collectives have mental state in any kind of realist sense. For certain groups it seems reasonable to interpret ends as mental states, at least in a functionalist manner of speaking; in human agents, what plays the functional role of ends in means-ends rationality is commonly referred to as desires, and accurately described as a mental state.[17] On the theory of collective ends considered here, as he discussion of collective ends attribution will clarify, collective ends can also be identified with, among other things, the reason for their inception, and hence are not necessarily identifiable with its mental states, metaphorically or literally.

On another clarificatory note, it is important to distinguish between an ex ante and an ex post perspective on means-ends rationality, characterized by different epistemic positions. To an analyst operating ex ante, it may neither be known which exact ends a group has, nor, for any given process, how its going to be realized in detail. Once both the group is identified and the process realized, it is possible that a process ex ante irrational according to the process model and an end attribution based on abstract norms turns out to be an optimal means to the actual ends of the group. Vice versa, a process that is ex ante identified as collectively rational may be realized in such a way as to not optimally further the actual ends of the group.

This does not imply that different concepts of rationality are in play; abstract norms such as consistency of belief or acyclicity of preferences are justified on the basis of concrete instances where violating them is a bad means to actual ends. A money-pump argument to reject cyclical preferences as irrational constructs a concrete scenario where cyclical preferences end up hurting an agent's ends (Davidson et al., 1955). Generic norms of rationality, collective or individual, are supposed to guide behavior and belief before it is known or even knowable which means are optimal, and sometimes even before an entirely concrete set of ends is epistemically accessible. But the defining case is the actual efficacy of means to concrete ends.

The investigation naturally starts with an aggregative approach, i.e. an attempt to construct collective ends from the goals of group members. The limitations of this approach lead up to a functionalist perspective, which turns out to be inappropriate for the normative project at hand. In the face of these difficulties, the discussion concludes with a discussion of several heuristical devices for collective ends attribution reconstructible from basic reason and scientific practice; these devices at least partially resolve the issues of the aggregative approach.

[17] How this attribution is to be understood precisely is a controversial question in the philosophy of mind, cf. Dennett (2009).

	A	B	C
Agent 1	1	2	3
Agent 2	2	3	1
Agent 3	3	1	2

Table 2.1: Preferences for Condorcet's Paradox

2.4.1 Aggregation Failure

The first problem when trying to aggregate individual level ends into collective ones is the need for homogenization: The ends across individual group members may be diverse in kind. This problem can, at least in principle, be solved by translating everything into preferences over accessible states of affairs.[18]

But as Arrow (1950) and many subsequent writers have shown, this preference-based approach combined with a few intuitively plausible conditions implies the mathematical impossibility of coherent aggregation. Let me walk through the simplest recurring example, generally ascribed to Condorcet (cf. List and Pettit, 2011, p. 47).

Imagine three members of a committee facing three possible states of affairs to choose from, labeled A, B and C. The preferences are depicted in Table 2.4.1, with 1 representing the most preferred option. What is sometimes called Condorcet's paradox is the following insight: imagine, that the committee votes first on A and B. Since agents 1 and 2 both prefer A to B, A wins by the majority principle. Now A is compared to C, where C wins. But if the committee went one step further and also checked the numbers on C vs. B, they found out that collectively, $B > C$. Hence, the resulting ordering is circular and therefore, any suggested choice is rationally objectionable on the basis of that ordering. Arrow and subsequent social choice theorists have widely generalized this result. The most famous theorem is Arrow's General Possibility Theorem (Arrow, 1950), which states that there is no aggregation rule that jointly satisfies the following conditions and creates a collective weak ordering over the states of affairs:[19]

1. Unrestricted Domain (UD): All ends, i.e. all preference orderings are allowed to be fed into the aggregation procedure.

2. Pareto Efficiency (PE): The procedure preserves a uniform preference or indifference over any two options. If all individuals either prefer A over B or are indifferent between the two, the aggregate cannot prefer B over A.

3. Independence of irrelevant alternatives (I): If two states of affairs A and B are ordered by the procedure in a certain way, that ordering is preserved under the addition and removal of further possible states of affairs C, C', ...

[18] This is the approach chosen by social choice theory in the style of Arrow (1950). For a survey of classical results, cf. Sen (1986)‘.

[19] For a more technical statement, cf. Morreau (2016).

4. Non-dictatorship (ND): There is no individual agent whose exact preference ordering forms the social preference ordering for all possible cases.

Many social choice theorists and philosophers have since been invested in relaxing such conditions to retain possibility results. However, this is not the subject here. Instead, I take the prevalence of impossibility results as one of several problems this approach generally faces as a universal approach to determining collective ends.[20] Let me consider these problems in turn.

First, any group for which the premises of the General Possibility Theorem hold cannot be attributed the ends that would result from aggregation. Note that this is conditional on all of the premises of the theorem being applicable. One common way out when employing an aggregation procedure to a concrete group is to relax UD: For a known, static group, one could check whether the kind of cycle arises that created the paradox in the initial committee example. For some groups it may also be acceptable to identify their collective ends with that of a dictator; social choice theorists argue against dictatorship assuming that the object of choice is the entirety of arrangements in a society. For more constraint groups, a general – and fundamentally ethical – argument against ends dictatorship is not following from the background assumptions of social choice.[21]

In general, however, it is a priori unknown to the analyst whether the Arrovian conditions hold or fail to hold for a certain group, and therefore an attribution of ends by, for example, majority voting on individual-level ends cannot be assumed to work.

Furthermore, the dynamic nature of group membership in many contexts threatens well-defined aggregation. The collective end structure identified may itself impact the future evolution of the group in terms of members; whose preferences would then have to be aggregated? A case can be made for considering all actual current, past and future members, only actual present members, or even all possible members of the group. Without additional contextual knowledge, it seems hard to recover a single static group for the application of aggregative procedures on a group's constitutive ends.

Next, there are plausibly groups for which collective ends stem from an entirely different source than individual preferences. Criminal juries, for example, plausibly have the collective end to produce accurate verdicts. But even if none of its members shares this goal, the jury as a group remains collectively irrational if it systematically fails to produce the right verdict.[22] Similarly, the ends of a political party cannot simply be broken down to the interests of its members; they extend at least to those of its voters, in addition, but those are not generally participating in the social processes delineating the party as a

[20] As a side note, List and Pettit (2011) have shown that an analog impossibility result holds for aggregating beliefs. That insight is important when it comes to judging theoretical rationality, even though ends in those cases may still be cast in terms of preference orderings. It excludes the possibility to assume that there will be just one unequivocal collective belief to be assessed.

[21] Commercial enterprises with a single owner provide an example; though one might disagree on the moral permissibility of ends dictatorship even in that case, it is prima facie plausible.

[22] Given, as usual, sufficiently benevolent environmental conditions. If the evidence is already completely biased when presented to the jury during the trial, it might be impossible to rationally arrive at the correct verdict.

group.

Note, that this objection relies on the plausibility of treating collectives such as juries or school classes as groups whose rationality can be judged without expanding the group to the larger embedding social environment. The ends attributed to a criminal jury are defined by the legal system, the goals of which are in turn defined by society as a whole. If it is not universally necessary to expand the scope of a group until there are no external factors left influencing its ends, aggregation cannot be applicable to certain groups – namely those not existing for the interests of their members.

As it is in some cases impossible to find any attributable ends by aggregation, in other cases the result is underdetermined.

Imagine two groups of agents, one comprised of software developers supposed to familiarize themselves with a particular platform for a large project, and the other comprised of scientists chasing down the structure of a particular molecule.

One collective goal that can reasonably be attributed to either group from this description is accuracy in their respective beliefs, understood as closeness to truth. For individuals, accuracy is often defined as quadratic distance of a degree of belief from the truth in formal accounts.[23] But given a distribution of individual beliefs, how is accuracy to be applied?

In the case of software development, it is crucial that everyone understands at least the basic properties of the development platform, since the eventual purpose of the team is to write code, much of which is interdependent in the sense that if one module feeds errors into the software system, the whole is flawed and may even fail entirely. Hence, to aggregate individual accuracy goals to a collective one, a maximin-norm is recommendable. Under such a norm, any social process employed is judged by the accuracy of the worst-performing individual, and therefore, this end structure suggests to maximize minimum accuracy as a collective end.

The scientists are in a very different situation; once the molecule's structure is uncovered, it does not matter how far off all the other scientists were – assuming that, once the structure is correctly identified, this result can easily be checked by other experts in the field. Hence, the normative standard to judge social processes in this case is instead by maximaxing: the collectively most desirable process is judged by the best performing agent, since eventually, the success of the project depends on that agent.[24]

What follows is that there is no unique way to aggregate the end of accuracy for a group; end attribution depends on details of the group, and cannot merely rely on their belief distribution and a priori plausibly generic norms of individual rationality.

A more technically demanding argument for this problem is to be found in the discussion of aggregating utilities under uncertainty presented by Mongin and Pivato (2016). They point out that under uncertainty, applying the genuinely collective normative standard

[23]Leitgeb and Pettigrew (2010a,b) offer an argument for this measure. It has been objected to as a unique measure of accurate belief by Levinstein (2012). For the purpose of my discussion, I will assume that there is a uniquely adequate interpretation of accuracy for individuals, since otherwise the collective problem becomes only worse.

[24]Note, that the best performing agent may not be same across possible social processes or even different scenarios for the environment under the same process.

of Pareto-efficiency is confronted with underdetermination: applying Pareto ex ante may lead to substantially different aggregate results than applying it ex post. Unless there is a conclusive a priori argument as to which version of the Pareto principle is adequate, there is once again no uniquely attributable collective end.

To summarize, aggregation of individual ends faces a wide variety of objections. Even though the objections are all limited to certain types of groups, aggregation cannot serve as a universal tool for attributing ends. Therefore, the inquiry has to be taken elsewhere for solutions.

2.4.2 The Intentional Stance

The aforementioned List & Pettit suggest a functionalist perspective to avoid the problems facing aggregation. They refer to it as the intentional stance, thereby referencing Dennett (2009). According to this approach, mental states such as beliefs and desires are to be identified with their functional roles. If a certain behavioral pattern is observed in an individual, the researcher identifies a set of beliefs and desires that would explain it given the individual's rationality. Whichever mental states thereby could explain the behavior are to be attributed to the agent.

Regardless of the predictive merits and demerits of such a view – which clearly are Dennett's main concern –, it fails to support a normative analysis. Rationality is what is to be adjudicated. But to identify a unique set of beliefs and desires[25], rationality is prerequisite. Put otherwise, without a further set of constraints, beliefs and desires are underdetermined by behavior. Therefore, the constraints provided by classical norms of individual rationality, such as the ones List & Pettit employ, cannot be assumed without creating vacuous circularity; thus, the functionalist approach cannot deliver ends in the required sense.

The proponents of functionalism with respect to collective agents partially recognize this problem, since they concede that agents regularly fail to behave rationally, due to the complexity of their environment and their own mental life, combined with various cognitive shortcomings (cf. List and Pettit, 2011, p. 31). However, they offer no solution. Basically, it would be necessary to specify those circumstances where behavior provides the basis to attribute collective beliefs and desires, and then to judge all other behavior according to the standard set thereby.

But this solution runs into two further problems. First, it still fails to address the problem presented by a multiplicity of available norms of rationality. It still has to assume that, for example, there is a clear-cut answer as to whether ex ante or ex post Pareto-efficiency are to limit the collectively rational.

But the choice of reference behavior to determine collective ends is itself unsound. Any criteria to distinguish reference situations from the ones to be judged are objectionable as arbitrary. This is, to note, the crucial difference to a merely descriptive context: there,

[25]Sufficiently unique for the explanatory purpose, since the resulting utility functions representing the agent's preferences are generally only unique up to positive affine transformation (Savage, 1972).

predictive power serves as an external criterion to evaluate categorization; whether a subset or all situations are supposed to enable attribution of beliefs and desires, the adequacy of that choice can be judged by the resulting predictive accuracy. When it comes to adjudicating rationality, no independent measure of success is available *before* normative judgment is passed. Hence, functionalism cannot solve the fundamental problem of ends attribution.

A further downside of List & Pettit's approach in particular is the limitation of its domain: many collections of interacting agents are disqualified as group agents proper. Traders in a common market, for example, are explicitly excluded. But a large part of welfare economics is invested in answering exactly the question which mechanism of resource allocation – a class of which markets are an instance – is most efficient; a question that is interpretable as asking for the mechanism's collective rationality. For any more general answer to the problem of ends attribution, this particular functionalist account cannot offer any help. To which devices are we left then?

2.4.3 Alternative Approaches

With two major candidates for a universal method of goal attribution found wanting, students of collective rationality are left with more limited tools. Those can be understood as heuristical devices, the applicability of which is limited to particular types of groups. Nevertheless, in lack of a more comprehensive alternative, they are useful tools to the inquirer.

The devices presented here are assumed to be reconstructions of both pretheoretical reason and scientific practices of goal attribution. They are meant neither as exclusive nor exhaustive, but to offer an idea of how to choose and justify a method of attribution a posteriori on fallible information about the group in question.

Explicit Statement One obvious approach to attributing ends is to simply *ask* the agent for them. For a number of reasons, this approach is often rejected for individuals. Economists in particular are frequently worried about either the ability or willingness of agents to reveal their actual preferences or beliefs in anything but sufficiently incentivized action.[26] For groups, the situation with explicitly stated goals and ends is different in several important respects.

(1) The form in which collectives explicitly state goals is often not merely verbal, but in more or less official documents. A research group has to state its ends when applying for funding, a party writes them down in a political program, and so on. Therefore, these ends are both less ephemeral, i.e. they generally persist for a substantial amount of time, and can be pointed out to the group who explicitly committed to them. This makes them both explanatorily – through persistence – and motivationally – through referentiability – more useful than the ends stated verbally by an individual.[27]

[26] Gintis (2009) argues for the use of relevant monetary payoffs in laboratory experiments for this reason.

[27] This may go up to the point of actual legal liability, but that is certainly a limiting case.

(2) As I argued above, the assumption that, under at least a limited set of conditions, actions reliably reveal preferences is not supported in the context of group rationality. The very existence of a phenomenon such as an unpopular social norm implies that it cannot be assumed for groups to end up with coherent patterns of action. Therefore, relying on explicit statement can be simply the best method available in the absence of a preferable alternative.

This method is of course limited in its scope: many informal and ad hoc groups the collective rationality of which is of great interest are not constituted in such a way as to provide a statement of their ends. Founders of a company, to give just a simple example, are in general not going to write down explicitly that they aim for profit – or anything else. So explicit statement cannot warrant the ascription of profit goals to commercial enterprises; it seems, however, generally justified to project this end onto a company, because of a different method of attribution.

Attribution by Categorization Many collections of agents can be *categorized*; they belong to a certain class of groups, such as schools, commercial enterprises, political parties or military units. For these categories, one can determine generic ends applicable to all proper members of that class.

Commercial enterprises, for example, are pursuing profit. They may pursue other ends, such as improving public health, creating jobs for unemployed members of the founders' family, or any number of additional goals that are harder to attribute. But if they fail to pursue monetary gain, this particular collection of agents fails to be a commercial enterprise in the first place.

The actual attribution of course needs to rely on a robust inclusion in the relevant category. There are a number of ways to determine membership. In the case of for-profit companies, the analyst can refer to their legal status. In other cases, categorization itself requires an argument based on a group type definition in social science, and those may change due to progress within those sciences. The definition of a socio-economic class, to name just one prominent category, has undergone substantial change since the days of Karl Marx.

A further difficulty with this approach is that it is sometimes unclear whether the analyst should rather give up on putting a group into a certain category or conclude that they are irrational. In many cases, the end they pursue is itself an important determinant of category membership – a problem mirroring the circularity argument against functionalism. In such cases, unless there is an independent determinant, there is no unequivocal conclusion on collective rationality on the basis of ends attributed by category. But our initial examples of schools, companies and political parties, all of which are normally legally constituted as category members, exemplify the relevance of this approach, even in the light of its limitations.

Aggregation Revisited The objections raised before led me to dismiss aggregation as a general method for ends attribution. But there are groups for which the conditions that

support the objections to aggregation fail. A disclaimer is in order at this point: the analyst has to have intimate knowledge of a group to judge whether these conditions hold or not, and therefore this approach is not only limited by the threat of impossibility and inadequacy, but also by the epistemic restrictions of the observer. But a plausible, though fictitious, example is easily found.

Imagine several large sugar companies trying to fix prices. Their situation can be described as an n-person prisoner's dilemma: for any single company, it would be most preferable if all the others asked high prices while they themselves sold for slightly less. As a consequence, if their product is assumed to be equivalent, everyone would buy from them, and they could turn in a large profit. But if all sugar companies followed that logic, they would all end up asking competitive prices instead of the monopoly prices they could ask if all would cooperate to fix the price.

Why is it possible to end up with monopoly-pricing if the would-be sugar cartel decided by majority vote? For simplicity, assume that the companies only consider three types of scenarios: monopoly pricing, competitive pricing and a situation with one defector charging monopoly prices and everyone else asking higher prices, leaving out mixed cases where some ask the monopoly price and multiple others charge less. Every company prefers monopoly pricing to both competitive pricing and any situation where someone else asks less than the monopolistic price while they hold up the higher price. The only scenario preferable to monopoly pricing is the one where the focal company defects while all others cooperate.

Hence, the second-most preferred option for all is monopoly pricing, and everyone disagrees on the most preferred option. Therefore, every other option is beaten in a majority vote by monopoly pricing, and monopoly pricing is the attributed end. Obviously, monopoly pricing is also Pareto-efficient for the cartel. The General Possibility Theorem does not apply due to the restricted domain of preferences, and the problem of aggregation under uncertainty cannot arise because of the underlying assumption of a shared assessment of the relevant probabilities for matters of fact.

With both impossibility and indeterminacy out of the way, there is still the problem whether a group is accurately described as serving some aggregate of its members' interests. This is why a business cartel is such a convenient example: its sole purpose is to further the private interests of the participants.

How many groups are actually analyzable by aggregation? I conjecture that there are very few for which such an analysis is as adequate as that of the sugar cartel, and it is not even necessary to point to impossibility results to support this point – even though they clearly make things worse. Imagine what might seem a typical example of an organization representing an aggregate of its members, a labor union. Historically, labor unions often pursued ends far beyond at least the immediate individual interests of their members. e.g. by also representing unemployed workers to some degree. This creates serious doubt whether an aggregative analysis would do the ends of this type of group justice.

Nevertheless, the case of a business cartel shows that general impossibility results should not let the student of collective rationality reject attribution by aggregation entirely; it just requires sufficient diligence to ascertain its applicability to a given case.

Inspecting the Process The examples discussed to showcase the indeterminacy of collective accuracy as a generic norm point to yet another method to figure out adequate ends: one can take a closer look at the social process. For the two epistemic groups – software developers and molecule-hunting scientists – the larger context of the behavior under consideration is referenced. But what is crucial is to understand certain features of social processes themselves.

For the software developers, the quality of their collective effort is assumed to be highly interdependent; all the pieces of the final complex software need to interlock and operate correctly. Furthermore, in this case it is difficult and costly to monitor the quality of any single agent's work directly to respond appropriately.

In the scenario of the molecule hunters, neither property features in the process: their work is assumed to be perfectly redundant and success or failure easy to detect on the level of individual belief. Therefore, an observer can support the arguments for a particular interpretation of collective accuracy to apply.

This method can only operate conditional on sufficiently detailed knowledge of the process. Furthermore, it may sometimes still not be sufficient to determine collective norms: the description of the two groups is simplified to arrive at a clear argument; in practice, to give just a simple example, the work of scientists is, to a certain degree, interdependent. Publication of a misleading result can hinder the efficacy of other scientists in their pursuit of truth.

Inspecting processes in more detail also opens up the possibility to detect potential sources of irrational outcomes. Failing communication within an organization, the prevalence of biases in deliberation or individual conflicts of interest are all potential symptoms of irrational processes. They are, however, only symptoms; without a previous attribution of collective ends, such observations cannot support a robust inference to irrationality. Instead, they should be viewed as diagnostic tools, relying on the attribution of maximally generic ends in the absence of more detailed knowledge about the group.

To summarize, inspecting the available social processes and the larger encompassing systems can allow the resolution of indeterminacy and therefore establish certain generic norms of collective rationality attributable to a group. In the given example, however, attribution depends on the prior plausibility of assuming accuracy as an end at all. The prior attribution is based on attribution by categorization, showcasing how these heuristics can be combined to surpass at least some of their limitations.

This leaves the somewhat disheartening conclusion that none of the above methods is universally applicable. For each one a description of a group can be found serving as a counterexample to its universality.

2.5 Summary

At this point, two important goals are met: there is a generic argument scheme, which can be filled in to instantiate an argument for any suggested case of collective irrationality. Plugging in the concrete details is, as the following chapters will exemplify, a substantial

task, but having a standard scheme eases the process of reconstructing the actual arguments made based on models, case studies or other potential methods. Our second major achievement is to more clearly spell out the problem of attributing ends to arbitrary collective bodies, and to offer at least partial solutions to this major problem of normative social philosophy.

Having established how to argue for the (ir)rationality of various social processes and mechanisms in groups, a further critical question arises: if the process-focused account given here is appropriate, is it a social philosopher's task to construct such arguments at all? Should it not be left to actual social scientists and psychologists? Insofar as expertise from social science and psychology is necessary, yes, but philosophical analysis can greatly contribute to the project. Philosophers are not confined to use results provided by other philosophers. They can build on results from empirical disciplines, as they historically often did.

What follows from the environment – process – ends account is that philosophers have to engage relevant empirical results when constructing models of environments and processes. The most genuinely philosophical project, evaluating behavior by normative standards, needs to build on process models as well as inspire their form, because the argument scheme requires the output of processes to fit into the evaluative algorithm suggested. In this spirit, the next two chapters provide a deeper analysis of the problems introduced informally in the case studies, but employing the tool of agent-based modeling and simulation.

The construction of models of the evolution of unpopular norms and the impact of strategic behavior on epistemic success in communities should be understood as an attempt to contribute to social theory in general. As any theoretician, a philosopher's arguments, theories and models have to stand the complementary tests of utility and adequacy to the target domain, but there is no reason to be held back by this truism.

Chapter 3

The Evolution of Unpopular Norms[1]

3.1 Introduction

The evolution and persistence of social norms is one of the core topics of social theory. Social norms form the core of grand sociological theories such as Parsonian systems theory (cf. Joas and Knöbl, 2013, chapter 2-4), and more recently, even proponents of rational choice (Gintis, 2009) and analytical sociology (Kroneberg, 2014) emphasize the explanatory importance of social norms. The meaning of the term social norm is therefore broadly construed. In general, it describes a complex of behavioral patterns, expectations and potential sanctions.

Traditionally, social norms have mostly been considered as solutions to problems of cooperation and coordination in social groups (Raub et al., 2015). Rational choice theory in particular often provides descriptions that make successful cooperation in human societies seem almost miraculous. Since the seminal work of Axelrod (1984), various models, particularly computational models, have been developed to explain the evolution and persistence of social norms and cooperation in a world of limited resources and egoism. [2]

But beyond this world of overall beneficial norms that allow their subjects to cooperate in complex social settings, there are also norms that appear to be largely detrimental to their subjects: so-called *unpopular norms.* These are norms that hurt the interests of the majority or even all its subjects, without being coerced onto the population from outside; for example, an enforced code of conduct in a prison does not constitute an unpopular norm, no matter how unpopular it is among the prisoners. If, on the other hand, the prison guards perceive a norm of strictness and hostility towards the prisoners and align their behavior with that norm against their actual private preferences, an unpopular norm has been constituted. In fact, this example is not far-fetched, since social psychologists discovered such a pattern in the U.S. penitentiary system (Kauffman, 1981).

How do such norms arise, and how are they maintained? In some cases, the answer is relatively simple. In a corrupt political system, bribery might be the established social

[1] The content of this chapter is in large parts published in Merdes (2017)

[2] For a recent example in social philosophy, see Skyrms (2004, 2014).

norm. Such a situation is easily modeled as an instance of a prisoner's dilemma: If firm A stops bribing while firm B goes on, firm A will loose valuable contracts and eventually might just go out of business. But, at least potentially, both firms would prefer a situation where neither A nor B would bribe (defect, in terms of the prisoner's dilemma). Such ambiguous cases can be understood in various ways.

One option is, contrary to Bicchieri and Fukui (1999), to argue that such cases should not be classified as norms at all. The expectations of the members of the relevant population might not be sufficient to constitute a social norm. To clarify this, it has to be laid out in more detail what exactly constitutes a social norm.[3] An alternative stance is to allow for some unpopular norms to be fully explained in terms of rational choice, while others require a more subtle psychological and sociological explanation, such as the mechanism of pluralistic ignorance. Case like these often appear quite puzzling to the observer.

Such puzzlement is effected most strongly in cases where unilateral deviation appears to be advantageous to at least some of the agents. And in fact, social psychologists have suggested norms of alcohol consumption at a college campus as a case study (Prentice and Miller, 1993). Their data supports the hypothesis that students systematically misperceive the preferences of their fellow students with respect to the excessive consumption of alcohol. Similar patterns have also been reported on sexual promiscuity (Lambert et al., 2003). In both cases, the behavioral expectation leads students to behave in potentially harmful ways against their own actual preferences. These cases call for an explanation in its own right.[4]

Bicchieri and Fukui (1999), among others, have argued that a major source of unpopular norms is the phenomenon of pluralistic ignorance. In a social group, behavior is often coordinated by perceiving behavior and inferring a preference to coordinate with from the observations. This process of social influence appears to be evolutionary successful, establishing widely accepted behavioral expectations with limited resources. But limited information or misinterpreted behavior can also start a cascade of misrepresentation and failed communication of true preferences, resulting in an unpopular social norm.

Once such a norm is thereby established, it reproduces the suboptimal behavior and maintains a system of misrepresentation of preferences. If a strict system of sanctions is implemented to protect the norm, it can survive indefinitely, as it appears to be the case with certain anachronistic religious norms. But if sanctions are weak, or powerful external events shake up the social system, an unpopular norm is expected to be very fragile. Once successful communication is enabled, the norm's subjects have strong incentive to deviate and dethrone the norm.

[3]There is also an extended discussion about the question whether norms can always be identified with game theoretic equilibria of some kind. It will be argued in more detail in Section 3.2 whether that is actually a reasonable assumption.

[4]It is actually quite difficult to ascertain for a specific case that unilateral deviation would be advantageous, since there might be subtle sanctions that weren't registered by any particular study. However, the above cases of alcohol and sex seem to be genuine cases of advantageous unilateral deviation, since the actual expectations differed so much from the perceived ones, removing the major motive for sanction. Of course, there is also the phenomenon of false enforcement (Centola et al., 2005), but the burden of proof rests with the party offering the less plausible prediction, which appears to be sanction in these cases.

The open question is how to formalize this informal theoretical account of the growth of unpopular norms. In the spirit of generative social science (cf. Epstein, 1999, p. 43): If you didn't grow it, you didn't explain the phenomenon. Agent-based modeling (ABM) provides an extremely useful tool to develop a model from informal theory. The main contributions of this chapter are thus

1. The development of an agent-based model providing a potential explanation of the evolution, maintenance and decline of unpopular norms
2. An analysis of the dynamics underlying the end result of an unpopular norm
3. Qualitative comparisons with empirical results from both survey and experimental research
4. The provision of an easily extensible baseline model, as shown by the addition of central influences

In terms of the three-partite argument scheme for collective (ir)rationality, the model formalizes a set of assumptions on the actions available to norm-choosing agents, their preferences and cognitive capabilities which constitute the the fixed boundary conditions. Within this norm-choice environment, the agents employ a process of restricted observation and communication; extensibility of the model is crucial to ensure its ability to respond to challenges not only to its empirical fit, but also to any normative inferences to be drawn, given the importance of possible alternative processes in the argument scheme. Finally, the discussion will be limited to a simple goal structure projected onto the agents, with pointers to plausible alternatives.

This chapter proceeds in the following way: First, the empirical literature on pluralistic ignorance is reviewed to extract a qualitative benchmark for the model, followed by a critical examination of an existing theoretical model and the concept of social norms assumed for the analysis. Next, a network-based ABM is introduced and analyzed in various simulation experiments. The chapter concludes with a consideration of the model's limitations, a short summary of the results and a review of open questions for future research.

3.2 Evidence and Theory of Unpopular Norms

3.2.1 Empirical Background

Pluralistic ignorance has been shown by social psychologists in a variety of contexts. Students of US universities misperceive the expectations and preferences of fellow students with respect to excessive alcohol consumption (Prentice and Miller, 1993) and so-called hooking-up behavior (Lambert et al., 2003). In the study by Prentice and Miller, the actual preferences on alcohol consumption were approximately uniformly distributed on their scale. The perceived preference of other students, however, turned out to be normally distributed with a significantly higher mean than the preference distribution – where the

preference of the other students is also interpreted by the study's authors as the expectation of subjects towards their peers.[5]

The transformation of a uniform distribution of individual preferences to a normal distribution of expectations can be explained as the students trying to coordinate on a shared normative expectation. The shift in the mean then shows the failure of the social influence mechanism to find the optimum; in that sense, the resulting norm is considered unpopular.

Similar patterns are also found outside college campuses. Prison guards in the United States turned out to misperceive their colleagues as way less liberal than themselves and than they are (Kauffman, 1981). The perception of support for segregation among white men is another case where the members of a group failed to recognize the actual distribution of preferences (O'Gorman, 1975). Bicchieri and Fukui (1999) argue that even norms of corruption can be the product of a social failure to reveal preferences.

Importantly, not all cases of pluralistic ignorance lead to the establishment of an unpopular norm. Bystander scenarios (Darley and Latane, 1968) can be explained by pluralistic ignorance, i.e. the failure to correctly infer another person's mental states from behavior, but the problem is limited to a specific situation. Pluralistic ignorance is a mechanism that potentially creates an unpopular norm, if it is sustained in a population over time. The studies cited above focus mostly on measuring the deviation between private preference and behavioral expectation. While this deviation is a valid operationalization of the construct of an unpopular norm, it doesn't provide understanding of the underlying dynamics of pluralistic ignorance, which are supposed to explain the pattern of expectations.

The study by Shamir and Shamir (1997) provides such a dynamic picture. They investigate the development of the political climate in Israel prior to the election of 1992. They show that private opinion became more and more friendly to the idea of returning territory, but the perceived political climate remained almost stable. Their explanation relies on the limited information on which Israeli citizens formed their opinions, in particular the strong influence of the incumbent conservative government. The spell was broken by the actual outcomes of the election that led to government turnover, therefore revealing the private preferences in a secret election.

Coming from a very different methodological angle, Gërxhani and Bruggeman (2015) observed such a lag in adapting expectations in an experimental setting. Constraints on communication allow a certain perception of social expectation to last when its actual support is decreasing. In both settings, the perceived shared belief is adapting eventually; but at least in the political climate case, a significant change took place only after an important external event. Thus, at least in some cases, an unpopular norm can survive almost indefinitely if unchallenged.[6]

[5]There are a number of motives to accord with other people's expectations: it can be necessary to acquire social status (or not to loose it), caused by fear of sanction or a basic preference for conformity. In fact, the different subjects of a social norm, unpopular or not, can be actually motivated to accord with expectations due to quite different factors.

[6]The case of political climate evolving under pluralistic ignorance does not necessarily constitute a social norm, unpopular or not. While it seems plausible to assume a correlation between the existence of

It is also worth noting that according to Shamir and Shamir (1997), there are two kinds of pluralistic ignorance: absolute and relative. In a case of absolute pluralistic ignorance, the population chooses the option that is dispreferred by the majority. Politics offer a simple example: when the party despised by the majority successfully gets elected in a two-party system, that constitutes a case of absolute pluralistic ignorance.

When a population of students perceives the mean preference for alcohol consumption by 2 points on an 11-point scale, it is in a state of relative pluralistic ignorance. Theoretically, relative pluralistic ignorance is the more fundamental phenomenon, whereas absolute pluralistic ignorance is rather a product of binary choices imposed on a population. It is easy to imagine how a relative misperception of the political climate could be transformed into a case of absolute pluralistic ignorance via binary choice: the misperception just has to shift the behavioral expectation across a party divide.

From these empirical insights, one can derive some important qualitative benchmarks for a simulation model that tries to explain unpopular norms by pluralistic ignorance:

1. It must be able to reproduce the basic pattern of a deviation between the mean of the preference distribution and the distribution of norm perception.

2. It should include variable constraints on the available information to constitute a model of pluralistic ignorance.

3. Ideally, it reproduces the dynamical behavior qualitatively, i.e. it exhibits the above-mentioned lag of the dynamic of the norm expectation distribution behind the preference distribution.

There is one feature of all these empirical studies that has not yet been mentioned, since it does not concern their descriptive content or explanatory value. The issues investigated all seem to be of high *normative* relevance. Excessive alcohol consumption, sexual conduct, the penitentiary system and Israel's territorial policy can all be considered ethical problems, too. While the authors do not explicitly claim anything normative, they can consistently be interpreted as ascribing collective irrationality to a group following an unpopular norm. If that reading is appropriate, it suggests that seemingly ethical problems are in fact failures of rationality and could be resolved by improving collective judgment. These potential moral implications are not the main concern of this chapter, but the discussion of the simulation results shall revisit this important issue.

3.2.2 Theoretical Models

The landscape of theoretical models of the evolution of unpopular norms is rather sparsely populated compared to its empirical counterpart. In fact, there are only two attempts to model that process formally to the best of my knowledge. The first one is a model by

certain social norms and a particular political climate, the study cited does not explicitly establish the existence of norms, but rather offers longitudinal data on pluralistic ignorance.

Centola et al. (2005) that relies on false enforcement as its main mechanism. False enforcement denotes the phenomenon that agents might enforce a norm they privately disagree with to prove their allegiance to the rest of the social group. While false enforcement is a potential explanation for unpopular norms, the mechanism is very specific and hard to detect empirically. None of the above-cited empirical research offers any evidence of false enforcement, making it relatively hard to validate or refute and improve this particular model.

The second model has been suggested by Bicchieri (cf. 2005, chap. 6). This model, referred to as BM from here on, relies on pluralistic ignorance in a binary choice situation only. The agents represented in BM might, for example, have to decide to litter the street or keep it clean, both of which could be the norm in a given neighborhood. The chosen action of an agent i is denoted by $x_i \in \{x^1, x^2\}$ which take on the values 0 and 1. Among other implications, the binary choice blocks certain impossibility results such as Arrow's theorem, which assume at least three options (Arrow, 1950). In addition, the agent has a private preference y and a degree of nonconformism β.

There is a "true majority" with respect to preference: any agent is a member of that majority with probability p. There are two kinds of agents: Conformists ($\beta = 0$) and trendsetters ($\beta = 1$). In addition, trendsetters lose θ if they suspend their choice. The overall utility of an agent's choice is described by the function

$$U_i = -(x_i - \hat{x})^2 - \frac{\beta}{2}(x_i - y)^2 - \theta \tag{3.1}$$

where $\hat{x}$ is the perceived majority. In accordance with the utility function, the model runs in two steps: Trendsetters have to make their choice in step 1 to maximize their utility because of θ, which is chosen appropriately. They have no majority to perceive and thus their choice is effectively guided by their y. In step 2, conformists make their choice. For a conformist, the formula simplifies to $-(x_i - \hat{x})^2$. It depends only on the perceived majority.

Bicchieri has shown that informational cascades take place in this model, both positive, i.e. revealing the true majority, and negative, leaving the population in a state of absolute pluralistic ignorance. In addition, it is possible that trendsetters do not provide a clear signal to conformists. In such a case, conformists fall back on a coin flip, resulting in a partial cascade.

While this model offers a first impression of how to formalize pluralistic ignorance to grow an unpopular norm, it has some significant shortcomings. The model does not meet the first of the above-stated benchmarks, since it cannot reproduce a pattern of relative pluralistic ignorance. The second benchmark is met, since the model successfully endogenizes informational constraints in its utility functions. However, such a conceptualization makes it difficult to model a structurally more complex communication environment. It would be desirable to decompose what is compounded in the utility function in BM, namely structural features of the informational environment, psychological features of the agents and variables that are to be understood purely in terms of rational reconstruction.

Finally, the model does not exhibit any real dynamics. There is only a two step process that does not allow for a more fine-grained analysis of the dynamics of the system. It is too

restrictive to assume trendsetters that are psychologically different. The role of an agent is often determined by structural features of the social environment rather than individual psychological features. That is not to say that individual differences aren't potentially important for the process, but they shouldn't be assumed to be a necessary condition for pluralistic ignorance.

These limitations notwithstanding, BM provides a starting point to develop a more sophisticated model of the process in question, a task the model proposed here is trying to accomplish. But before the discussion can proceed in that direction, it needs to be stated explicitly how the concept of a social norm is understood.

3.2.3 Social Norms

Given the reliance on Bicchieri's model of pluralistic ignorance, her definition of social norms is a natural starting point. According to that account (cf. Bicchieri, 2005, p. 11), a social norm is a behavioral rule R applicable to a situation S representable as a mixed-motive game. It exists in a population P, if there is a sufficiently large subset $P_{cf} \subset P$, such that for each $i \in P_{cf}$, the following conditions hold:

1. i knows about R and its applicability to S.

2. i prefers to act according to R if i beliefs there is a sufficiently large set of agents who follow R in situations of type S and an analogous (not necessarily identical) subset of P exists, which expects i to conform to R in situations of type S, potentially also willing to punish i for violating the norm.

This definition certainly encompasses some of the important features of social norms also noted, for example, by Elster (2000). They are socially shared, i.e. not simply individual habits, they come with normative expectations unlike simple collective patterns of behavior, and they do not require codification. In fact, the expectation of sanction is not a necessary condition for a social norm to exist. Note that, according to this definition, a norm can exist but not be followed, since the preference is conditional. While this definition is prima facie quite plausible, there are two major points in need of closer inspection.

First, it offers a *rational reconstruction*, which, according to Bicchieri (cf. 2005, p. 11f.) implies that the beliefs and preferences mentioned in the definition are not necessarily identifiable with beliefs and preferences within the agents. At a first glance, that is an important concession to remain plausible, since Elster (1990) argued extensively that some norms are not plausibly explained rationally, in particular certain norms of revenge.[7]

But this concession of course also raises the question what it actually is in any given concrete case that explains the existence of a norm, and the motivation to follow it. This is not a problem in principle, but it highlights an issue potentially concealed by a rational reconstruction: there are likely numerous different explanations for various cases of social

[7]Note, that Elster himself concedes that *some* rational reconstruction will always be possible. But in some cases, this reconstruction will not be explanatorily valuable.

norms, and to offer an actual explanation, one has to specify what takes the place of the rational implementation of preferences in the rational reconstruction.[8]

Second, the assumption that norms only govern phenomena that are faithfully represented by mixed-motive games is too restrictive. Gender-role norms, for example, are often to be represented by asymmetric coordination games.[9] A lot of norm-governed behavior might also be better described by a decision under risk or uncertainty without a strategic element, where the social component enters only through an additional norm *requiring* an agent to punish.

This idea of a norm that requires an agent to sanction is not dissimilar to that of a meta-norm (Axelrod, 1986) requiring agents to punish agents who fail to sanction norm violations, since it is logically independent of the actual norm.[10] This implies the possibility of a scenario where a social norm on non-interactive behavior is in place, but the punishment norm, which would allow to interpret the whole as a mixed-motive game, does not exist in the population. As an example, consider norms against masturbation in a world where no one is expected to actually sanction it. One would have to construct a seriously convoluted argument to represent the norm-governed situation as a strategic interaction.[11]

Elster (2000) avoids such problems by giving an account in terms of what is *not* a social norm – i.e. legal norms, habits, moral norms – and providing paradigmatic norms – social norms against cannibalism or incest – to offer positive guidance of what is to be understood by the talk of social norms. Where does that leave the modeler, intending to explain the emergence of a particular kind of norm? While there is no convincing philosophical definition stating the necessary and sufficient conditions available, one can build on Bicchieri's account by relaxing the problematic components and adding the caveat that it might not offer a perfect demarcation criterion for the class of phenomena accurately referred to as social norms.

Therefore, for the purpose of our analysis, a social norm is a behavioral rule R for situations of type S, for which the second condition in Bicchieri's definition is fulfilled. In addition, this rule can allow for some "slack": Even if a situation is classified as an instance of S, there can be room to behave slightly deviant without violating the norm. This is a useful clarification for cases of fuzzy norms.

For example, norms of fairness often do not exactly specify what a fair split is, as experiments based on the ultimatum game suggest (cf. (Roth et al., 1991)). This fuzziness is likely a tribute to the limitations of human norm-followers, both with respect to their epistemic powers and their ability to control their actions.

[8] In this essay, I will actually follow the suggestions towards the end of Elster (2000) and rely on motivation by conformism; however, this is probably not universally adequate, and can also be dissected more closely into conformity due to factors like fear of sanction, striving for status or simple habituation.

[9] It is also not apparent how role-differentiated norms in general are to be stated in this terminological framework, but I will assume for the sake of argument that it is in principle possible.

[10] Note that Axelrod's definition of a social norm differs significantly, since it only refers to the frequency of norm-following and punishment. While punishment is not necessary according to the definition employed here, behavioral patterns are not sufficient.

[11] It might not be impossible to do so, but importantly, it would not be a faithful model of the type of situation.

A final remark on this notion of a social norm: it may, for any practical case, be difficult to establish the existence of a social norm, since it is impossible to spell out the meaning of "sufficiently large subpopulation" unambiguously. But ambiguity seems unavoidable, since there is, analogous to the concept of a heap, no non-arbitrary amount of deviation marking the applicability of the concept of a social norm. But for the purpose of explaining clear-cut cases of unpopular norms, potentially contentious borderline cases do not pose a problem. Hence the discussion can now turn to a formal model.

3.3 Model

The basic process of the model consists of agents entering a social network and choosing their observable behavior from a range of options. They know the behavior of their neighbors in the network and their own private preference, but neither the global distribution of choices nor anyone else's private preferences. The theoretical assumption behind this model outline is the bounded rationality of the agents: they have limited information and limited capability to make their decision, and therefore rely on a simple behavioral heuristic to consider their preferences and the cues provided by a small number of other agents.

The model description therefore encompasses two major subprocesses: the decision rule for the agents and the network growth algorithm determining the evolution of the social structure.

3.3.1 Decision Under Social Influence

Every agent has two properties: their private preference, which is assigned on initialization, and their chosen behavioral norm, which is visible to other agents. Which agents actually perceive an agent's norm choice is defined by the network structure. Since the model is meant for relative pluralistic ignorance, the preference is chosen uniformly at random from $[1, \ldots, 11]$. This kind of scale is able to represent the commonly used Likert-scales from research on pluralistic ignorance and unpopular norms. When an agent enters the network, they have to choose a behavioral norm, again from $[1, \ldots, 11]$.

The goal of the choosing agent is to accord with the prevailing norm. Since the agents themselves are part of the population at large, their own preferences have to figure in the decision. The rationale behind this is that the agent tries to fulfill the social expectations based on other agents' preferences. They cannot access them directly, and therefore use their behavior as a proxy. The only exception is the agent's own preference, which is assumed to be known.

This leads to the stipulation of the following utility scheme: For an agent i with the preference $p_i \in [1, \ldots, 11]$, the behavioral norm choice $c_i \in [1, \ldots, 11]$ and the reference population of agents N, the goal function to maximize is defined as

$$G_i(c_i) = -\left| \frac{1}{|N|+1}(w_i(i)p_i + \sum_{j \in N} w_i(j)c_j) - c_i \right| \tag{3.2}$$

where $w_i(j)$ is a potentially time dependent function describing the weights assigned by agent i to another agent j. This equation is very similar to the procedures proposed by Lehrer and Wagner (2012) or Hegselmann and Krause (2002). Their models are targeted at opinion dynamics, but at least the descriptive component of normative expectation can be redescribed in terms of belief. The main structural difference to these models is separation of a privately held preference and an observable choice. This differentiation is what allows the representation of a failure to reveal the private component to coordinate efficiently. Otherwise, the process can be understood as a special case of belief formation under social influence.

Within the model, some constraints on Equation 3.2 are necessary. First of all, it will be assumed that all the weights are equal, representing that the agents have no reason to give a particular agent more weight. In an environment with actual power differentials, this would be an invalid assumption, but in the model environment, it is a reasonable idealization. Second, the agents' information is limited to their network neighbors. This constraint is in fact a crucial component of the model, and leads to the localized version of conformism that agents are actually able to implement:

$$G_i(c_i) = -\left| \frac{1}{|\Gamma_i| + 1}(p_i + \sum_{j \in \Gamma_i} c_j) - c_i \right| \tag{3.3}$$

where Γ_i denotes the neighborhood of i. The problem with using this Equation as a proxy for the actual payoff is the potential for admitting too much influence by extreme positions; this is a simple property of the average, and since the average to be taken by the agents would be only on a restricted amount of information, choosing on this basis might exaggerate conformity effects. To avoid this problem, the actual behavioral rule is determined from

$$G_i(c_i) = -\left| \frac{|p_i - c_i| + \sum_{j \in \Gamma_i} |c_j - c_i|}{|\Gamma_i| + 1} \right| \tag{3.4}$$

which informally is the average distance instead of the distance from the average. A further advantage from the point of view of bounded rationality is that this proxy utility is maximized by the median which requires less arithmetic than computing a mean. Equation 3.4 defines the actual choice behavior of the agents under social influence.

Note that an agent without a reference group will always behave on the basis of their private preference only. Thus, the decision rule is homogeneous across the population, including an initial set of unconnected agents. This avoids the assumption made in BM that trendsetters are psychologically special. Instead, the difference is explained by social structure alone. Both pure explanations are likely incomplete, but since representing social structure explicitly is one of the main motivations for the current model, the design focuses on this aspect omitted in BM.

It is not necessary to imagine the decision as conscious, despite the talk of computation and choice. The model is supposed to be instrumentally valid, but the choice procedure can also be interpreted realistically in various fashions. Given the interpretation of c_i as

the choice of a behavioral pattern potentially sanctioned by a social norm, the decision procedure can be interpreted in at least three different ways.

First, it can represent an agent's anticipation of sanction for deviation. This anticipation creates an incentive to conform as closely to the perceived norm as possible, given the agent's limited information. Less rationalistic, the agent could also have a metapreference to conform to its neighbors for the sake of coordination itself. Finally, the agent could also assume that other agents are onto something, in the sense that their behavioral choice might be more beneficial. This interpretation of course implies that an agent's preferences is not everything there is with respect to judging the efficiency of the established norm, which is likely true for some subject matters, but shouldn't generally be assumed and goes strictly beyond the assumptions used here.

3.3.2 Network Growth

There are various different processes to construct networks with certain statistical properties. A problem with classical random graph models (Erdös and Rényi, 1960) or small-world networks (Watts and Strogatz, 1998) is that they are not designed to grow dynamically. While it is possible to reformulate them in a way to address this issue, there is also an elegant and empirically more plausible alternative in the preferential attachment model (Barabási and Albert, 1999).

The preferential attachment algorithm starts with a small initial network of m_0 nodes. Additional nodes then become iteratively attached to the network by randomly generating m edges with existing nodes. The new edges are not chosen uniformly at random, but the probability depends on the current network structure. More precisely, the more network links an old node i already has, the more likely it is that a new node will attach itself to that node. Formally, the probability that a freshly created link includes i is defined as

$$p_i = \frac{k_i}{\sum_j k_j} \tag{3.5}$$

where k_i denotes the degree of i. In the model of unpopular norm growth, this algorithm is slightly modified by adding the constraint that there are never two links between two agents. This constraint is added to retain the assumption of equal weight in the decision rule. Besides its algorithmic simplicity, this procedure has several further advantages:

1. Its statistical properties are quite well understood. For example, Barabási and Albert (1999) have shown that the degree distribution approaches

   $$P(k) \propto k^{-\gamma}$$

   with the parameter-dependent constant γ for large numbers of nodes.

2. It has been shown that the degree distribution generated by this model fits well with empirical data on certain real-world social networks (Barabási and Albert, 1999; Albert and Barabási, 2002).

3. There exist useful generalizations of this model that make it possible to adapt the growth algorithm to different applications with special requirements (Albert and Barabási, 2002).

As the decision rule under social influence, the preferential attachment procedure lends itself to various behavioral interpretations. One can just understand it instrumentally, but an advantage of this model is the possibility to provide multiple realist interpretations. The two most relevant interpretations for our purposes are a more rationalistic and a psychological version.

From the perspective of rational choice, it can be an efficient heuristic to acquire information to only inquire the choices of the most well-connected members of a network. The assumption is that (a) creating links is costly and (b) well-connected nodes are statistically more representative with respect to the relevant information. These assumptions are not explicitly implemented in the overall model, but they justify preferential attachment as a potentially efficient heuristic.

An alternative interpretation assumes that human agents tend to prefer to connect to well-connected individuals. Being central in one's social network often represents high status, power or charisma, making it desirable to be socially connected to a highly connected individual. For the purpose of modeling, it is sufficient to note that there are multiple plausible interpretations of the behavioral rules employed. Any real-world system following preferential attachment will likely include a mixture of social and psychological mechanisms implementing preferential attachment without challenging the validity of the model.

3.3.3 Simulation Algorithm

At this point, the parts of the model can be assembled: It is initialized with a set of agents who decide on their behavior purely on the basis of their preferences. They become connected afterward to make the probabilities for the preferential attachment algorithm well-defined. Note again that agents are not heterogeneous with respect to their decision functions, only using different information. After the initialization step, agents are successively attached to the network and make their decisions according to the information they acquire from their initial neighborhood to create an overall distribution of norm choices. The following pseudocode specifies the model's behavior:

1. Initialize m_0 agents' preferences by drawing them uniformly at random out of $[1, \ldots, 11]$, and choose the preference as the observable behavior.

2. Connect the initial set of agents.

3. Repeat:

   (a) Create a new node i, with its preference p_i chosen uniformly at random.

   (b) Add i to the graph by preferential attachment.

(c) Compute the median behavior choice in i's neighborhood to determine c_i.

A few remarks on the simulation model are appropriate before the analysis turns to its surprisingly complex behavior. The variables that a simulation experiment can directly intervene on are m_0 and m, thereby to some extent controlling the quality and amount of information available to the agents that enter the model later on. However, the stochastic elements of the model are influential, requiring the simulator to run large numbers of simulations to identify interesting cases; to a significant extent, the explanatory factors are to be found in the early dynamics of a model run rather than in the model parameters. Therefore, the analysis will start out with a number of single-run analyses to exhibit the relevant qualitative behavior, followed by a thorough exploration of the parameter space and the crucial factors leading to an unpopular norm.

The results in the following section are based on a Python implementation of the algorithm above which can be accessed at `https://www.openabm.org/model/5289/version/1/view`.

3.4 Results and Discussion

3.4.1 The Emergence of an Unpopular Norm

The first question to be answered is if an unpopular norm can actually emerge in the model. The pattern that serves as a qualitative benchmark for an unpopular norm due to pluralistic ignorance is provided by the alcohol consumption norms reported by Prentice and Miller (1993): There should emerge an approximately normal distribution of behavioral expectations and therefore choices from an approximately uniform distribution of preferences, but with different means.

Figure 3.1 depicts the result of a simulation run exhibiting the described pattern. The parameter configuration is $m_0 = 40$, $m = 10$ and the simulation is run for approximately 10^3 timesteps, thereby increasing the population size to 1000 agents. The distribution of behavioral choices has the form of a steep binomial distribution that has the normal distribution as its limit. The steepness signifies a high degree of coordination and the difference of 0.73 between the means of the two distributions represents the unpopularity of the norm on which the population coordinated.

For a complete interpretation, let me reconsider the concept of a social norm that has been assumed and match the components of the definition with the results. The possible choices should be understood as various options for actions, available to become rules for behavior. The conditional preference for norm-following is assumed in the agents' decision rule: they try to estimate a rule which factually has the most followers, and by their decision rule they assume others to expect to follow and are therefore motivated to follow the rule themselves, as expressed by their particular choice.

The final piece, and what is revealed by the simulation, is to show that there is a sufficiently large part of the population sharing the same beliefs on actual followership and expectation. The distribution of choices in Figure 3.1 exhibits this high degree of shared

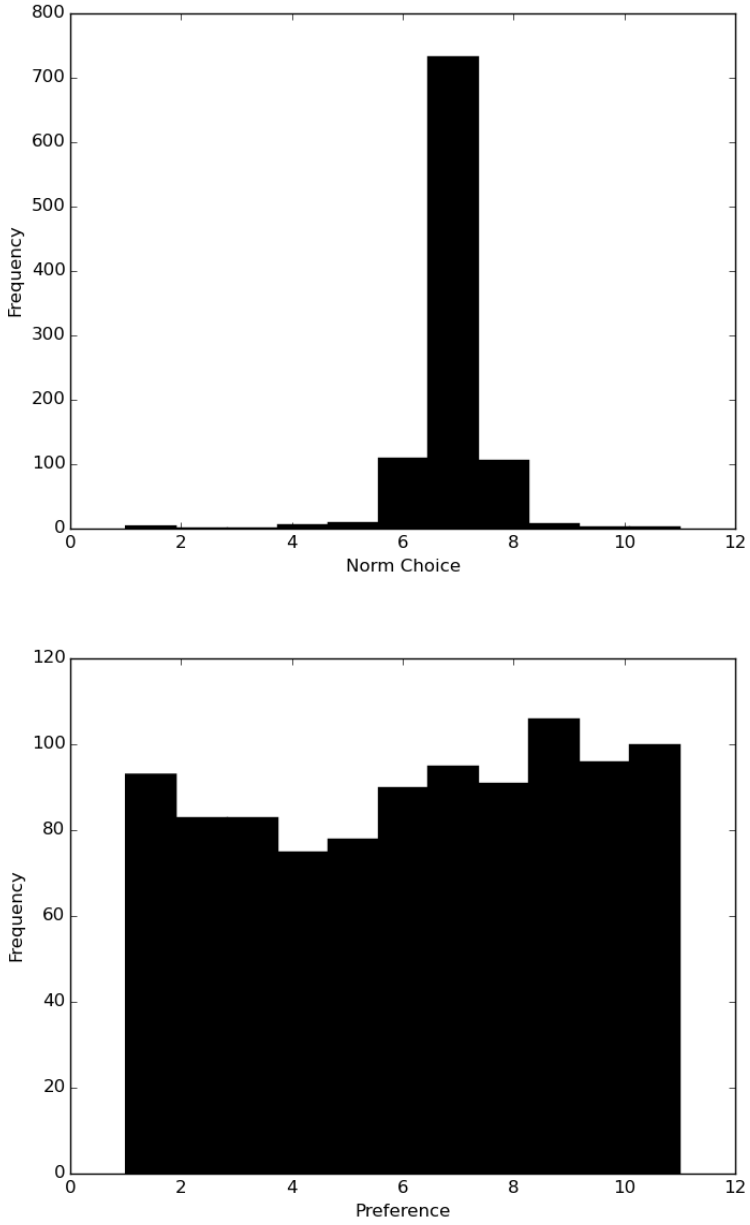

Figure 3.1: Simulation results depicting the difference between the distribution of choices (upper figure) and preferences (lower). Note the difference in scale between the two distributions.

belief, and thus clearly is not one of the potentially problematic borderline cases for the assumed definition of a social norm. If, in addition, the norm is allowed to be slightly fuzzy, the agents only off by one in the spectrum of potential options can even be counted into the subpopulation of norm-followers, making the case for the emergence of a social

norm even more clear-cut.

The results of a single run are useful to showcase the model's behavior, but a more detailed exploration of the distribution of outcomes is provided in Section 3.4.3. Before the discussion turns to a more detailed analysis of the dynamics of unpopular norms and the overall behavior of the model, a few remarks on the interpretation of this first result are in place.

First, it is at this point easy to see how a binary choice imposed on the simulated population could generate absolute pluralistic ignorance. Imagine that the scale of choices and preferences represents a simple political spectrum from left to right. But whereas political preference is diverse, the political system supports only two parties. Now assume that everyone votes for the party that is closest not to his or her preference, but to their behavioral expectation for the population, for example because the election is not secret.

Furthermore, assume that party A's program exactly mirrors the mean of the preference distribution (6), but party B is more radical and positions itself at 7.4. In this scenario, the majority of agents would vote for B, therefore creating a choice that exhibits absolute pluralistic ignorance. A more extreme case of pluralistic ignorance could lead to an even more extreme situation, and is exacerbated by the assumption that everyone votes for the closest party – an assumption that is not obviously true, but quite standard in voting models (Black, 1948).

Second, the level of coordination achieved in the simulation is quite surprising. There is no belief revision involved, and there is a steady inflow of conflicting private preferences. The communication structure of the model seems to enable such a high degree of coordination with very small effort. This relationship between network structure and outcome will be explored more closely in Section 3.4.3, but it is crucial to understand how a social process can become established that *sometimes* produces highly suboptimal results: It does not deterministically produce inefficient outcomes, but oftentimes allows for successful coordination using a minimum of both cognitive and communicative resources.

3.4.2 The Dynamics of Pluralistic Ignorance

One of the advantages of a simulation model that successively emulates a social process is access to its dynamic features. These features can be compared to known empirical properties of the process to increase the model's credibility. The empirical finding referred to is the lag of change in norm perception and adaption behind changes of the distribution of preferences.

The model has no mechanism for change in attitudes or preferences, but the actual distribution of preferences approaches the distribution defined by the underlying generative procedure over time, due to the growth of the network. Likewise, the distribution of norm expectations evolves dynamically with the addition of new agents. If pluralistic ignorance occurs in the model and potentially leads to an unpopular norm, a lag as described by Shamir and Shamir (1997) and Gërxhani and Bruggeman (2015) should be observable. This is indeed the case, as Figure 3.2 shows.

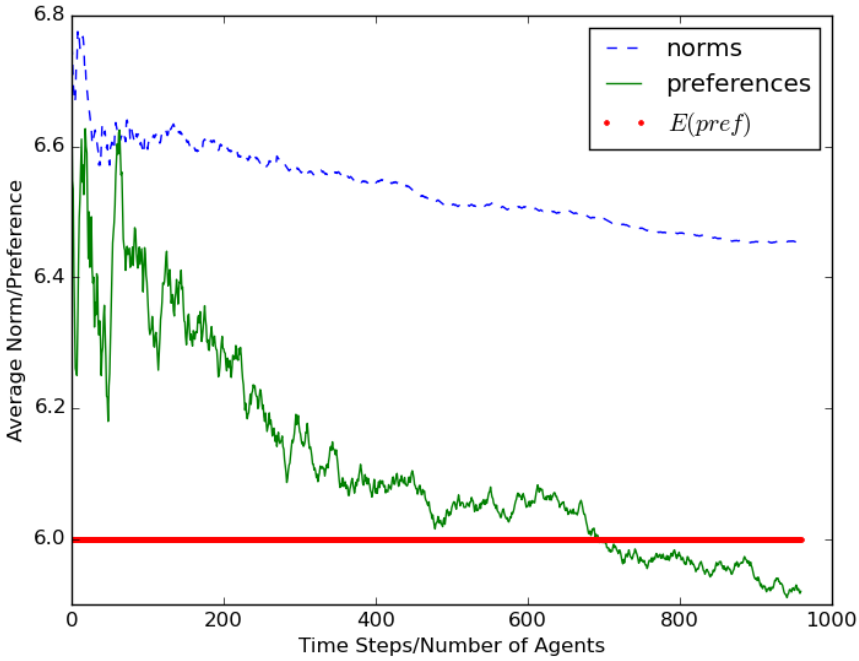

Figure 3.2: Example dynamic evolution of an unpopular norm generated with the simulation model.

The evolution exhibits the characteristic lag of the perceived norm behind the preferences. The mean of preferences over time approaches the mean of the underlying uniform distribution.[12] The perceived norm, however, lags behind this development and gets stuck on a suboptimal option. After a highly volatile initial phase, the lag becomes visible: The mean of the actual preferences starts to drop (modulo random variation) towards the underlying distribution's mean. At this point, the average of choices also starts to drop, but way slower, remaining at a significantly higher value for the rest of the observed time interval, while steadily decreasing in the direction of the preference distribution's mean.

This fits observations reported by longitudinal empirical research. In the example of the political climate in Israel, private preferences evolved towards what became the actual election results and came fairly close, while the perceived political climate did not change significantly until the election results became public. Since such an external event does not occur in the model system, it reproduces the process until the election, where the preference distribution changes without an appropriate change in the perception of the situation.

[12]The mean of realized preferences actually drops below the theoretically underlying mean in this particular case, as a result of variation in the underlying pseudo-random number generation process. Note, that the deviation is actually very small (< 0.1), which is not an unlikely deviation from the actual mean for a medium-sized sample.

The source of this time lag in the model is the combination of conformism with a lack of belief revision. As Gërxhani and Bruggeman (2015) argue, belief revision in the real world is delayed by uncertainty. This is very much in line with both the observation by Shamir & Shamir and the simulation outcomes, and gives an overall justification of the assumption that there is no belief revision on the simulation's time scale. The process starts out with a non-representative initial agent population with respect to their preferences.

They form a norm that fits their distribution of preferences. Agents entering the group later on adapt to their norm due to social influence. Then, over time, the preference distribution approaches the underlying uniform distribution. Notably, the distribution of norm expectation evolves in the general direction of the preference distribution but is inhibited by the implicit conservatism of the underlying mechanism and the informational restrictions imposed on the agents.

The model is also able to reproduce other scenarios than the emergence of unpopular norms. Two instances of convergence of the norm to the underlying preference distribution are depicted in Figure 3.3. It is interesting to note the differences between these scenarios. In the run depicted in the upper graph, an initial misrepresentation of the population-wide preference distribution creates a deviation of the behavioral expectation, and therefore a suboptimal norm.

However, the mean of the instantiated preferences then drops sharply, and pulls the norm towards the mean of the underlying preference distribution. Thus, in this case, the shift from one misrepresentation to the other allow the mean of normative expectations to converge comparatively quickly to the population-wide mean of preferences. It is important to note that the deviation started out relatively small, making the initial situation of ignorance easier to overcome.

More surprising is the case depicted in the lower graph of Figure 3.3, where the convergence of the mean of preferences to that of the underlying distribution lags behind the norm as the mean of norm expectations is approaching the underlying preference mean. This can be understood as inverse pluralistic ignorance, since the initial nodes of the network are more representative of the underlying distribution of preferences than the subpopulation that exists after a few dozen time steps.

This suggests the possible normative result that the social process described by the model is not in and for itself collectively irrational as it might appear. For some configurations, the results are maybe even surprisingly close to the stipulated optimum, in particular given the uncertainty. The discussion will return to this normative aspect in Section 3.4.5.

The capability of the model to generate a variety of phenomena surrounding social norms very different from unpopular norms of course raises the question under which conditions the process runs into a problematic, since unpopular, norm and when it generates a reasonable or even favorable outcome given the assumed preference distribution. To address this issue, it has to be investigated which variables within the simulation model can explain pluralistic ignorance and thereby the evolution of an unpopular norm.

This task will be undertaken in Section 3.4.3, but the candidate hypothesis that will be tested there could be stated the following way: Given an initial population not representative of the underlying preference distribution and a network evolution that facilitates the

influence of these initial nodes providing misleading cues, an unpopular norm emerges.[13] Besides testing this hypothesis, the sensitivity analysis has to explore the impact of variation in the parameters m and m_0, which control the network growth algorithm.

A side remark on the stochasticity of the model and potential correlations in the target system: Random elements of the model are not necessarily related to actual random, unknown or unpredictable factors. For example, there might be patterns relating political radicalism and influence, positive or negative. The model's randomness is meant to allow a variety of different social settings, generating quite different patterns and exploring their consequences.

3.4.3 Sensitivity Analysis

The Importance of Network Structure

A first important question is how relevant the network structure is for the model's behavior. The preferential attachment mechanism creates networks with a relatively small number of highly connected nodes, so-called hubs, while most agents have a degree close to the minimum set by m. While the model configuration only provides an approximation of the limit degree distribution for relatively small k, it already assures the general pattern of a small number of hubs and a large number of peripheral nodes.

Up to this point, it has been tacitly assumed that the informational constraints posed by this network structure are vital to the model's behavior. The assumption can be operationalized for testing by the following hypothesis:

$H_{Network}$ The perceived norm in a given network is positively correlated with the preferences of a small number of hub nodes.

The rationale behind this hypothesis is that most agents make their judgment on the basis of perceiving mostly hub nodes. They do not only perceive hub nodes – at least they perceive themselves – but hub nodes' preferences are overrepresented. Figure 3.4 depicts the relationship between the average population norm and the preferences of the top 5% of agents with regard to their degree in the network.

A simple linear regression for the influence of the hubs' preferences gives us approximately a slope 0.8 with an r-squared of 0.52 and standard error statistics of 0.05, with $p \approx 5.5e^{-33}$. Hence, the average preference of the network's hubs provides a good predictor of the resulting population-level norm. To better frame this result, a second regression relating the average population preference to the average norm can be fitted.

The plot exhibits a large amount of variance, with the following regression statistics: The slope of the regression line is 2.1, with a value of 0.08 for r-squared and a standard

[13] Note that the model has a built-in self-correction mechanism, since for finite m and m_0, new agents entering the system will always assign some weight to themselves and therefore new information on the true underlying distribution of preferences keeps entering the process, leading to a correction in the long run, notwithstanding the potential for a strongly mistaken norm at intermediate timesteps. The model itself is always evaluated with respect to short to medium term results, since those are of more significance to the intended target system of groups forming social norms.

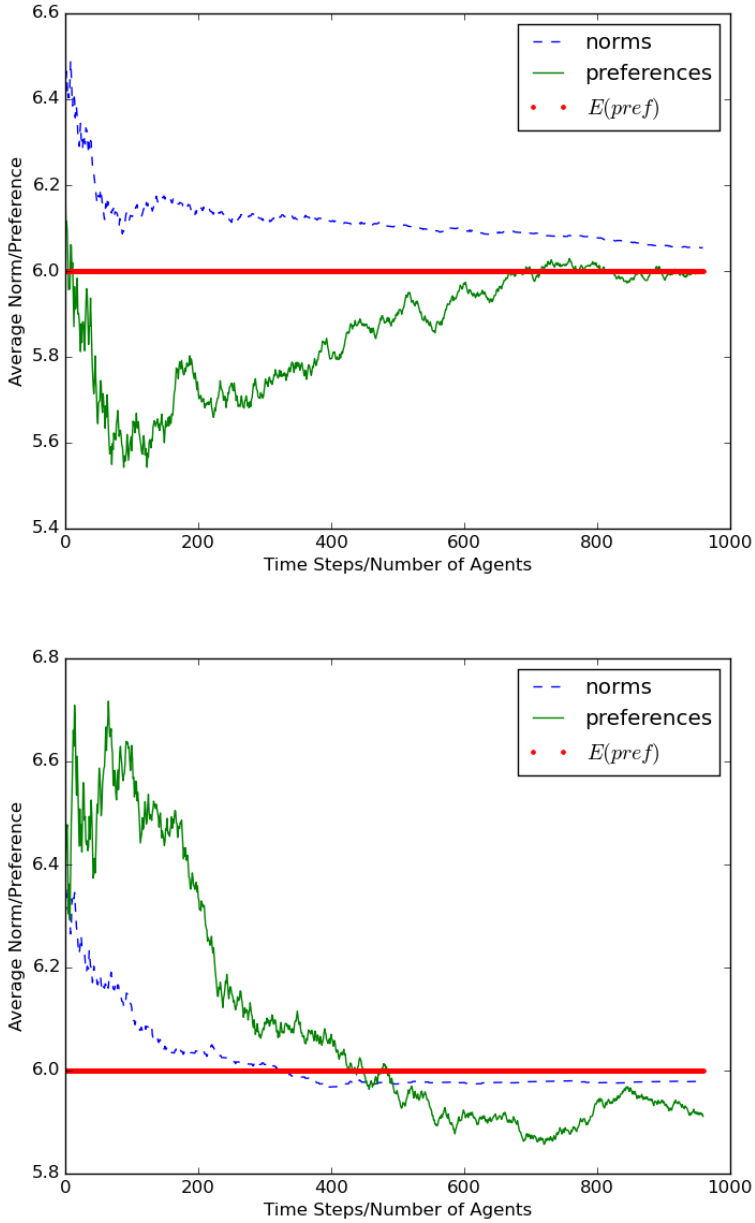

Figure 3.3: Two cases of approximately successful convergence to the optimal norm defined as the mean of the underlying preference distribution.

error of 0.52. Comparing these two statistical models confirms the assumed importance of central nodes for the resulting norm in a given simulation run. The actual preference distribution's influence is strongly perturbed, making it effectively useless to explain the variation.

To summarize, well-positioned nodes are crucial to the evolution of norms in a world of

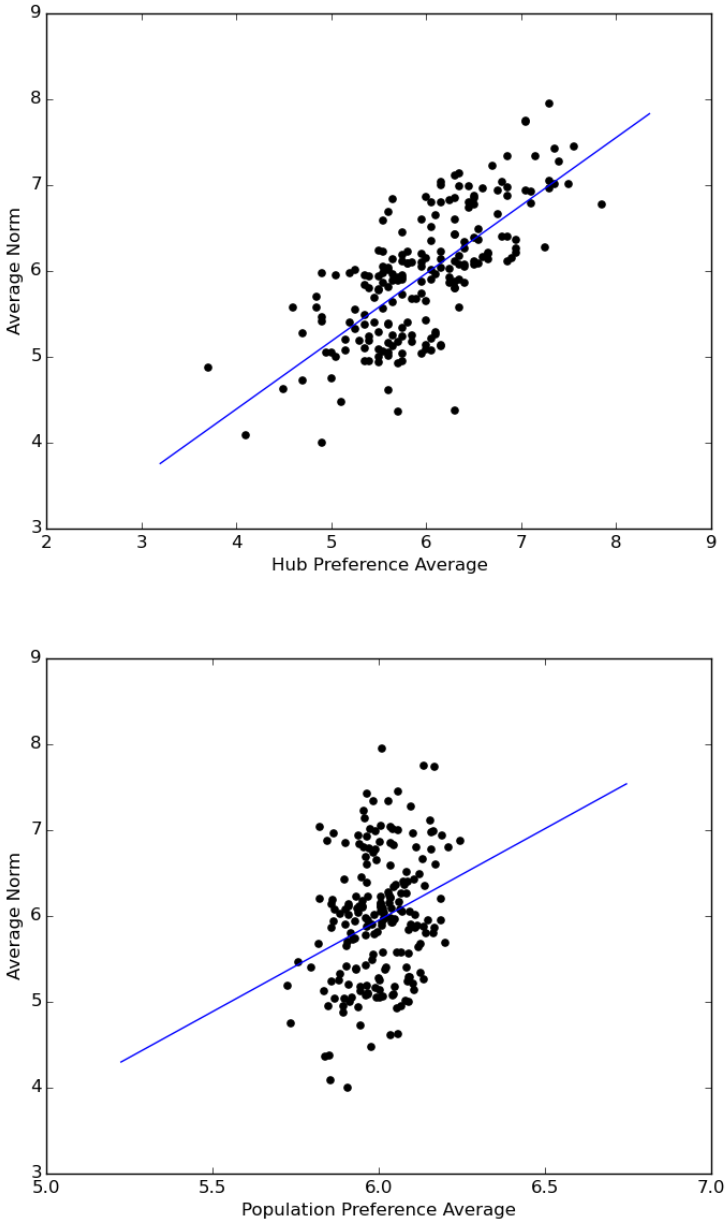

Figure 3.4: Average hub (top 5% degree nodes) preference vs. average population norm and average population preference vs. average population norm, together with the regression lines.

restricted information and a need for adapting behavior to perceived norm expectations. With respect to real-world phenomena, network hubs can be interpreted differently to account for different application domains. For example, in the analysis of the political climate in Israel, Shamir & Shamir point to the important role of the incumbent government

with its superior ability to communicate a certain opinion as the norm.

Switching contexts, while Prentice and Miller (1993) do not give this explanation explicitly, they mention the influence of fraternities on the perception of students. Again, an established institution allows its leading members to overrepresent a certain behavioral norm to the overall population. In the case of fraternities on a campus, they might also have additional influence on the perception of students in the general public, creating an additional indirect effect on prospective students' expectation and perception of campus norms. The discussion shall return to this issue when exploring the effect of explicitly modeled central influence.

However, even without the assumption of institutionalized influence, the case of a college campus can be framed in the model's terms to understand the application of preferential attachment in a practical example: When a new batch of freshmen enter the school, they have approximately no social ties. When entering the social network at the campus, it makes sense for them to orient towards the most visible senior students, since those already know the social rules in place, which the freshmen have an interest in learning.

In the real world, many links are probably formed simultaneously, and the probability to connect to fellow freshmen is higher than assumed in the preferential attachment model. Nevertheless, a process of continuous addition of nodes to the social network with the probability of new links skewed towards the incumbent population is a reasonable model of our stylized description.

Parameter Space Exploration

What remains to be analyzed is the model's behavior across a fine-grained variation of the parameters. Especially the potential interplay of m and m_0 has to be taken into account. These parameters control network growth, and therefore the available information. Larger m improves the access of new agents to the existing network, while m_0 controls the likelihood that the initial population is representative. The larger the initial sample, the more likely it is representative of the generating distribution.

Figure 3.5 depicts an exploration of a larger part of the parameter space. The parameters m and m_0 are varied between 1 and 2 respectively to 50. The upper graphic shows the maximum difference between the mean norm and the mean preference over 5 runs, the lower one depicts the average of that difference. This difference provides a rough measure of the failure to identify the optimal norm in a given simulation run.

Descriptively, there are two main features shown in the graphics: The lowest values of pluralistic ignorance, signified by the lighter blue, are concentrated in the domain of low m and large m_0. There is a strong increase (signified by orange and red) towards smaller m_0, and a smaller, but clear increase when both m and m_0 grow larger. It is also relevant to observe that cases of stronger pluralistic ignorance are scattered within areas of weaker ignorance, pointing to the volatility of the underlying process.

To interpret these observations, I shall start by noting that pluralistic ignorance occurs consistently over a wide variety of parameter configurations. While being mostly a negative result, it is important to show the robustness of the process to justify the use of single run

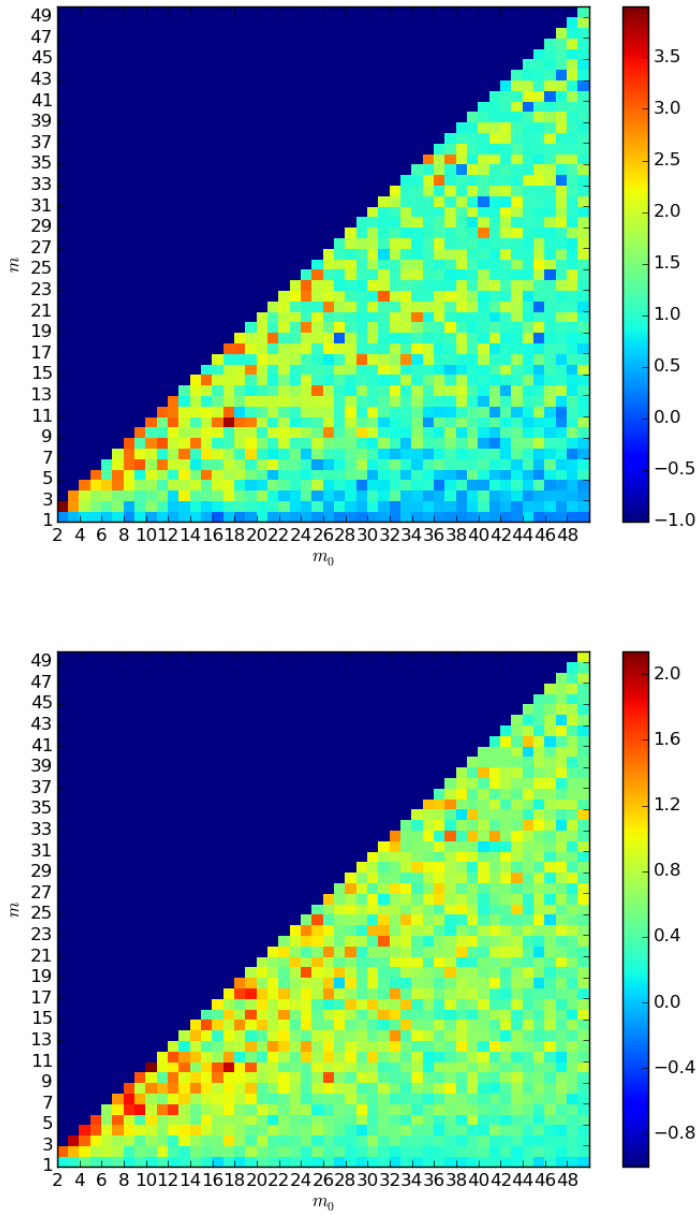

Figure 3.5: Parameter space exploration of m and m_0. The top figure shows the maximum difference between the mean norm and mean preference at the end of the run over 5 simulations, the bottom figure shows the corresponding averages. The dark blue squares represent impossible parameter configurations.

analysis. But still, the extent of ignorance and its variability differ across the parameter space.

Configurations with very small m are less prone to pluralistic ignorance, but the most

extreme scenarios seem to appear more often for relatively small values of both parameters. It appears prima facie counterintuitive that smaller m improve the performance when m_0 is large, since larger values of m imply more information for the agent. But in the case of a small initial population that is misrepresentative of the actual preference distribution, a large degree parameter just reduces the weight of the preferences of new agents compared to the incumbent population, thereby cementing the unrepresentative distribution of the first m_0 agents.

This pattern can be explained more directly from the decision rule, since larger m automatically reduce the weight agents assign to themselves and increase the relative importance of the current agent population. There are three important takeaways from the sensitivity analysis for the general validity of the model:

1. Pluralistic ignorance occurs under a variety of different structural conditions imposed on the network. This justifies the generalization from the behavior under a specific parameter configuration to the properties of the simulation model more generally.

2. It shows that the skewed degree distribution maintains pluralistic ignorance when connectivity is increased. This complements the above experiment on the relevance of hubs: Since the number of hubs remains approximately stable due to the scale-free character of the degree distribution, the influence of hubs is sustained even in scenarios with a larger total number of sources of information.

3. The whole process is overall fairly volatile relative to the network structure. This also implies that to potentially evaluate the risk of the emergence of an unpopular norm, a researcher would have to know not only the network structure and the distribution of preferences, but also how preferences map onto the structurally important nodes in the network.

3.4.4 Central Information

The influence of centralized agencies like government-controlled or otherwise centralized media seems to have an important impact on the emergence and sustenance of a state of pluralistic ignorance and, in turn, unpopular norms. Other examples where a central agent or agency is able to communicate a behavioral norm to a whole population are centralized churches and media empires holding a quasi-monopoly on a certain widely consumed medium.

In our basic model, this kind of influence does not exist. It is decentralized, and the initial nodes, which are most likely to become important hubs in the network, are independent in their preferences. This is once again similar to many models of opinion formation (as the basic version of the Hegselmann-Krause-model, see Hegselmann and Krause (2002)), but in either case an unrealistic assumption. In particular, there exist extensions to the Hegselmann-Krause-model of opinion formation to include various central sources of information (Hegselmann and Krause, 2006, 2015).

Therefore it is an important step to amend the model to represent central influences, which can be achieved by adding a single node to the network which is visible to everyone from the beginning on. Figure 3.6 shows the outcomes of simulation runs, again with $m = 10$ and $m_0 = 40$, one with a central influence $I = 6.0$, which is identical to the mean of the distribution from which agents' preferences are drawn, and a second one with $I = 8.0$, representing a perturbing central influence.

These outcomes can also be compared to the case without any central influence shown in figure 3.4. The resulting norms are significantly more clustered around the mean of the preference distribution in the case of representative central information, as one would expect. In the case of a misleading central agent, again unsurprisingly, the perceived norm shifts towards this agent's behavior. In addition, choices are clustered around integer values due to the agents' adherence to the median perceived norm that is necessarily discrete on an integer scale.

It is interesting to note that these effects are pronounced despite the fact that the central information is only 1 in 11 neighbors in the given configuration. This explains the variation in the outcomes, but it also confirms the immense influence of a central agent in a scale-free network.

An interesting feature of this model of centralized sources of information is that it also generates an indirect effect. When a new agent is added to the network and gathers information about the current norm, it is directly influenced by the central media node. But there is also an indirect effect, since all the neighbors it retrieves information from have been influenced by the central agency before. Therefore, the cumulative effect of such an agency goes beyond giving information, it also creates a certain informational background across the agent population it influences. In this manner, centralized agencies can be highly influential in creating a certain political, social or cultural climate, without even being perceived as a more important source of information than normal agents close in the social network.

3.4.5 A Normative Perspective

Up to this point, the normative assumptions making an unpopular norm collectively irrational were largely implicit, since the analysis focused on the genesis of norms that happen to be unpopular. But evaluative statements are clearly pervasive in the analysis of unpopular norms. By definition they are norms that are not in the interest of the majority of their subjects – a point that is also mostly implicitly present in the empirical literature. The selection of examples includes norms of corruption, alcohol abuse and racial segregation, which are often not only detrimental to their subjects but also considered social problems externally. In some cases, they might even be considered ethical problems. The model and the arguments based on its behavior are, however, open for the application of various normative standards.

First of all, it is necessary to clarify what the agents' actual underlying preferences are, and what is part of their strategy. Agents prefer to behave according to their built-in ordering of options: an agent who favors option 5 on the model's scale would prefer 5 to

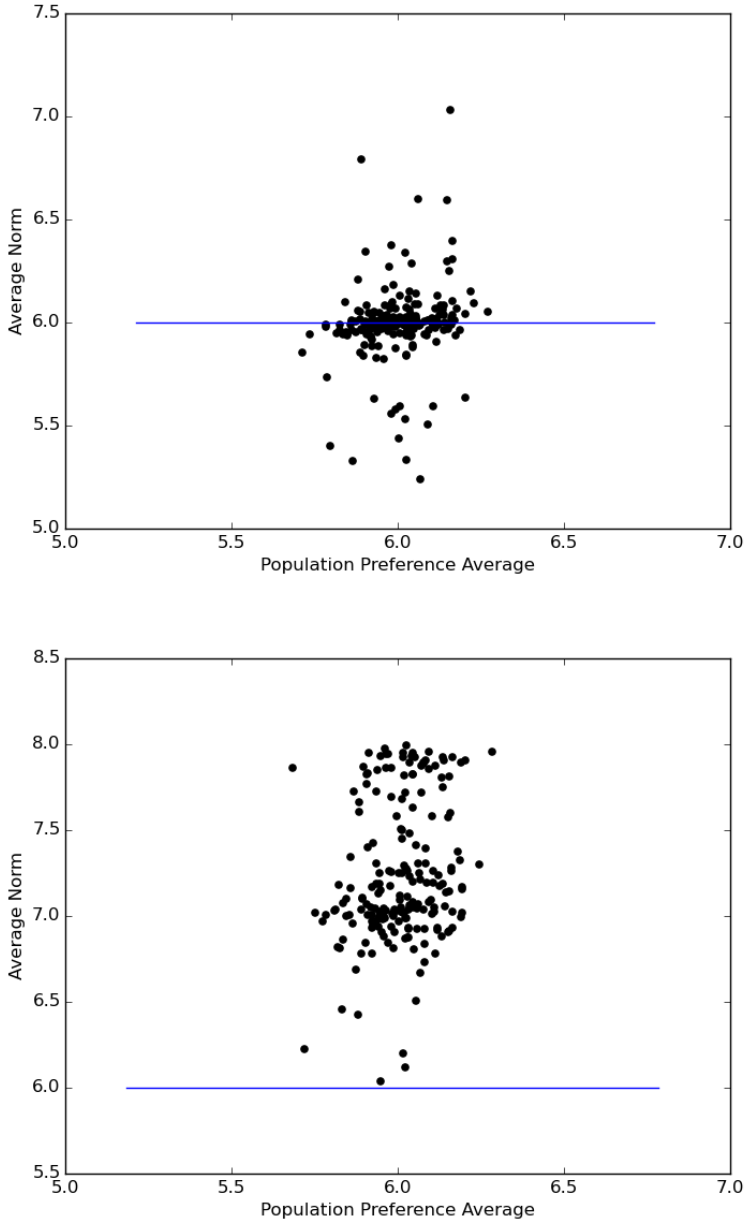

Figure 3.6: Results of 100 simulation runs with either representative influence (top) or misrepresenting influence (bottom). The straight line shows the mean of the theoretical preference distribution.

everything, 4 and 6 to anything but 5 and be indifferent between 4 and 6, and so on.

In addition, a social component has been assumed: a good outcome, under this assumption, will always engender a high degree of coordination. This factor can be understood as part of the agentic preference structure, too.

For example, a member of an emerging political movement might have a very specific preference on the subject matter, but he or she also has an interest to be part of a democratic majority to realize something as close as possible to his or her interest. Equation 3.2 can actually be reinterpreted as a definition of an agent's payoff, since it takes into account the degree of coordination as well as individual preference. The weights can express the different importance of coordination with different people – for example, coordinating with your co-workers or family seems very valuable, while being aligned with all the distant members of your political community seems less important. For the current discussion, the assumption is $\forall i \forall j w_i(j) = w_j(i)$.

The last implicit assumptions concerns the features of the utility functions resulting from the above preferences. First, it is assumed that the distance from coordination figures proportionately. This assumption can easily be relaxed to construct alternative normative standards. Quadratic distance measures are a natural candidate, since they put more weight on extreme deviations. In a population with a uniform preference distribution, this change would not be particularly relevant. But if there is a substantial minority on one extreme of the spectrum without a counterpart on the other end, quadratic measures would favor different points of coordinations.

Second and less innocuous is the implicit assumptions that one can simply weigh all agents equally in judging the norm. In Section 3.1, the argument relies on simply counting the agents who landed on a certain choice. But if the collective rationality of the social norm is to be judged purely on its maximization of its subjects' utilities, intercomparability is notoriously problematic.[14]

But normative standards not requiring intercomparisons are also often unsatisfactory. Pareto-efficiency provides a standard example. An outcome is considered Pareto-efficient if at least one agent strictly prefers it to its alternatives. Given the assumptions on the preference structure, that would make any spot on the spectrum of choices Pareto-efficient if it is occupied by at least one agent. Under this standard, actually collectively irrational social norms were practically impossible.

Also, social norms could be considered collectively irrational not directly because of the preferences of its immediate subjects, because their preference structure might itself be an unfortunate consequence of their irrational social norms. Such group-level adaptive preferences [15] may be considered irrational by the surrounding larger community or an impartial observer, but that judgment could not be passed purely on the actual preferences of the agents in question.

To summarize, there are strong, but not universally compelling reasons to apply a standard similar to the one assumed in my discussion on unpopular norms. Constructing the standard from preferences, assuming intercomparability and valuing coordination are valid across a significant number of applications – I believe that they are reasonable for all the examples discussed in this chapter – but I concede that there are likely extreme cases of unpopular norms, such as foot binding in the traditional Chinese culture (Willer et al.,

[14]For a discussion, see Sen (1986).

[15]See Nussbaum (1997) for an introduction to adaptive preferences.

2009), where different normative standards need to be considered.

3.5 Challenges and Model Limitations

While the suggested model meets all the benchmarks set in the introductory section and thereby provides a valid theoretical model of the evolution of unpopular norms under pluralistic ignorance, there are some important challenges and limitations to the model that ought to be mentioned. They fall in three broad classes:

1. Idealizations that lay bare the dynamics of a single social process, isolated from the various influences interfering with it in the real world
2. Simplifying assumptions that enable the computational analysis of the model
3. Domain restrictions

Important instances of the first kind of idealizations are the lack of belief revision, the exclusion of external events, and the backward-looking, i.e. non-strategic behavior of the agents. It is possible to relax all these assumptions – in fact, it would be highly interesting – but adding these factors would create a new model of multiply intertwined processes. Such extensions could easily be added to the model, as the exploration of central sources of information has shown, but they go beyond what the current investigation set out to achieve. In particular, if one wants to extend the model to arbitrary time scales, it becomes crucial to include factors like belief revision and external events. Insofar, these idealizations can also be understood as domain restrictions.

Besides explaining the phenomenon by the most simple fundamental process sufficient, there are additional reasons to adopt these constraints. The effects of belief revision, for example, seem to depend heavily on the particulars of sequencing. How often and when belief revision would take place would be crucial to its effects on the process. Note that this kind of idealization is also responsible for the impossibility of quantitative fitting of data. To account for real-world data, it is usually necessary to include not only a multiplicity of processes, but also to include a number of theoretically meaningless parameters for calibration. The model is not supposed to be used in such a research agenda; its main purpose is to provide a formalization of social theory to allow for the exploration of its implications.

The second kind of limiting modeling assumption, on the other hand, could be relaxed not by fundamentally changing the model, but by gradually varying certain factors. In the decision procedure, the main assumptions concern the choice set and the weighing of other agents. With respect to the choice set, it has been assumed that it does not make a relevant difference to have an 11-point scale to having a 7-point scale. While this seems plausible, it has not yet been explicitly shown. Varying the weights of various agents is a more intricate problem, since there are infinitely many options to model differential weighing. Agents could weigh themselves disproportionately, they could emphasize the importance of well-connected nodes in their neighborhood, or intentionally de-emphasize them.

For the network structure, a very basic version of preferential attachment has been chosen. But there are many alternatives to growing an artificial social network. As the analysis of hub influence has shown, the model's behavior depends on the stratification of the network into hubs and peripheral nodes, but there are alternative algorithms to generate such networks that have not yet been tested.

The results of the analysis of preferential attachment allow speculations that can be tested by future research. The most common alternatives would be random graphs and small-world networks, as mentioned above. The prediction for a Erdös and Rényi (1960)-network is straightforward: The binomial degree distribution implies that hubs with high degree are less likely than in the scale-free preferential attachment network. Therefore, pluralistic ignorance becomes less likely.

The case of a small-world network is more complicated. Informally, if the network evolution starts from what will become the bridging agents, these agents would be more influential then a random agent from within a cluster. The effect should be smaller than in the case of preferential attachment, since most of the additional influence of the bridging agents is indirect, and therefore fairly weak under the assumption of equal weights. Given no alternative assumption on weights, pluralistic ignorance seems once again improbable. The degree distribution of a small-world network falls between that of a ring lattice and the aforementioned random graph depending on the parametrization, and both assign low probabilities to large degrees – as compared to the scale-free distribution generated by preferential attachment. A more interesting perspective on a small-world based model might be to compare the within-cluster distributions of behavioral expectation with the global preference distribution.

Finally, the restricted domain is not so much a limitation, but rather a necessity of scientific modeling. The model imposes two major constraints on the class of potential target systems: First, the model has to behave approximately as described by preferential attachment. To even judge this, it has to be assumed that the network grows sufficiently large: Small kinship groups, for example, are certainly out of scope. The empirical literature on preferential attachment has found fitting power law distributions in scientific collaboration and citation networks as well as the collaboration network of movie actors (Jeong et al., 2003). Furthermore, the World Wide Web (Donato et al., 2004) and the link structure of Twitter (Java et al., 2007) have been found to follow power laws in their degree distribution.

There possibly is some selection bias in the tested examples, since these are all networks on which data is comparatively easy to obtain. Keeping this in mind as a caveat, one can tentatively generalize: where agents are able to choose links without strict spatial or formal limitations and linking oneself to well-connected nodes appears advantageous in some sense, the network structure arising from preferential attachment seems to be an empirically adequate model. Note that a network evolution following the rules of a formal hierarchy will often generate a highly skewed degree distribution, too, if it is assumed that links represent the perception of behavior and are usually not bidirectional. Thus, there might be a broader class of applications, but for the discussion in this essay, the applicability of preferential attachment limits the model domain.

Second, as mentioned before, the target system's time scale shouldn't render it absurd to exclude external events or belief revision. For example, if there is a major election, a media campaign, or intense public discourse, the model may only be applied to the process before that event. In general, the model fits best when a relatively decentralized process of opinion formation about behavioral expectations takes place over time, with more and more agents entering the system.[16] Norm formation within some growing grass-roots political movements or the workplace population in an expanding company can provide examples for processes operating in accordance with these constraints. Note that application is meant in a qualitative sense, not a quantitative analysis or prediction.

3.6 Conclusions

3.6.1 Summary

It has been shown that the evolution of an unpopular norm via pluralistic ignorance can be modeled by the combination of agents deciding under social influence given the informational restrictions of a social network. Relative pluralistic ignorance, denoting a deviation of the overall perceived distribution of norm expectations from the population's actual preferences, emerges in the model under various parameter configurations. Furthermore, the model turns out to be more general, as it also allows for the emergence of efficient norms in a population of conformist, boundedly rational agents.

The driving force is the influence of hubs in the network, which can be either misrepresenting the actual distribution, representing it correctly for the subpopulation at a given point in time, or even be representative of the underlying distribution of preferences for the whole population when the current subpopulation is still way off the unfolding preference distribution. These phenomena lend themselves to interpretations as hubs being a reactionary or avantgardist elite – with elite defined in purely relational terms. The statistical analysis of the relation between hub preferences and the resulting population-wide norm supports this interpretation.

As further exploration has shown, pluralistic ignorance can arise in networks of varying structural parameters. It occurs most pronounced in configurations where information is particularly sparse and non-representative. More representative information, however, can be provided by central agencies in a minimally intrusive way. At the same time, such a powerful central influence can also perturb the evolution of a norm in a detrimental way, according to the standard provided by the population-wide preference distribution. This analysis connects the decentralized model to those case studies that feature central influence at crucial points. Especially in political contexts, central agencies are pervasive. But also for small, naturally decentralized groups, establishing a central information instance can

[16] Agents entering the network can in some cases be taken literally, as in the example of a school or university. But in many cases, it simply represents the sequential choice of agents who are already connected; then, the creation of links represents not links in the social networks but a relation of having observed someone else's choice.

be utilized as an intervention.

These leads back to the normative question. The overall process of norm formation described by the model enables coordination on a social norm beneficial under a variety of assumed goal structures. The main shortcomings of the process as it stands is the lack of a secondary process monitoring the quality of the outcomes and triggering interventions when necessary, and moderately benevolent conditions of communication. However, the process fails as a collectively rational procedure of norm formation when information is too sparse or the impact of a subset of agents is overwhelming. This relates the collective rationality of the process to the distribution of environment configurations to be assumed.

3.6.2 Future Directions

There are some natural extensions to the model that are to be explored in future research. Mechanisms of belief revision, the inclusion of external events or a more extensive study of the robustness of the results under varying network topologies are all valuable options to increase the theoretical understanding of unpopular norms and social influence mechanisms more generally. Another interesting route to consider is a hybrid approach that combines the mechanism of social influence with a more strategic behavioral model. This kind of hybridization is also one of the current problems for formal models of opinion formation (Mäs, 2015), which are closely related to the dynamics of norm expectation and behavior.[17] This issue can be understood as an instance of the more general problem to combine strategic, forward-looking behavior with backward-looking adaptive behavior in agent-based modeling and simulation.

Besides improving the descriptive accuracy and the explanatory power of the model, future research should explore normative consequences of the model under a wider variety of possible evaluative standards. There is an apparent tension between the abstract character of stylized agent-based models and the need to devise concrete, quantifiable interventions to ensure collective rationality.

But in that respect, model-based arguments are very similar to ones based on informal theoretical accounts. In their discussion of availability cascades, a related phenomenon of apparent collective irrationality, Kuran and Sunstein (1999) make no actual quantitative claim. Based on an empirically supported theory of individual behavior, they develop a theoretical account and suggest concrete interventions – without providing any quantification of the effects.

Both approaches fit into the general argument scheme for collective rationality. However, formal models provide a more transparent way to express theory and draw conclusions. As long as we are interested in counterfactual scenarios to solve social problems as they are represented by unpopular norms, agent-based computer simulations provide an important tool not only for explanation and description, but also to develop interventions and support normative arguments.

[17] Chapter 4 suggests an approach to the problem in the context of belief aggregation.

Chapter 4

Strategic Belief Formation[1]

4.1 Introduction

Whereas the previous chapter focused on a generic mechanism to coordinate collective behavior – social norms – I shall now turn to a different range of problems: in a group pursuing epistemic ends, potentially besides a variety of other motivations, how can information be effectively aggregated by a decentralized process implemented by boundedly rational agents? This description already gives a sketch of the decision environment for the arguments to come. Agents have limited access to information from the world, and are therefore confronted with the problem of incorporating information provided by their fellow agents, but without immediate access to *their* full information about the world.

As a consequence, agents can engage in the *strategic* presentation and transmission of their state of belief. Even if agents are not themselves motivated to influence others, they might be forced to take precautions against strategic manipulation. These are, in a rough sketch, the outlines of the social process under scrutiny.

With respect to evaluative standards, I argue that while accuracy is of crucial importance for paradigmatic epistemic communities such as science and truly valued by scientists, other ends may figure in an evaluation of collective rationality. Even more importantly, accuracy as an individual-level epistemic norm fails to translate to a unique norm of collective belief evaluation. To exemplify the problems involved and ensure the relevance of the discussion, science in its current form provides the domain of real-world targets for my models and the arguments constructed about them.

Science is a social epistemic enterprise. The increasing scope and complexity of both scientific theories and methodologies requires division of labor (Kitcher, 1990) both vertically, spanning the range from theorists to experimentalists and the engineers constructing the necessary machinery, and horizontally, by tackling the same problem with a variety of approaches. This cognitive division of labor creates an ever increasing need to incorporate socially available results from other researchers.

Agents finding themselves in this condition face the question of how to aggregate avail-

[1]The content of this chapter is in part published in Merdes (2018)

able information and how to transmit their results, given that everyone in the scientific community of their field finds themselves in the same predicament. As suggested by this description, this epistemic condition creates, among a variety of other problems, the possibility of strategic behavior (Strevens, 2003). More precisely, it can incentivize a strategic presentation of one's own beliefs about significant scientific hypothesis, given a limited amount of directly accessible data, the potential for others to incorporate one's expressed beliefs, and a diverse set of motivations.

Consider the following stylized, but not implausible scenario: Imagine a collection of scientists, all working on the question whether a particular psychiatric treatment is significantly superior to the null treatment. Furthermore, assume that some of the agents involved are invested in psychodynamical approaches to therapy, while others are otherwise proponents of behavioristic treatments. Some agents are solely motivated by their desire to uncover the truth, others would like to further their career by appearing – and maybe being – particularly original. Again others deem all of their colleagues highly reliable, while others have their reservations.

All being psychologists, they also not only know that there are both rational grounds and a psychological tendency for human agents to be socially influenced in their beliefs, but they are also aware that such agents have the capacity to engage in strategic behavior: they will – within their capabilities – anticipate the reactions of their fellow epistemic agents and act to further their ends in the light of this behavioral insight.

One possible consequence is the exaggeration of the relevance of one's observational results by employing methods such as p-hacking. Not all ways of emphasizing one feature of one's results over another are, at least prima facie, epistemically vicious in the same way. But they all hold the potential to mislead other researchers, and give rational grounds to discount socially available information.

A few examples from scientific practice highlight the relevance of the topic, as well as the difficulty to unequivocally assess potentially strategic behavior. Publication bias, i.e. the tendency to only publish statistically significant results (Auspurg et al., 2014), provides a useful working example; here, the editor, in anticipation of the reception by the scientific community, creates an incentive structure that motivates scientists to selectively publish their results or even compromise their data analysis.[2]

A less benign example are the case studies of industry-funded biased pseudo-research analyzed by Oreskes and Conway (2011). In their case studies, a subset of scientists are motivated in their research entirely independent of the epistemic values internal to the scientific community. Such cases emphasize the need for robust strategies to incorporate socially available information, since it is in practice difficult to simply identify biased agents and to dismiss them entirely.[3]

The goal of this chapter is to construct a model of the problem of decentralized aggregation with varying motivations across agents, and to employ that model to support

[2]Gigerenzer (2015) argue that a subset of research on human rationality provides another example of, at least allegedly, selective reception and reporting of results.

[3]For a formal analysis of the impact of biased agents in epistemic networks, cf. Holman and Bruner (2015).

the following consequence of the social epistemic agent's predicament: there is a trade-off between behavioral strategies which lead to epistemically optimal social outcomes under benevolent circumstances and strategies more robust to variation in the group's motivational structure and communication patterns. Furthermore, an ecological analysis using techniques from evolutionary game theory suggests that under a wide variety of circumstances, accuracy-rewarding institutions can diminish strategic exaggeration. However, mixed reward schemes, which are often inevitable in science, leave substantial space for manipulative communication within the epistemic community.[4]

The chapter proceeds in the following way: In Section 4.2, an agent-based model for motivated exchange of beliefs is constructed. Section 4.3 explores the dynamics of the model and provides support for the trade-off between accuracy and robustness. The analysis is extended to an evolutionary approach in Section 4.3.3, and the case study of science is complemented with a brief sketch of an application to legal epistemology in Section 4.4. Section 4.5 discusses some of the important challenges to the suggested model. Section 4.6 provides a summary of results and an outlook for future extensions and applications of the model.

4.2 Deliberation Game

4.2.1 Iterated Opinion Expression

Science constitutes the paradigm of collective epistemic enterprises. From a comprehensive point of view, scientific research is not a uniform process, but consists of a variety of activities, such as experimentation, literature review, direct discussion as on a conference or workshop, indirect disputes via publications and so on. Any complete model of the scientific enterprise therefore would need to be extraordinarily complicated. Instead, social epistemologists focus their modeling efforts on specific features pertinent to one particular mechanism or process in the target system. In the following inquiry, the process under scrutiny is the potentially strategic exchange of expert opinion on a specific hypothesis of interest.

The process of collective opinion formation is reduced to two subprocesses: data collection and repeated opinion expression. Initially, every agent receives a noisy signal b_i, drawn from a beta distribution centered on the true value μ.[5] Depending on the discipline at hand, an agent can be interpreted as either a single person, a small research team, or an entire lab. The signal represents either the probability that some hypothesis H is true, or the true value of a variable of interest; either interpretation is consistent.

After the data collection phase, agents are allowed to express their belief. There are multiple ways to model the initial expression $s_i(0)$. One could allow the agents to reveal

[4] For alternative, but related models of strategic behavior in opinion dynamics, cf. Hegselmann et al. (nd) and Eger (2016).

[5] Any other non-degenerate probability distribution will serve the same purpose; the beta distribution has simply been chosen for technical reasons.

their signal or at least do so with some additional noise due to the imperfections of human communication. Instead, some of the agents might be perceived as starting off from an entirely random position, based on previous imprecise predictions or as an expression of external motives . Complete randomness is of course an idealization, but it constitutes a useful limiting case.

After the first step of aggregation, agents are allowed to engage in repeated updating of their expressed opinion $s_i(t)$ to incorporate information made available in the previous time step. The iteration allows the agents, among other things, to acquire higher-order evidence. Cognitively limited agents will not always epistemically profit from the iterative structure, but it is an indispensable feature of scientific debates.

The strategy space in the model ranges over sequences of s_i, and the agents have access to their own signal b_i, and some or all choices of s_j made before. What is missing to complete this simple deliberation game are appropriate payoff functions.[6]

4.2.2 Payoff Schemes

The scientific community, like most social groups, is not subject to only one scheme of rewards, even if it was only for the individual differences between scientists and their motivations. This is partly due to the influx of non-epistemic values. But it is also a consequence of the variety of epistemic values (Kuhn, 1977) and the difficulty to always identify the epistemically valuable directly. Looking purely at accuracy, it is often possible only in hindsight to attribute the comparative value of predictions. In particular when it comes to the prediction of rare events, it can be difficult to assign credit before the fact.[7]

To accommodate such problems – and also because scientists are psychologically functioning human agents – science relies on social judgment to a significant degree.[8] At least three relevant components for determining a scientist's reward can therefore be identified: accuracy, coordination with their epistemic peers, and the impact of their results on the scientific community.[9]

The deliberation game model suggests a very simple but common approach to modeling

[6]One might notice that the model is structurally very similar to linear averaging models, and wonder whether the opinion of an agent is represented by b_i or s_i, or put otherwise if there is an actual dynamic of evolving opinions, or just an adaption of the beliefs expressed to other agents. A more appropriate interpretation for an application to science is that of an actual evolution of opinions, the possible importance of their initial signal to an agent notwithstanding. There are different target systems the model is applicable to, such as political discourse, where often an evolving compromise is better understood as just a convergence on expressed opinion, with no underlying revision of private belief.

[7]For clarification, the need to assign credit long before the predicted event is supposed to take place arises frequently in fields like climate science or high energy physics, where predictions extend far into the future or theoretical progress precedes technical realizability by decades.

[8]This is the crucial difference between models explicitly incorporating agent motivation and standard models in opinion dynamics such as Lehrer and Wagner (1981). In applications of such deliberation processes, it is acknowledged that strategic behavior poses a problem (Regan et al., 2006).

[9]Accuracy stands in for a number of plausible epistemic values discussed e.g. by Kuhn (1977). The reason is both ease of quantification and the almost universal acceptance of this particular virtue.

accuracy, namely by an agent's distance from the truth.[10] It has been argued that proper scoring rules, such as the squared distance measure, are an appropriate way to evaluate accuracy (Leitgeb and Pettigrew, 2010a,b; Brier, 1950).

This characterization of accuracy suggests a reward scheme specified by the following payoff equation:

$$u_i(s_i) = -(s_i - \mu)^2 \tag{4.1}$$

where i is an agent, s_i their expressed opinion, and μ the correct value of the variable of interest or degree of belief warranted by the data.

Under this payoff function, agents have an incentive to reveal their signal on the first step of the deliberation game, and from thereon always choose the average of the revealed signals, under the joint assumptions that everyone shares the same reward scheme and common knowledge of rationality.[11] There is no new information entering, there is no incentive for an agent not to reveal, and signal variance is assumed to be homogeneous, making the linear average the maximum likelihood estimator. In particular, it does not matter how many iterations are run, since there is no dynamic after the first two steps.

But the value of μ is generally unknown – that is why a scientific investigation is undertaken in the first place. Therefore, the community cannot simply implement Equation 4.1 as their reward scheme. To alleviate such accessibility problems, the second reward scheme to be examined is based on coordination with the scientific community. Note that agreement-based rewards need not be understood as irrational or arational: they can be based in rational reliance on the approximate accuracy of the epistemic community.

There are multiple ways to model a coordination-based reward scheme. Consider the following options:

$$u_i^c(s_i) = -\left(s_i - \frac{\sum_{j=1}^{n} s_j}{n}\right)^2 \tag{4.2}$$

$$u_i^{ad}(s_i) = -\frac{(s_i - b_i)^2 + \sum_{j\neq i}(s_i - s_j)^2}{n} \tag{4.3}$$

$$u_i^{da}(s_i) = -\left(s_i - \frac{b_i + \sum_{j\neq i} s_j}{n}\right)^2 \tag{4.4}$$

where again b_i is i's private signal and n is the total number of agents. The squared difference is taken to avoid discontinuous payoff functions and to maintain structural similarity as compared to the accuracy-based payoff. The first equation reproduces equal-weight linear averaging under synchronous choice, and is considered the baseline option. The second and the third incentivize the agent to either minimize the distance to the average or the average distance, though with the additional twist of inserting their own signal instead of

[10] As Douven (2010) points out, beliefs in this kind of model can either be interpreted as degrees of belief or estimates of a certain variable of interest. For the purpose of the analysis of accuracy, it is unnecessary to distinguish between those two interpretations.

[11] The strategy profile where all agents play this strategy is a Nash equilibrium. However, the strategy is not dominant: if all other agents play, for example, by uniform randomization, the agent should stick to their own signal.

their expressed opinion. The rationale is that, while agents are assumed to be rewarded for conformity with the expressed opinions of others, they retain the signal they have privileged access to; they treat this signal as constituting part of the collection of expressed beliefs to be in agreement with.

It is noteworthy that all three options imply different optimal solutions in a population of agents homogeneously subject to the respective payoff scheme. The baseline scheme is trivially maximized by perfect coordination on any value. Differentiating and synchronously maximizing the p_i^{ad} is also maximized by perfect coordination, but in this case specifically on the median of the b_i. Unlike the baseline scheme, it is not trivial for agents with limited information to coordinate on this point, since it requires them to estimate the private signals of other agents correctly; but since they do not have access to the other agents' signals directly, they can only estimate on the basis of already expressed opinion.

The optimal solution to the third scheme differs more substantially, since it deviates from perfect agreement. In equilibrium, every agent plays a mixture of their private signal and socially available information. For example, in a simple three player case with complete information and common knowledge of rationality, a signal vector of $[0.1, 0.5, 0.9]$ will result in an equilibrium strategy profile of $[0.4, 0.5, 0.6]$. Because of this maintenance of diversity in opinion, strategies inspired by this payoff scheme can be described as limited or cautious conformism.

One final relevant component is missing so far, and it concerns a scientist's interest to coordinate the community on *their* result. The immediate psychological interpretation makes such a reward scheme appear epistemically irrational. Instead of trying to incorporate all socially available information, it incentivices the scientist to (at least publicly) engage in a form of motivated reasoning to lead their fellow scientists as closely as possible to agree with them. This raises the questions why actual reward institutions carry significant resemblance to this apparently epistemically problematic scheme.

A plausible answer, if one does not want to resort to the claim that science is merely a playing field of social forces and in important respects devoid of any special epistemic motivations, can be found in the reconstruction of opinion leadership as a stand-in for the epistemically valuable. Coordinating the community on one's own signal functions as a proxy for the originality of scientists and their results.[12] Rewarding scientists for originality, as they are for example by the priority rule (Strevens, 2003), seems a legitimate, if sometimes troublesome, component of the payoff structures in epistemic communities.

Formally, there are once again multiple options of representing what can be called idiosyncratic or originality-based payoff depending on the context. Consider the following

[12] A recently discussed example of this substitution is the priority of (Sakoda, 1971) over (Schelling, 1971) described by (Hegselmann and Flache, 1998) and in more detail in (Hegselmann, 2017), where the suggested models are similar though different in subtle details, but the community allocated all their acknowledgment to Schelling and coordinated on his results.

two options:

$$u_i^{sda}(s_i) = -\left(b_i - \frac{\sum_{j=1}^{n} s_j}{n}\right)^2 \tag{4.5}$$

$$u_i^{sad}(s_i) = -\frac{\sum_{j=1}^{n}(s_j - b_i)^2}{n} \tag{4.6}$$

Even though these are in important ways different functions, they have quite similar implications with respect to an agent's behavior. The reason is that an agent can only directly influence the term $s_i - b_i$ with their strategy choice. Of course their ability to choose s_i incentivizes other agents to take i's possible behavior strategically into account, but this is merely an indirect effect. When it comes to payoffs, the first equation, which once again penalizes the distance from an average, puts more weight on extreme deviants, whereas the second function puts more weight on the median agent.

In terms of interpretation, the first payoff scheme benefits normal, non-extraordinary science. If one imagines that, as in Weisberg and Muldoon (2009)[13], not all agents operate under the same scheme, this would give the idiosyncratic agent an incentive to reduce the deviance of others from their personal standpoint.

The second scheme creates much less of an interest in the normal scientist to pull deviants towards the center. However, it is important to note that either payoff rule differentiates between a one-sided deviation and a scenario where more extreme agents exist on either side of the focal agent. That is plausible, since it both provides reason to belief in one's own accuracy, according to the argument offered for social aggregation and positions the agent well to receive acknowledgment and funding from both ends of the spectrum.

What the payoff scheme cannot describe is the agents' actual capability to influence the distribution of expressed opinions. This depends crucially on (1) the agent's assumed cognitive abilities and (2) the behavioral patterns of the population the agent inhabits.[14] From hereon, I assume that scientists are in important respects limited in their knowledge, memory and cognitive ability, which suggests the transformation of the deliberation game into an agent-based model, featuring boundedly rational agents imperfectly trying to operate under the introduced payoff schemes.[15]

4.2.3 Boundedly Rational Agents

Given the basic structure described by the deliberation game and the possible reward schemes, there is a multitude of options to model the behavior of scientists. The objectives

[13]Note that this model itself has been under severe criticism. cf. Thoma (2015) and Alexander et al. (2015). However, the basic structure of the model and the problems it addresses still provide a valid point of reference.

[14]As before, other relevant components have to be set aside, in particular the impact of material resources on the actual outcomes.

[15]These assumptions block the game theoretic solutions partially discussed before, since they conflict with perfect information and common knowledge of rationality.

of this particular inquiry are to create artificial scientists who follow simple rules, but whose behavior can still be interpreted as strategic. To achieve this, the model follows Eger (2016) in using naive best-response learning; an agent employing this learning mechanism acts at $t+1$ by choosing an option that would have been optimal at t. Naive best-response learning itself is thus a backward-looking pattern of behavior, which requires minimal amounts of memory and knowledge.

The strategic component enters on the level of interpretation. Imagine a naive learner behaving as dictated by the following equation:

$$s_i(t+1) = \frac{\sum_{j=1}^{n} s_j(t)}{n} \qquad \text{(Social Learner)}$$

which is obtained by applying naive best-response learning to the baseline coordination rewarding payoff scheme. Intuitively, this rule defines an agent who takes expressed opinions, e.g. papers and conference talks, very seriously and assumes that any announcement of belief carries valuable information.

Now consider another agent who anticipates this behavioral pattern and believes that the actual payoff scheme in place is a combination of social coordination and idiosyncratic reward; such an agent could be modeled by the following rule:

$$s_i(t+1) = \begin{cases} 1 & \text{if } n * b_i - \sum_{i \neq j} > 1 \\ 0 & \text{if } n * b_i - \sum_{i \neq j} < 0 \\ n * b_i - \sum_{i \neq j} s_j & \text{otherwise} \end{cases} \qquad \text{(Social Influencer)}$$

The first two cases of the equation ensure that the agent will not express an inadmissible belief, the third case controls the behavior for the rest of the domain of belief. The resulting behavior depicts an agent who is interested in getting the current state of collective beliefs closer to their initial signal. Regardless of the exact source of this motivation, the agent can only directly influence their own strategy to move the average closer to their target; since they look for the average, that results in pushing their expressed opinion beyond their signal. Alternatively, the resulting behavioral pattern can be interpreted as strategically anticipating social influence on other agents, such as the simple learners above.[16]

Yet another agent might anticipate the idiosyncratically motivated behavior, but also believe that accuracy pays off and socially available information needs to be included to maximize it. Assuming that the agents are only capable of naive-best response learning, what would be a good compromise between robustness to unwanted influence and utilization of socially available information? Once the agents are situated in a mixed population they have little knowledge about, there is no trivial answer to this question.

The discussion of various reward schemes suggests employing a constrained version of learning from other agents:

$$s_i(t+1) = \frac{b_i + \sum_{j \neq i} s_j(t)}{n} \qquad \text{(Cautious Social Learner)}$$

[16]The factor n before the signal b_i might strike the reader as odd; it is a consequence of optimizing the payoff scheme the social influencer is based on, and effectively corrects the weights such that agents compare their signal to the average opinion within the community.

Cautious social learners give significant weight to the opinions of their fellow inquirers, but reserve additional weight for their own signal. This models, for example, an experimenter who is willing to take the claims published by others into account, though still trusts results the most they created in their own lab, measured with their own instruments and statistically analyzed themselves from the raw data to the published result.

As the number of other researchers providing results increases, their own relative weight shrinks. Hence, they are willing to correct their beliefs in the face of an overwhelming social consensus, but as long as the community engaged with a topic is small, they treat socially available information with modest suspicion without entirely ignoring it.[17]

While social learners of various kinds are the most interesting class of agents, the model is easily extensible to, for example, a steadfast agent:[18]

$$s_i(t+1) = b_i \qquad \text{(Steadfast Agent)}$$

which is behaviorally equivalent to a model suggested for charismatic leaders (Hegselmann and Krause, 2015). But in the context of the deliberation game, it is subjectively justifiable by the agent's attempt *not* to be influenced by agents misrepresenting their belief; this behavior represents the limiting case of constrained incorporation of social information. In practice, such an extreme case can only rarely occur, since the agent would just not be perceived as part of the scientific community under most circumstances. But it provides an approximately true model of, among other things

1. extraordinarily reputed researchers, who are able arrogate to themselves such behavior, at least for a limited time span,

2. agents who enter the population from the outside and therefore are at least initially presumed to be proper members of the community, but for some reason, e.g. an externally funded agenda, are not actually interested in anyone else's results,

3. agents who are so deeply embedded institutionally that they cannot be easily ignored or pushed out of the process, such as a powerful journal editor or the head of an important research institute.

These examples are all violating social epistemic norms, and while there are scenarios where being steadfast is the most reasonable response, this is not by accident: it is at odds with participating in a social epistemic process. An agent who is *systemically* steadfast in the sense of the behavioral rule stated above could just as well pursue their inquiries on their own if they searched for truth. Therefore, while the model is able to represent steadfast agents, they will be left aside within the following analysis of model dynamics.

[17] This particular strategy actually lives in a continuum: by introducing additional weight parameters, one could adjust the exact degree to which the agent responds to socially available information. Having no explicit weights results naturally from the payoff scheme, but this shouldn't be taken as a fundamental limitation to a cautious learning strategy.

[18] cf. Elkin and Wheeler (2018) for discussion of steadfastness as a response to disagreement

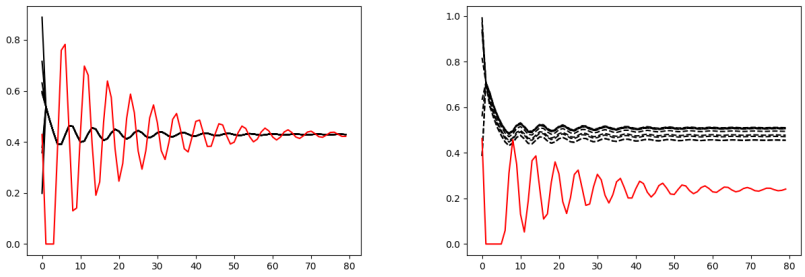

Figure 4.1: Opinion dynamics under a single opinion leader, simple learners (left) and cautious learners (right).

4.3 Results and Discussion

4.3.1 Model Dynamics

The original motivation for constructing a framework to model strategic behavior in opinion dynamics is a concern with the interaction between influentiable and intentionally influencing agents. To understand the basic dynamics, it is useful to focus on a small number of stylized examples.

First, consider a population of basic social learners (Eq. (Social Learner)) combined with a single opinion leader (here modeled by Eq. Social Influencer. A typical run is depicted in Figure 4.1, together with the same scenario using cautious social learners (Eq. Cautious Social Learner). Before going deeper into the interpretation, a few basic observations are in place. There is little complexity in the opinion evolution in the social learner scenario. They effectively converge immediately the way the behavioral rule is defined; this simplistic mode of convergence can be modified by basic manipulations of the learning rule, as explored in Appendix 4.7. In practice, social learning will often take place at a slower pace, but with similar consequences. Second, the expressed opinions of cautious learners generally fail to fully merge, due to their tendency to retain part of their initial signal throughout the process. As individuals, however, they converge to a steady s_i.

Furthermore, the opinion leader exhibits cyclical behavior. This is due to the overadaption of social learners to the belief represented by the opinion leader, which triggers that agent to steer them backwards again after causing them to move beyond its preferred opinion. This phenomenon occurs regularly, but there are also simpler dynamics where the population converges to equilibrium without oscillation.[19]

These scenarios provide stylized descriptions of certain phenomena in the process of scientific discussion. The most obvious application comes from externally motivated research trying to promote certain outcomes, exemplified by instances of corporate sponsored re-

[19]Whether there is oscillation or not depends on the initial distribution of signals, and the type of social learners. Cautious learners create less oscillation, since they are less prone to overadaption.

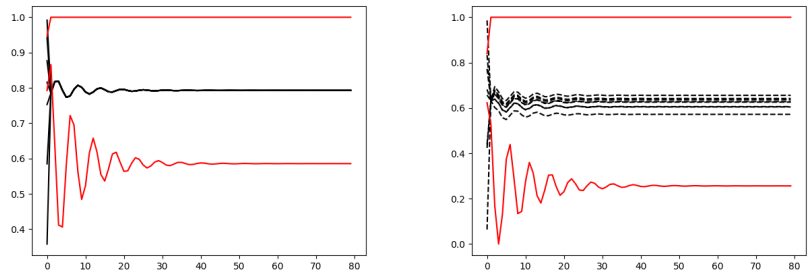

Figure 4.2: Dynamics under dual influence, simple learners (left), cautious learners (right).

search in medicine (cf. Holman and Bruner, 2015, for a discussion and alternative model). The high degree of specialization in science both requires the utilization of socially available information and makes the influence of a single agent plausible. It is also worth noting that the manipulative scientist in the model is *not* trying to steer the group to an empirically entirely unwarranted conclusion, but rather to whatever happens to be the information they actually gathered themselves. The interested external party could simply have chosen to sponsor the scientist with the most convenient results, without rendering them entirely intellectually dishonest.[20]

In practice, corporations may also resort to more extreme measures (cf. Oreskes and Conway, 2011, ch. 1), which could easily be represented by setting the initial signal of the manipulative agent to an arbitrary value instead of the noisy signal. But in the case of pharmacology, a company often has an incentive not to put forward arbitrary claims. On the one hand, they might be interested in downplaying side effects of a new drug and exaggerating its efficacy, on the other hand, they have to avoid legal liabilities.

The cyclical change is the most puzzling feature with regard to this interpretation. If an agent's declared opinion would oscillate in such an obvious way, that person or lab may quickly loose reputation. But in practice, it is quite difficult to determine whether someone is deliberately exaggerating their findings in different directions, or is just trying to weigh in all available data. Furthermore, while it is obvious from the modeler's stance that no new information enters during the modeled process, the same is not assumed to be true from the participants' perspective.

Throughout the modern era, science was rarely influenced as heavily by a single agent as in the single influencer scenario. Even relatively small, specialized groups tend to exhibit a competitive dynamic of multiple influential agents. Once again, the analysis has to focus on simple cases. Therefore, a second potential opinion leader is introduced, trying to steer either a group of simple social learners or cautious learners. Typical dynamics are depicted in Figure 4.2.

Figure 4.2 showcases the basic dynamic of scenarios with two potential opinion leaders. A competitive dynamic ensues between the leaders with both of them exaggerating their

[20]This is not to say that the scientist is not to some degree epistemically blameworthy, but their claims are still connected to actual empirical results and not, for example, based on makeshift data.

results towards the extremes of possible opinions. If the two opinion leaders are initially both on one end of the range of signals and pulling towards the same general direction, the dynamic shifts to the range between their two signals after an initial phase of the opinion leaders pulling in the same direction. The initial phase resembles the scenario with only one influencer.

As the simulations are set up, there is an inherent asymmetry due to the expected value of the signal distribution at 2/3. This incentivizes the opinion leader on the lower end of the range to oscillate more often, since there is more room to exaggerate successfully and turn back afterwards.

Once again, the difference between simple social learners and cautious learners is easily observable: simple learners end up in perfect agreement with each other, while their more epistemically conservative counterparts tend to retain disagreement. As expected, the latter keep up disagreement in a competitive scenario, too. To illustrate the kind of debate, consider the development of sociological theory from the middle to the end of the 20th century, the history of which it is extensively reconstructed by Joas and Knöbl (2004).

There are two extreme answers when it comes to the question of what makes a good sociological explanation, and sociological theories are often substantially defined by their stance on this methodological question. On one extreme, sociological explanation has to reduce to individual attitudes and actions entirely; macroscopical entities cannot figure in genuine explanations.[21] This end of the spectrum is probably best represented by neoclassical economists, but rational choice sociologists take this position within the sociological discourse.

At the other extreme, individuals almost fall out of the explanatory picture. The paradigmatic proponent of such an account is the German sociologist Niklas Luhmann, whose position is referred to as radical functionalism. On his account, what explains social phenomena is the functional requirements of social systems and their subsystems.

Much of the theoretical developments in sociology can be interpreted as working on the middle ground between these extreme positions, representing the larger mainstream that is represented in the simulation model by social learners. Over the course of time, either of the extremes more successfully influences thoses agents occupying the middle ground, creating attracting forces to one extreme or the other. At least in this application, the scenario featuring restrained social learners who end up not fully converging offers a more compelling representation. Even the oscillatory behavior becomes intelligible under this interpretation, since agents are not necessarily to be understood as individual sociologists, but can represent whole theoretical schools, who shift around their position to a limited degree over time due to the influence of different individuals within the school.

So far, the analysis focused on exploring and describing possible dynamics, but not much of normative relevance has been said. But of course the model strongly bears on the question how strategic opinion expression impacts the accuracy of an agent population. At

[21] Even defenders of a very strong version of this kind of individualism can of course accept the possibility that such explanations are pragmatically useful. Such explanations are, however, always assumed to be reducible to agent-level explanations.

this point, the environment and process components of the argument are established, and I turn to the evaluation of the collective behavior depicted, arguing for a general limitation to optimal strategy choice in the assumed environment.

4.3.2 An Accuracy/Robustness Trade-Off

Informally, it is easy to present a line of reasoning supporting the claim that there is a tension between the most effective social learning strategy under optimal conditions as opposed to a strategy that is less accurate in a wide variety of social epistemic situations, but more robust to imperfections within the environment. The reason is that the best strategy under optimal conditions takes into account all socially available information, and since circumstances are optimal by assumption, this leads to the most accurate state of belief.[22] At the same time, such a very credulous strategy performs poorly under less favorable circumstances. A strategy that instead discounts socially available information to some degree is less prone to social influence in general, which is advantageous under suboptimal epistemic conditions.[23]

This argument can be recovered within the framework of strategic opinion formation by the following simulation experiment: the populations of the above scenarios are run for a large number of times, in the initial step – unbeknownst to the agents – all revealing their signal, which provides a lower bound for the error scores relative to the signals they actually received. Error is measured by quadratic distance to the true value of the variable of interest represented as the mean of the noisy signals the agents receive.

It is noteworthy that under these assumptions, there remain multiple ways of measuring accuracy: for example by either computing the error score of the mean of the agents' beliefs, or by calculating the individual quadratic error scores and taking their average. In the discussion so far, more emphasis was put on individual accuracy in the social process, but for certain scenarios it is more interesting to take a measure of collective accuracy into account in addition or even instead. The results are summarized in Table 4.1.

The left-hand values in the first data row provide the benchmark accuracy attainable by aggregating the signals actually received by the population. Since agents initially reveal their signals, the squared error of the mean belief would be the error made by a population comprised entirely of simple learners, because such agents would simply converge to their average signal. A remark to avoid misunderstanding is in place: it is not known to the agents that everyone reveals their private signal in the first round; otherwise, the social learners would ignore any opinion expressed later in the interaction. Furthermore, nobody has information about the strategies constituting the population of the current game. Without this assumption, the agent should try to filter out the beliefs expressed by agents attempting to exert social influence.

[22] A similar argument has been put forward in a different formal framework investigating epistemic norms (Mayo-Wilson, 2014).

[23] This argument is also similar to the claim defended by Zollman (2007) that restricting the communication structure of a community can protect them against fallacies related to premature convergence.

	Simple Learners / Single Leader	Cautious Learners / Single Leader	Simple Learners / Competing Leaders	Cautious Learners / Competing Leaders
$t = 0$	0.006/0.056	0.006/0.055	0.006/0.56	0.006/0.055
$t = 10$	0.048/0.051	0.031/0.036	0.024/0.060	0.017/0.052
$t = 20$	0.054/0.055	0.034/0.038	0.025/0.060	0.017/0.052
$t = 80$	0.055/0.055	0.034/0.038	0.025/0.060	0.017/0.051

Table 4.1: Accuracy scores (squared error of belief mean/mean of squared error of beliefs), averaged over an ensemble of 500 runs.

In all scenarios, the existence of strategic behavior leads to higher error scores, i.e. less accurate beliefs. Comparing the scores of populations of simple learners with those of cautious learners, the argument goes in favor of cautious learners. But this type of agent has worse results in expectation in the ideal scenario of a population consisting entirely of cautious learners, since they do not converge to the mean of their signals, which is the best estimator of the true value. Thus, cautious learners are worse under optimal conditions, but are able to outperform simple learners under strategic influence.

There are many other elements of a social epistemic situation that can hamper the effectiveness of a strategy fully incorporating socially available information. Miscommunication, dependencies between the signals of some of the agents or simply highly unreliable agents could lead to problems similar to those posed by strategic interaction. But it is plausible to conjecture that these additional hindrances support complementary formal arguments to support the informal claim of a trade-off between robustness and accuracy.

The data also suggests a second interesting argument concerning the relationship between a single agent trying to influence the group, and two competing opinion leaders. Competition drastically improves the error scores, as the last row exhibits. This holds since competition creates a mainstream between the most extreme positions, and since the exact expressed opinion of an opinion leader still depends on their own underlying signal, the mainstream ends up at more accurate results when incorporating information provided by two competing influences.

As in the single influencer condition, it is advantageous to be more cautious, which is less obvious than in the basic case, because it implies discounting the strategic agents, but also agents who are partially revealing their signal. But under the model assumptions it is still recommendable to discount this information, since it indirectly duplicates the opinions of the strategically motivated opinion leaders.

Thus, contingent on the prevalence of strategic opinion expression, competition increases accuracy. This result, unlike the trade-off between maximal accuracy and robustness, is more dependent on the specific modeling assumptions. For example, the model does not represent the allocation of resources, which is likely affected by competition.

Other authors have argued that competition also is conducive to an optimal allocation of scientists – one of the major resources in science – to projects (Kitcher, 1990), but there

is significant room for doubt: if science is at least partially governed by similar forces as commodity markets, market failure is also a possibility to be taken seriously. Hence, the recommendation of competition in epistemic enterprises is strictly conditional on the appropriateness of the stylized model in a given application.

Finally, as I alluded to before, there are competing norms to evaluate the resulting pattern of belief resulting from a decentralized communication process. The argument for a trade-off between robustness and accuracy is based on an individual-level assessment of the agents aggregated by calculating the mean. This argument carries over to several other evaluative standards, such as maximizing the minimum accuracy or maximizing the best accuracy within the group of learners given the opinion leader is excluded: simple learners fully merge, and therefore there is no difference between minimum, maximum and average. Cautious learners, in comparison, converge but fail to merge in general, and therefore perform worse under norms such maxi-min, where the worst performing agent is considered.

But it is possible to imagine epistemic goals invalidating the argument. Imagine an application, where convergence itself is extraordinarily valuable. In that sense, what the above argument establishes is not the universal impossibility to devise an optimal strategy in decentralized information aggregation, but really only a specific problem for achieving consistent levels of accuracy across a variety of unknown environments.

4.3.3 The Evolution of Misrepresentation

So far, these results exhibit the dynamics of an opinion expression process, limited more or less to one specific question. It offers the observer a sketch of relevant behavioral patterns on the level of a concrete controversy; the next step is to take a more macroscopical perspective.

In the social process of science, agents are not merely engaging in inconsequential debates, but some scientists are more successful in the long run than others. But according to what metric? One obvious measure would be individual-level accuracy, as described by Equation 4.1. This is the most obvious choice when implementing a reward structure for the social system of science. However, it is not the only relevant consideration. For starters, it is important to find *significant* accurate results (Weisberg and Muldoon, 2009) and often to be the first to find them, i.e. to be original (Strevens, 2003).

The following analysis takes significance for granted (otherwise agents would not engage in social exchange of information at all). Inaccuracy is taken as the squared deviation from the epistemically best possible result given the collection of signals provided. Given that all agents draw from the same distribution and therefore their signals share the same variance, the best attainable degree of belief is the mean of initial signals. Though generally plausible, sometimes it does not matter what would have been the best judgment possible given the evidence, but accuracy is judged after the fact by comparison with the actual value of the variable under consideration, or either truth or falsity of a hypothesis.

In the instances of the deliberation game we are interested in, the expected difference is small, but in particular conceptually, the two standards should be kept distinct. Fur-

thermore, in a different application context, contingent on the assumptions on the relation between true value and received signal, the difference can be increased arbitrarily.

Originality is less obvious to be defined within the model. Part of the difficulty to frame originality in the model is that the represented process brackets the more creative parts of science, such as theory construction or experimental design.[24] The component depicted by a strategic model of opinion formation most faithfully represents part of the consolidation phase of scientific knowledge, where experimental results and theoretical interpretations are on the table, but have not yet been aggregated and unified into a canonical account.

But there is still room to identify certain outcomes as instantiating originality. An agent can at least appear as the source of an eventual consensus (or partial consensus) opinion. In terms of the model, they can be rewarded for minimizing the distance between the final opinion distribution[25] and their original signal. The rationale is that it would seem to the community as if an agent on whose signal they converge had been correct all along, and by analogy this argument can be expanded to cases of incomplete convergence (and in principle, even fragmentation or polarization).

Against the background of these payoff schemes ranging across multiple instances of the opinion expression game, it becomes possible to take an evolutionary perspective on the behavioral strategies defined in the previous section. The question becomes how the composition of a given agent population evolves conditional on a combined reward scheme of the form

$$r_i(s_i) = \alpha(s_i - \mu)^2 + (1 - \alpha)(b_i - \frac{\sum_{j=1}^{n} s_j}{n})^2 \tag{4.7}$$

with $\alpha \in [0, 1]$. The scheme weighs the squared difference between the truth and the focal agent's chosen strategy by α to represent rewards for being accurate. The second component, weighted by $1-\alpha$, describes being payed off for moving the population average opinion to one's own initial signal. This represents, within the constraints of the model, rewards for originality.

For the purpose of evolutionary analysis, discrete replicator dynamics without mutation is used (adapted from Weibull, 1997, p. 113) where $x_p(t)$ is the proportion of agents employing strategy p, e.g. cautious learning, at time t:

$$x_p(t + \delta) = \frac{1 - \delta + \delta r_p(t)}{1 - \delta + \delta \bar{r}(t)} x_p(t) \tag{4.8}$$

where $r_p(t)$ is the expected reward for playing p at t and $\bar{r}(t)$ the average reward across the population at that time. The presence of δ implies that only a proportion of δ agents is exchanged each time step.[26] Importantly, payoffs are determined from the opinions expressed in the final stage of the opinion expression game.

[24] For account of the ingenuity required in these components of science, cf. Kuhn (1970).

[25] This measure is represented by pairwise squared distances to retain structural similarity to accuracy.

[26] Besides renaming some of the functions for convenience, this version leaves out the background reproduction rate β present in Weibull's formulation, which is assumed not to be relevant here. This assumption is an idealization that might warrant relaxation to check robustness elsewhere.

The evolution described by this model corrects a strategy's proportion for the next generation by comparing that strategy's average payoff to the population average payoff. If a strategy is highly accurate (if that is determining the payoff), but the majority of the population outperformed it, its prortion decreases. Since the adaption of successful strategies takes time, only δ percent of the population is exchanged in one time step. In terms of science, tenured faculty tends to persist – for better or worse – until their retirement, and can stick to any strategy that is not too inferior compared to the population.

Not only does this model of strategy evolution possess the virtue of simplicity, the discrete time evolution also mirrors funding and hiring cycles in academia, which are often aligned with teaching or central funding allocation. Mutation is left out of the picture mainly to restrict the parameters space; failure to imitate correctly[27] as well as experimentation with new (or currently unused) strategies could be included for a separate investigation.

The setup for the following simulation experiments is as follows: at each evolutionary step, the population is split into groups of 10 agents playing the opinion expression game for 50 timesteps. Then, the payoffs are computed from final expressed beliefs per group and aggregated for the whole population to calculate the strategy proportions for the next round, with $\delta = 0.1$. Assuming an initial equidistribution, the set of four strategies – naive learner, cautious learner, strategic exaggerator and steadfast agent – would result – contingent on the number of agents to be assigned a strategy – in a large number of possible scenarios. Since the focus of this essay is at its core on the effects of attempts in strategic influence, the analysis focuses on the three types of scenarios containing exaggerators plus one other type and a fully mixed population. The results are depicted in Figure 4.3.

A few cautionary remarks are in place before interpreting these results in detail. The data contains a significant amount of random variation, meaning that it should be read qualitatively rather than quantitatively. In particular, the rate at which the strategically misleading population goes extinct varies quite heavily, and any result close to 0 represents a scenario where it will go extinct eventually. Second, while the results for three of the four cases exhibit strong similarities in their macroscopic behavior, the underlying mechanisms differ in important ways. Given these cautionary remarks, we can turn to the interpretation.

For all populations under consideration, increasing α drives down the success of strategic exaggeration quickly: the more accuracy-based the reward structure is, the less does strategic exaggeration pay off. What exactly quickly means in this context depends on the projection of generations within the model onto any given target system.

Take, for example, the following interpretation: A single opinion expression game represents a highly focused, contained debate, i.e. one carried out at a specialized conference, potentially carried on in the conference proceedings. Rewards had to be distributed shortly thereafter, for example in the form of further conference invitations. Extinction is implemented by omission of invitation, leading to lack of recognition. In such a case, a run of the simulation could be mapped onto a normal scientific career (50 conferences being quite

[27] I thank Kevin Zollman for pointing out the relevance of mutation for the most accurate representation.

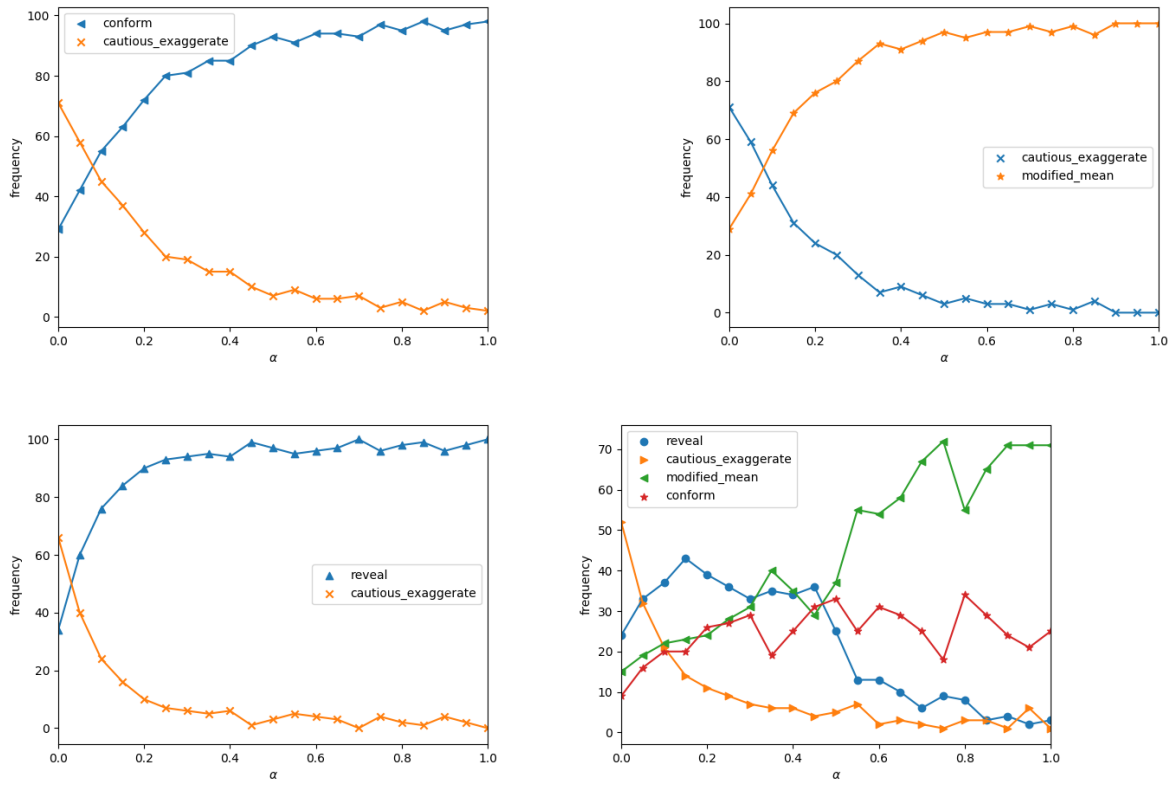

Figure 4.3: Strategy evolution for varying proportions of accuracy-based and self-centered coordination payoffs.

low of an estimate). Such an interpretation is of course not uniquely adequate, and also has further implications.

The need for a fairly immediate allocation of payoffs reduces the access to reliable accuracy estimates, since in many cases, the debate simply cannot fully be settled in an agreed-upon set of results. If the debate concerns speculations about the outcomes of easily implementable experiments, accuracy could be very significant in determining rewards, but often selection committees have to stick to proxy criteria, such as apparent or actual originality and community impact.

This provides a useful frame to the central result contained in all of the above datasets: while strategic exaggeration is soft to evolutionary pressure in an accuracy-driven environment, where accuracy is unavailable as a criterion, it can be successful. But even in such a setting, it cannot take over the whole population, since it creates evolutionary pressure on itself. This leads to the obvious question of what, in any given case, determines the evolutionary dynamic; and at least in the first three scenarios, the question can be answered employing the previously established understanding of microscopical dynamics.

In all three cases, there are four distinct types of subpopulations generated from the overall population for a generation. For the naive learner scenario, these are pure groups of social learners, pure groups of exaggerators, groups with a single exaggerator, and groups with multiple exaggerators. I discuss their effectiveness in finding the best estimate of the truth in turn.

1. A pure group of naive learners converges to the best estimate possible based on their signals, namely their average.

2. A pure group of exaggerators contains two isolated individuals, one on either extreme of the opinion spectrum. These two exaggerate beyond their signal towards the extreme, while the rest of the agents converges in the center, creating suboptimal outcomes for the extremists.

3. A mixed group with a single exaggerator converges on the exaggerator's signal in the long run, leaving all members with the same accuracy-based payoff determined by the exaggerators signal quality.

4. A mixed group with multiple exaggerators behaves essentially like a pure group of exaggerators, with the two most extremely signaled exaggerators receiving strong negative payoffs.

Overall, exaggerators cannot outperform naive learners in terms of accuracy, explaining their rapid demise in scenarios where accuracy is crucial to the reward scheme. In scenarios that are based mostly or entirely on coordinating the group on one's own signal, exaggerators are able to profit in subgroups otherwise consisting of naive learners, while naive learners never perform particularly well on this originality measure. Therefore, originality payoffs are relatively even throughout all groups except for those mixed groups with only one exaggerator who drives the dynamic.

The analysis for cautious learners is very similar, the main difference being worse payoffs in pure groups of cautious learners and lower maximum social coordination payoffs due to non-merging.

The case of steadfastly revealing agents is slightly different. These agents basically represent the case of asocial scientists, who completely filter out socially available information. The effective result in evolutionary terms is very similar, but the underlying dynamics differ. Under social coordination payoff, basically everyone receives low payoffs – as compared, for example, to a group of naive learners – since the population is constituted by agents who are not desiring coordination. This makes it virtually impossible for strategic exaggerators to succeed. At the same time, steadfast agents also perform poorly, since they leave all socially available information on the table, even if they partake in an instance of the opinion expression game where nobody engages in strategic exaggeration.

In the more accuracy-driven part of the parameter space, however, exaggerators once again are on the decline, since their strategy drives them away arbitrarily from their own signal, leaving them at a less informed position as the partially informed retaining at least their own signal.

Finally, the mixed population of naive and cautious learners, influencers and steadfast agents gives an impression how more complicated mixtures of strategies can evolve. The basic pattern is similar: for low α, there is little to be gained by trying to learn socially, and strategic exaggeration pays off, but never takes over the whole population due to the cost influencers inflict on each other.

Interestingly, it seems that cautious learners become most successful when accuracy pays off more and more. This is a deliberate product of the finite time span under consideration. In the long run, both steadfast agents and exaggerators are driven to extinction in such a scenario. But as long as such agents are part of the population, cautious social learning provides a competitive advantage over naive social learning, since it discounts misleading signals to a limited extent as a consequence of discounting all socially available information.

This matches the accuracy estimates from the previous section and precisely mirrors the advantage of cautious learners in an ensemble of games with a static population. Why are we licensed to put the equilibrium consisting only of naive social learners to one side? There are two reasons, one more technical, the other immediately regarding the target.

On the technical side, any significant amount of mutation added in can at least temporarily shake up the naive learning equilibrium. Furthermore, even a marginal component of coordination-based payoff creates an advantage for cautious learners and steadfast agents, though as the above two-strategy populations show, either case leaves exaggerators under strong evolutionary pressure. Given these complexities, it seems more informative to look at the out-of-equilibrium dynamics, which depend less on assumptions of non-mutation and similar random variation.

Second, when it comes to the target of scientific group deliberation, discussion and opinion exchange, it is a very real possibility that strategic exaggerators are fed into the system at a steady rate. Such influx might be the product of big money interests or the personal engagement of scientists with new projects. As an example, economists enter-

ing the debate on failures of rationality in cognitive psychology may not be subjected to the same selective pressures as the resident psychological researchers in this debate; their rewards are still determined by the substantially different standards for research in economics. Therefore, the out-of-equilibrium dynamics are a highly relevant part of the outcomes provided by the model.

Summing up the results of evolutionary analysis, there are good and bad news. On the one hand, the possibility of strategic exaggeration limits the rationality of naive social learning, instead suggesting to discount socially available information. But the model also supports the conclusion that an orientation towards accuracy strongly limits the efficacy of strategic exaggeration in the long run.

Furthermore, the demise of socially ignorant steadfast agents underlines that the utilization of at least some socially available information actually benefits agents in terms of accuracy. Of course the simple strategy of cautious learning employed here is in general not optimal, but it showcases an exemplar of a species of agents who are partially robust against strategic manipulation while still incorporating information provided by their social environment.

4.4 Alternative Application: Jury Deliberation

So far, the analysis interpreted the model as representing scientific discourse. Science provides a particularly interesting case, since strategic behavior is often considered unimportant where everyone shares the same epistemic values.[28] Therefore, applying it to science without simply denying the importance of epistemic values to scientists appears a particularly challenging enterprise and therefore a particularly strong corroboration of its power.

However, it is interesting to look at other applications, where a strategic interpretation of opinion expression is less controversial. Legal epistemology offers a prime example. Since it is generally the best known, I shall consider the example of a jury trial roughly in accordance with the legal system of the USA, and will focus on defense, prosecution and the jury itself, leaving the judges out of the picture for simplicity.

The scenario of jury deliberation can thus be reconstructed in terms of the deliberation game model: the hypothesis in question can be the defendant's guilt, or anything more complicated, such as the amount of damages to be paid to a plaintiff or the likelihood that a particular relevant event has taken place. Depending on that, the values would represent the credence of agents involved in "guilty", or their best estimate of the variable in question, normalized to the interval $[0, 1]$. Lawyers on both sides have the explicit mandate to drive the opinion to whatever their party in the trial deems accurate; though they are of course not legally allowed to lie, as argued in the case of science, there is a lot

[28] There is a tradition in particular in the sociology of science that more or less flat out denies the actual relevance of epistemic values. While it is correct that science is a social system, I conjecture this approach generally underestimates the actual relevance of epistemic values due to its failure to acknowledge that suboptimal outcomes may be the consequence of epistemically virtuous behavior on the individual level.

of leeway to represent the information available to oneself without straightforwardly lying.

What type of agent would best represent the jury? That depends on the particular instance; some jury members might themselves be motivated to take the role of opinion leaders, it might contain simple and cautious learners, and, making deliberation particularly difficult, it could contain agents entirely steadfast in their opinion. All these types of agents could, for example, be found in the famous theatrical depiction of jury deliberation by Rose (1997).

Model time maps onto the range of the whole trial under this interpretation. During the public part of the trial, jury agents receive their signal – potentially a combination of prior beliefs and evidence provided throughout the trial. The deliberation phase may either be mapped onto the isolated phase of jury deliberation, such that strategic influencers are only other jury members, not the lawyers. Alternatively, the lawyers could be represented as the strategically representing agents.

The jury deliberation case therefore needs to model the initial signals as correlated: though the lawyers may have information unavailable to anyone else through their clients, and the jury may not have access to all the same evidence as, e.g. the prosecution, they are learning largely the same evidence as their initial signal. This modification is, however, easily implementable.

Thus, as it turns out, a jury trial is another natural target for the deliberation game model. Let me end the sketch of this application to legal epistemology with two observations about the implications for a jury trial.

First, the equilibria of the game incentivize the lawyers to exaggerate, if the mean of the jury agents deviates from their preferred position. To illustrate this, assume that the jury consists of simple learners and converges to 0.5. The preferred position of the prosecution is 0.9, for the defense it is 0.1. To maximize, the prosecution has to play 1 and the defense has to play 0, since otherwise, the resulting point of convergence of the jury agents would move away from their respective preferred positions. While this is a particularly simple case, the example can be generalized. If the other party to the trial reveals their preferred position, it is individually rational to exaggerate under the modeled assumptions.

This result might not be particularly surprising, since it is akin to standard problems in bargaining. However, it takes a different turn here because of its epistemic implications.

This leads to the second observation, concerning the relevance of accuracy in rewarding agents within the legal system. While there have been large waves of exonerations long after the fact, there is generally no remotely immediate way to check the correctness of the judgment passed in a trial; otherwise, much of the rationale for running the trial would dissolve. But as the evolutionary analysis in the context of science shows, to create pressure for epistemically purer agents, accuracy cannot be replaced with simple proxies in the payoff scheme. But in the legal system, accuracy-based payoffs are extremely difficult to implement. Among other things, jury members often will serve a very small number of times.

Of course this whole argument comes with the important caveat that it relies on a highly stylized model of deliberation and opinion aggregation. Therefore, it is recommendable to be very cautious in claiming external validity of any kind. Since much of the debate in

legal epistemology has to be executed on similar levels of abstraction due to the sheer lack of data (Laudan, 2006), the argumentative support can still be a valuable contribution to an empirically difficult topic.

The adversarial system in particular creates strong strategic incentives easily overlooked when focusing on formal constraints on the procedures – a dangerous omission, which the deliberation game model exposes: under the right circumstances, as I argued, competition increases accuracy. This insight, in its informal way, could be identified as the original motivation to establish an adversarial system of defense and prosecution. But it is entirely unclear whether the actual world instantiates the right conditions, and therefore, whether the overall effects of competition in this context are beneficial or detrimental.

4.5 Challenges in Strategic Opinion Modeling

The stylized nature of models of social opinion dynamics generally limits their quantitative fit. In particular, this makes their application in prediction inherently problematic. But besides these obvious standard limitations of highly stylized theoretical models, there are certain challenges for models such as the deliberation game. Discussing these issues both improves understanding of the model as well as its scope, and has implications for opinion modeling in general. Therefore, it is worth investigating two such challenges more closely.

First, a very broad interpretation of the term "strategic behavior" is assumed throughout the discussion. In standard game theoretic analysis, agents are assumed to be perfectly forward-looking. Boundedly rational agents modeled and simulated here are rather backward-looking, adaptive agents under that regime of terminology. Hence, they can only be considered strategic in a limited sense requiring further specification.

The boundedly rational agents can be interpreted as behaving strategically in two ways: (a) While it is desirable to introduce more forward-looking agents, Hegselmann et al. (nd) argue that perfect forward-looking behavior is excessively demanding in non-trivial models of opinion dynamics. Constructing agents with explicitly stated motivations and a behavioral strategy – naive best-response learning – that converges on an optimal strategy for many games without overly taxing the cognitive abilities of the agents presents a substantial advance towards fully forward-looking agents. (b) Strategies employed by boundedly rational agents are constructed to allow for an interpretation that assumes more strategic depth than the equations themselves unequivocally state. Part of an agent's motivation to utilize, for example, a strategy of cautious social learning lies in their anticipation of potential strategic misrepresentation of signals. Even though their behavior on the level of the internal model of the agents is entirely backward-looking, the exogenous strategy choice can be understood as anticipating the potential behavior of competing agents.

The second point of significant theoretical worry is the inability of the modeled agents to ever *learn* one another's strategies. There is some plausibility to this idealization, since it is highly problematic to draw conclusions about another agent's honesty based solely on the difference between one's initial signal and expressed opinion. If the agent happens to have received a misleading initial signal, this introduces a bias in the evaluation of everyone

else's reliability.[29]

But the model dynamics suggest that a perceptive agent should be able to identify suspicious patterns; most obviously to an external observer, the oscillating movements of strategic influencers seem highly characteristic, even though the caveat mentioned above holds for potential confounding factors in a more descriptively rich account. Even if it turns out that there is no robust rule to learn another agent's authenticity, it would be of great value to understand the impact of this additional mechanism – or mechanisms – on the process of opinion formation. For now it has to be left for future research.

4.6 Conclusions

The analysis and discussion of an agent-based model of motivated exchange of beliefs supports three main claims: First, the introduction of various motivations and strategic responses in the reach of boundedly rational agents creates a rich dynamic of beliefs. Depending on the composition of the population, phenomena from oscillating fluctuations to convergence, fragmentation or polarization occur. Second, the model formalizes and confirms the informal claim that there is a trade-off between robustness and accuracy across epistemic environments.

Third, the evolutionary study provides an argument that even mildly accuracy-driven reward structures create substantial pressure on strategic manipulators. Their prevalence may, however, be a result of the system moving out of equilibrium, an influx of such agents, or their entirely independent motivation by non-epistemic motives.

The first two insights in particular shed light on the example of researchers in clinical psychology that provided our starting point: given the suggested assumptions on the composition of their population, they have good reason to discount socially available information to some degree. One important caveat applies: even with perfect information about agents' motivations, the model should not be expected to quantify this qualitative result, given its highly idealized nature.

With respect to the argument scheme of environment, social process and collective goals, the model and its analysis further details some possible arguments and their internal structure. Starting with the last and simplest point, the analysis applies a plurality of possible evaluative norms. Coordination, accuracy and originality can point in different directions when it comes to the evaluation of epistemic performance. Usually, accuracy is presumed to be the norm of choice in epistemic circumstances, but as I sketched, there is no unique translation of accuracy from the individual to the collective variant. In total, different normative standards clearly support different normative conclusions on the social process.

[29] Note that this problem can be circumvented if both an agent's initial reliability estimates and signal are constrained to be accurate to a minimum degree. For a model of expectation-based updating of reliabilities, see the competing accounts of Bovens and Hartmann (cf. 2003, ch. 3) and Olsson and Vallinder (2013). However, neither of these approaches helps under unfavorable conditions (Hahn et al., 2018).

In terms of the social process, this study focused more on the individual behavior constituting the social process. Varying strategies can be understood as individuals choosing different strategies, but it may also represent social planning influencing the agents' motivation by selection. It also showcases that certain negative arguments, in this case, that neither learning strategy supports a universally optimal social process, can be constructed without specifying all possible social processes. For practical purposes, such arguments are of immense importance, as an inspection of all possible alternative processes is not generally a realistic goal.

Finally, switching from the analysis of individual model runs to an evolutionary perspective showcases how variation in environmental conditions can strengthen arguments by showing a sufficient degree of robustness. In our case, evolutionary analysis supports the claim that cautious learners can be advantaged over naive learners, yet only in out-of-equilibrium dynamics. Generally, any controversial assumption on the decision environment could be tested like that while still representing the static features of the situation under consideration relative to one instance of the argument scheme.

There are several paths for future research in the framework presented, and I shall briefly suggest some of the most promising ones. The theoretically most provocative assumption of the model is the inability of the agents to *learn* their colleagues' strategies within the model. The consequences of including such learning procedures could improve model performance – or lead to further unwarranted discounting of socially available information.

Furthermore, interpretation of the model is not inherently limited to scientific communities. Social groups face epistemic problems from areas as distant as legal epistemology, political philosophy and economic theory. An adapted and reinterpreted version may yield a potential explanation of the polarization of elites in a democratic society, but the possible scope of applications is vast and varied.

4.7 Appendix: Robustness

As the next chapter explores in more detail, the robustness of theoretical ABM to relaxations is a standard part of its analysis. Some of the more ambitious options are mentioned in the model limitations and future directions sections. What I shall consider in the following sections are moderate variations on the naive best-response learning algorithm. The reason to focus on this issue, conditional on how central the learning mechanism is to the model, is that rule's rigidity. While it is very easy to implement, plausibly not only on a digital computer but also in a human being, there is a lot of room for small error, variation or slow adaption to the strategy profile confronting the agent. Therefore, I replicate a part of the above results using relaxations of naive best response learning, which are called softmax and partial best response learning.[30]

[30]I thank Kevin Zollman for suggesting these alternative learning rules to corroborating the results of the main analysis.

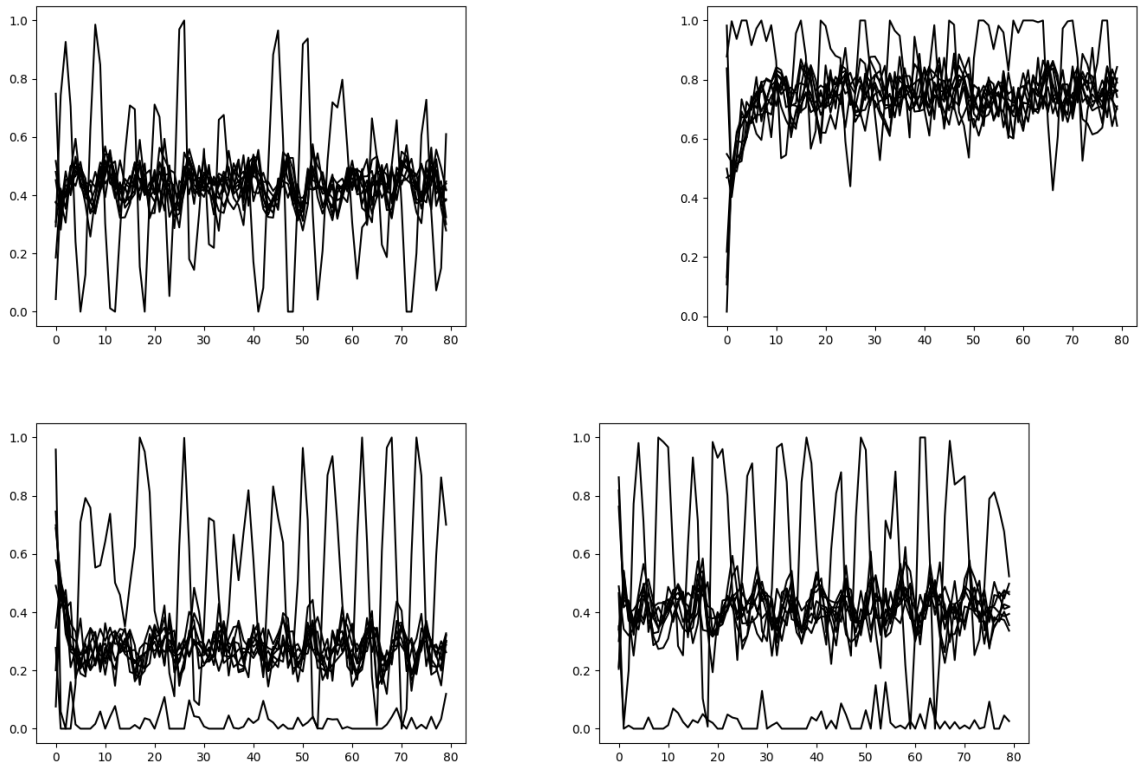

Figure 4.4: Opinion dynamics under softmax learning, naive learners with single leader (upper left), cautious learners with single leader (upper right), naive learners with two leaders (bottom left), cautious learners with two leaders (bottom right).

4.7.1 Softmax Learning

Softmax learning can be understood as taking the naive best response and adding some random noise to the choice. For the purpose of the following robustness checks, instead of playing the naive best response, agents draw from a normal distribution with their best response as the mean. To remain within the modeling assumptions, the choices are cut off at 0 and 1. The variance is set to 0.05, which already creates a highly visible effect, but without turning the agents' actions into a merely random signal yet.

Random variation translates into added noise as compared to the results under best response learning. In general, agents converge not to a fixed value, but to a limited range; however, the high responsiveness of the opinion leader (or leaders) can create repeated strong shakeups. With respect to accuracy (depicted in Table 4.7.1), the overall results are slightly worse scores, but the differences to the naive best response learning results are minimal.

I restrict the robustness tests for the evolutionary process to the fourth and most universal scenario. Once again, softmax learning does not change the qualitative picture.

	Simple Learners / Single Leader	Cautious Learners / Single Leader	Simple Learners / Competing Leaders	Cautious Learners / Competing Leaders
$t = 0$	0.006/0.056	0.006/0.057	0.006/0.56	0.005/0.057
$t = 10$	0.045/0.052	0.033/0.042	0.023/0.058	0.019/0.054
$t = 20$	0.052/0.058	0.036/0.045	0.024/0.058	0.020/0.054
$t = 80$	0.052/0.059	0.037/0.045	0.025/0.059	0.020/0.055

Table 4.2: Accuracy scores (squared error of belief mean/mean of squared error of beliefs), averaged over an ensemble of 500 runs, using softmax instead of naive best response learning.

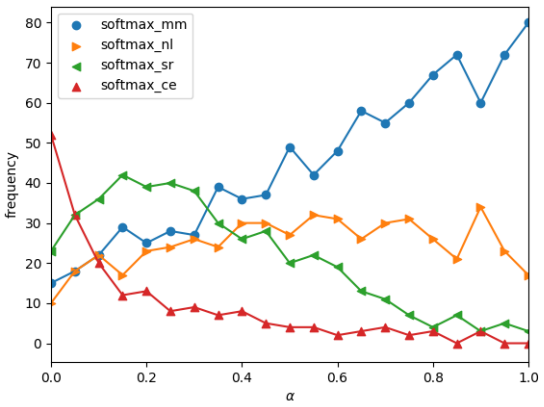

Figure 4.5: Evolution of a fully mixed population under softmax learning.

	Simple Learners / Single Leader	Cautious Learners / Single Leader	Simple Learners / Competing Leaders	Cautious Learners / Competing Leaders
$t = 0$	0.005/0.058	0.006/0.056	0.005/0.57	0.005/0.056
$t = 10$	0.032/0.038	0.022/0.029	0.020/0.054	0.018/0.051
$t = 20$	0.047/0.048	0.031/0.035	0.023/0.056	0.019/0.052
$t = 80$	0.056/0.056	0.036/0.040	0.025/0.057	0.020/0.052

Table 4.3: Accuracy scores (squared error of belief mean/mean of squared error of beliefs), averaged over an ensemble of 500 runs, using partial best response learning.

This should not come as a surprise, given the above explanation of evolutionary results based on the microdynamics and the similarities regarding those same dynamics.

4.7.2 Partial Best Response Learning

Whereas softmax learning may represent, for example, communication error of some kind, partial best response learning provides a model of more conservative updating, be that for reasons of cognitive capacity or an explicit decision of the agents not to get on the bandwagon too quickly. Instead of playing their naive best response, agents choose a convex combination of their previous action and their naive best response. The general form for partial best response learning is stated by Equation 4.9. For the following simulations, both factors are evenly split.

$$o_i(t+1) = \alpha o_i(t) + (1-\alpha) NBR_i(t) \tag{4.9}$$

where $NBR_i(t)$ is i's naive best response to the opinion profile at t and $\alpha \in [0, 1]$ determines the relative weight of i's previous opinion.

The main result for the dynamics (depicted in Figure 4.7.2) is a smoother evolution of opinions. There is less oscillation and the difference between naive and cautious learners is diminished, since all learners are now moderately conservative.

The differences in accuracy to naive best response learning are marginal. The only interesting systematic observation – besides establishing the robustness of the claims made on the basis of naive best response learning – concerns the advantages in the early stages of the process. The above-mentioned dampening of exaggeration and oscillation is responsible for those improvements. The observed reduction in inaccuracy is further increased by the superlinearity of the employed measure of accuracy.

The same remarks on evolutionary dynamics made for softmax learning also hold for partial best response behavior. A difference not efficacious in the evolutionary study here concerns the out-of-equilibrium behavior of partial best response populations: conditional on the population composition, partial best response learners converge either significantly faster or slower than both of the other mechanisms under consideration. For example,

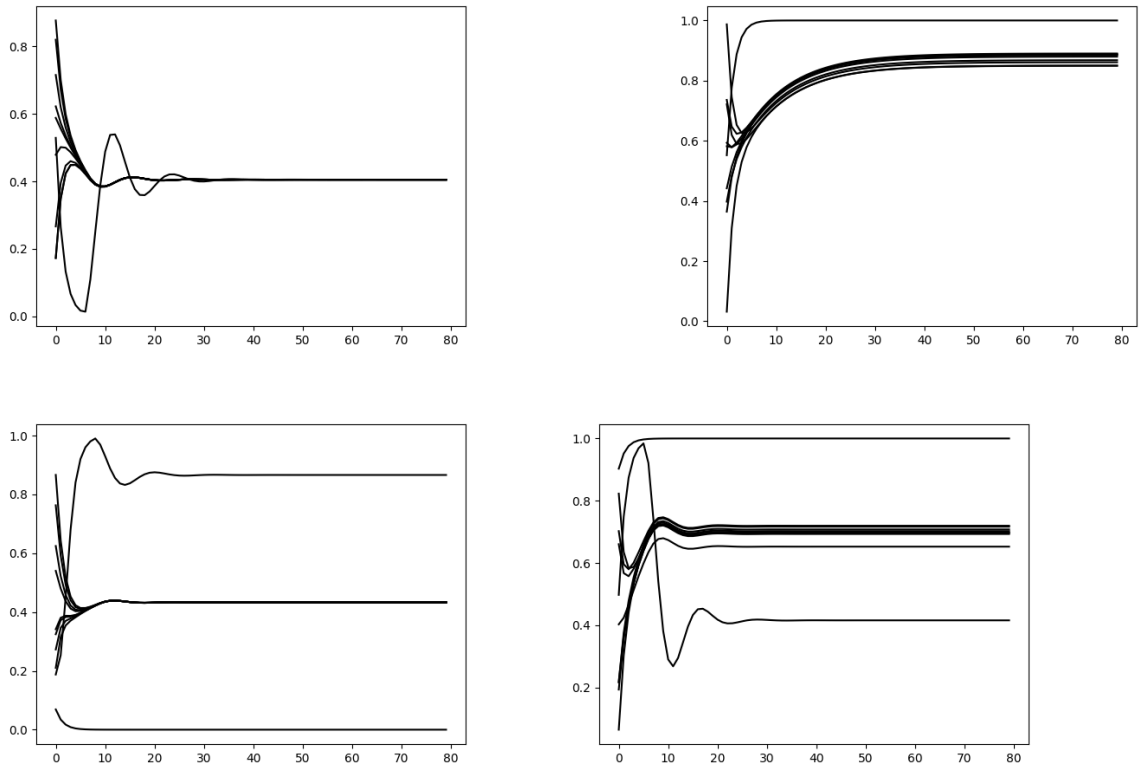

Figure 4.6: Opinion dynamics under partial best response learning, naive learners with single leader (upper left), cautious learners with single leader (upper right), naive learners with two leaders (bottom left), cautious learners with two leaders (bottom right).

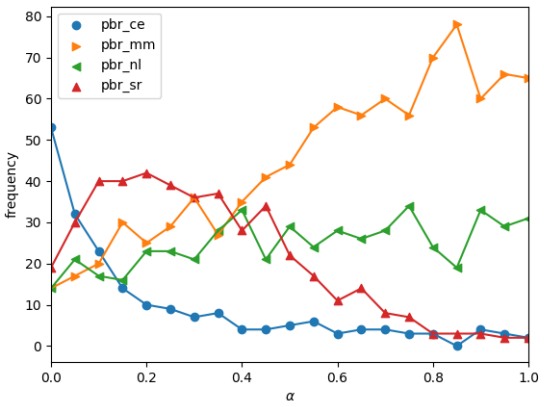

Figure 4.7: Evolution of a fully mixed population under softmax learning.

opinion leaders have an increased ability to coordinate groups, while a pure population of simple learners converges more slowly.

Let me conclude on two final remarks on robustness. First, the results are mostly negative, but not entirely. That is particularly satisfying, since it implies that the assumption of strict naive best response learning is not critical, but at the same time, relaxing it impacts the results to a degree, depending on the exact parameters chosen for the alternative learning rules. In other words, the model is sensitive to modifications, but in a rather continuous, expected fashion.

Second, the robustness results themselves may appear in need of robustness checks, since as rely on the choice of a single value for the respective additional parameters, variance in noise and the combination factor α. But these values are not chosen arbitrarily: increasing the variance merely increases noise, thereby turning the dynamics more and more into a random process independent of its previous states. Decreasing variance lets the dynamic once again approximate naive best response learning, and the chosen value exemplifies the presence of relevant, but limited random variance.

Similar reasoning applies to partial best response learning: increasing conservatism slows the process down further, magnifying the effects already visible at 0.5, while reducing it once again moves the dynamics closer to naive best response learning. The alternative learning algorithms are specific continuations of naive best response learning. Therefore, it is sufficient for the robustness argument to test for a small number of values.

Chapter 5

The Epistemology of ABM

5.1 Introduction

Engineers should be able to build their own tools. Analogously, philosophers should be able to analyze the epistemology of their methods. The epistemology of computer simulations in general has been discussed extensively in philosophy of science[1], but agent-based models and simulations as a subset allow for a more specific analysis.

Agent-based models differ in several respects from the simulation models underlying standard arguments in the philosophy of computer simulations. Those models are usually taken from physics or climatology, and they consist of numerical approximations to a system of differential equations.[2] Agent-based models, on the other hand, are largely not constructed by approximating a pre-existing mathematical model.

ABM often are the formalization of informal social theory or the modeler's pre-theoretic intuitions.[3] Therefore, it is impossible to quantitatively evaluate the error between model and theory. As a consequence, ABM should be considered on equal footing with the often more general, but only qualitative, informal theory. If the computational model of a pendulum introduces too large of an error relative to the underlying theoretical model, it is a defective tool. The same scenario never arises with ABM.[4]

Furthermore, agent-based models are often inherently discrete – or could be formulated as discrete without meaningful loss – removing yet another common source of error between mathematical and compuational models. The modeled agents are usually supposed to have finite memory, finite choice sets and finite computational capacity. There is no underlying theory that requires the assumption of continuity. Even time in an ABM need not be continuous, since all that is relevant is the smallest time interval the agents perceive.[5]

[1] Cf. Winsberg (2010) for an influential monograph, or Saam (2015) for a more recent overview

[2] See Beisbart (2012) for a standard example.

[3] The main class of exceptions to this generalization are models constructed by relaxing game-theoretic models, such as the study by Axelrod (cf. 1984, in particular ch. 2–3).

[4] Of course, an ABM can be rendered defective by *empirical* inadequacy.

[5] Agent-based models on opinion formation, such as the one discussed in the previous chapter, often assume continuous levels of belief; however, in these cases, continuity is generally an idealization, while

For these differences one should expect agent-based models and simulations to be an interesting special case of simulations; a special case that allows defending stronger claims than for computer simulations in general. In the following sections, these four questions in particular will be discussed:

1. How does the use of digital computers impact understanding?

2. Can the robustness of simulation results provide confirmation?

3. Which relationship holds between ABM and controlled experiments with respect to the inferences they support?

4. How are normative models in particular evaluated as sound or adequate?

5.2 The Challenge from Opacity

A running computer simulation, whether it implements an agent-based model or a system of approximated differential equations, performs an excessively large number of basic logical and arithmetical operations within the machine's processing units. This capability to crunch the numbers at a superhuman rate is what enables computer simulations to deliver results that are otherwise unattainable. However, it has been argued that this exact power also comes at a substantial epistemic price, namely the opacity, i.e. epistemic intransparency of the method (Humphreys, 2009).

To fully grasp the threat posed by opacity to theoretical ABM, it is necessary to first explicate the term a little further. Following Beisbart (2012), I assume that the epistemic power of a computer simulation can be captured in its reconstruction as a deductive argument; the initial conditions and the transition rules of the model constitute the premises, and the resulting sequence of states follows. Deductive arguments per se are not intransparent, but in principle open to human inspection.

Furthermore, the premises are also not problematic, since they are, even on the level of machine language, still inspectable in principle. This leaves two loci as the possible source of opacity. Either the sheer length of the conclusion poses a problem, or the actual execution of the argument does. If it is assumed that the conclusion consists of *all* states of the model during the simulation in a large conjunct, it is plausible that the resulting proposition becomes practically impossible to maintain by a human agent.

The mere length of the conclusion only creates opacity insofar as our tools to analyze it, i.e. descriptive and inferential statistics, also count as opaque. For sophisticated statistical techniques, as various kinds of machine learning, the claim to opacity comes down to the same as for computer simulations in general. But for theoretical ABM, the sequence of states is actually sufficiently simple to omit tools themselves prone to the challenge of opacity. Theoretical models of opinion dynamics are often limited to a single state vector

actual agents likely hold beliefs from only a finite set.

evolving over time, which can be visualized without loss.[6] Thus, in the general case, the sheer number of conjuncts in the conclusion could threaten opacity, but theoretical ABM are highly resistant to this problem, as their opacity would be tantamount to the opacity of any large set of data or the methods for its representation.

What about executing the argument? Beisbart (2012) argues that the argument underlying the simulation is actually executed by a coupled system of machine and scientist. The machine is, at least in some cases, indispensable to the execution; otherwise, the simulation itself could simply be eliminated in the context of justification, even if it played a crucial role in the context of discovery. Therefore, the argument is opaque to the scientist in isolation, and only transparent to the coupled system.

At this point, it is worth stopping and asking: Why is the scientific community actually interested in epistemic transparency? There are at least two plausible answers, which may both be true at the same time. Epistemic transparency enables other scientists to check results. The natural analogy is the surveyability of mathematical proofs: another competent mathematician should be able to follow a proof, i.e. it should be transparent to them.[7]

This rationale for transparency is compatible with considering theoretical ABM transparent, since another coupled system of a different scientist and a different machine has the capacity to replicate simulation results. This is not to say that there are no practical problems in following another researcher's simulation-based argument. Lack of documentation and limited computational power are standard problems in ABM replication, but they do not threaten their fundamental epistemic status, just as various limitations in resources and documentation do practically, but not in principle threaten the epistemic status of laboratory experiments.[8]

However, the second reason to require transparency references the scientists themselves as a psychological agent. Under a certain account of science, the researcher does not only want to acquire the power to predict events and successfully manipulate the world in technical applications, but also aims for an *understanding* of the causal processes in the world. Understanding by individual scientists, however, is severely limited in the coupled-system-account: their psychological state alone may or be not be in what Reutlinger et al. (2017) call the state of "grasping", since this is not an implication of the coupled system being in such a state.

What understanding means precisely, and therefore what is required to achieve it, is controversial both among scientists and philosophers.[9] Controversy notwithstanding, all

[6]There is, as Kuorikoski et al. (2010) point out, a risk to overestimate one's understanding of a model in the face of compelling visualizations. But their worry concerns a sense of understanding the *causal* process within the model, while I am at this point only interested in the intelligibility of the sequence of states.

[7]On a related note, mathematicians have, for this exact reason, been suspicious of automated proofs (De Millo et al., 1980).

[8]Note, that this argument fails for certain other simulation models, as replication can become in principle impossible due to the need for parallelization and adaption to a specific high-performance computing machine.

[9]See the discussions by Reutlinger et al. (2017) and De Regt and Dieks (2005).

these accounts agree that understanding is a *pragmatic* notion, and therefore a lack of understanding fails to pose a threat to the justification of conclusions derived by agent-based simulations. This observation, though correct, is insufficient to address the original worry.

The problem is not that inferences based on ABM are not justified – which is closely related to the peer checks previously discussed – but that they fail to achieve a different epistemic goal regarding the mental states of scientists. The scientist tries to understand a phenomenon by grasping a correct explanation.[10] Unfortunately, the often employed notion of "grasping" is notoriously vague and supposedly philosophically primitive.

This vagueness makes it difficult to assess the understanding provided by ABM. For example, does it count as grasping an explanation in the relevant sense if the relation ranges over explanations, phenomena, and coupled systems of computers and scientists instead of scientists alone? To circumvent this problem, I suggest to spell out the motivation for understanding more precisely.

I conjecture that scientists are striving for understanding to acquire a justified *sense of understanding*. The sense of understanding is the purely subjective state of mind that human agents can enter on grasping an explanation, regardless of its correctness. Whether the sense of understanding is justified depends on the precise account understanding; on Streven's "simple view", for example, it requires the scientist to entertain an explanation and for that explanation to stand in a certain relationship to the world (**?**). On another account, such as the contextual account suggested by De Regt and Dieks (2005), the conditions differ. But in all cases, there is a notion of legitimate understanding, to which scientists are implicitly committed, but which is only contingently linked to the sense of understanding.

Note that this account includes a psychological claim: scientists are motivated, at least partly, by their desire to experience a sense of understanding, though constrained by some set of justificatory criteria. The history of science seems to stand witness to this claim, but its truth remains a contingent empirical matter.

If this claim is correct, the use of theoretical ABM poses no threat to understanding. Theoretical ABM can clearly create a sense of understanding.[11] Part of the reason is that the transition rules are interpretable in terms of human behavior, and the simplicity of the models allows the analyst to inspect not all, but still a substantial portion of the executed operations. The fact that no translation between the theoretical model and its algorithmic implementation is necessary contributes further to the sense of understanding by removing technical hindrances.

Of course this does not ensure that the criteria for actual understanding are fulfilled. But as argued above, there is no reason to believe that partial opacity to the human scientist threatens the justification of their inference. Whether such an inference is actually

[10] This is the core of Strevens "simple view". Reutlinger et al. suggest certain important refinements, without changing the core of the theory. (Reutlinger et al., 2017). DeRegt and Dieks also discuss the opposing position of Trout (2002), whom they trace back to the view presented by Hempel (1965).

[11] The discussion of simulation results in Axelrod (cf. 1984, ch. 2) and Hegselmann and Krause (2002) stand witness to this claim.

justified depends on the relationship between the explanation embodied in the simulation and the model's target, and imposes no additional conditions on the mental state of the scientist. Therefore, theoretical ABM can succeed in providing that good that is sought for in understanding.

5.3 Robust Phenomena in ABM

Agent-based modelers are often expending substantial effort to establish the robustness of their results. Variations in assumptions, sensitivity analysis and cross-implementations are standard techniques in the field.[12] But what is supposed to be established by robustness results, what are their limitations, and when should results be fragile to modifications in the model rather than robust?

The analysis of robust consequence of models has a history that ranges back at least to Levins (1968), who suggested robustness analysis as an important tool in theoretical biology. Models in population biology, for example, often have to resort to various strong idealizations, approximations and tractability assumptions to yield solutions. As these assumptions are not themselves justified directly, the model's epistemic status depends on its robustness against variation in these arbitrary assumptions.

Agent-based models may appear to differ, since they often are constructed as simulation models right from the start, and therefore are less reliant on tractability assumptions. However, they are still very much reliant on both idealizations and technical assumptions of various kinds, and the need for robustness is anything but lessened in ABM.

My intention is not to analyze the ongoing controversy on robustness analysis and its epistemic role in general [13], but instead I shall focus on two particular questions: (1) Is it possible to *confirm* the existence of phenomena by establishing robustness across models? (2) Under which conditions should the results of a model break down, or put otherwise: to which variations in modeling should the results be *fragile*?

5.3.1 Non-Empirical Confirmation

At a first glance, the idea of non-empirical confirmation is anathema to modern science. As important as theory is, it only acquires credibility by *empirical testing*. The elegance, consistency, scope and any other virtues of a theory, that is the standard assumption, cannot step in for thorough empirical *confirmation*.[14] The literature on robustness analysis thus understandably either expresses incredulity at the suspected suggestion of non-empirical confirmation (Odenbaugh and Alexandrova, 2011) or tries to establish the increase in credibility provided by robustness analysis as different from confirmation to avoid a seemingly

[12] Axelrod (1997) offers a good example of how difficult it can be to achieve even the simplest form of robustness, namely replicability.

[13] Main positions in the debate can be found in Holman and Bruner (2015)

[14] I presume that accounts like naive falsificationism remain discredited due to arguments along the lines of Lakatos (1968).

absurd claim.

But on closer inspection, confirmation by robustness is not nearly as absurd, at least in the realm of agent-based modeling.[15] Agent-based models of opinion dynamics serve as the running example. A central objective of such models is to establish the conditions under which polarization arises. As Mäs and Flache (2013) point out, various mechanisms have been suggested to explain polarization phenomena: bounded confidence in other agents, negative influence and the homophilous choice of communication partners provide prominent examples.

Now consider the following thought experiment: imagine a medieval scholar, who, by a time traveling accident, gets hold of a digital computer and an introductory textbook on agent-based modeling and simulation to acquire all the basic skills to build the above mentioned models on his own – and so he does. Next, the scholar wonders about the patterns of opinion in the villages surrounding his monastery; given his insight that a wide variety of mechanisms could bring about polarization, is he warranted in increasing his belief to find patterns of polarization, and thereby, acquire confirmation for the hypothesis "for all villages X, opinion polarization is likely to occur"?

The answer is yes, for the following reason: Assume that the scholar attributes some probability $p(M_1)$, $p(M_2)$, $\ldots p(M_n)$[16] for all the mechanisms which are potentially at work.[17] His simulations show that all these mechanisms generate phenomenon R, patterns of polarization. Furthermore, the different mechanisms are not counteracting each other to the best of his knowledge. Therefore, any additional mechanism added to the scholar's knowledge base increases the credibility of finding polarization in any given village.

The point is strongest when a phenomenon is considered impossible or at least extremely unlikely, such as the existence of Giffen goods in economics.[18] A Giffen good is defined as one for which demand increases with rising prices. If a phenomenon is deemed impossible for the lack of a plausible explanation, its existence should become more likely to an agent first encountering such an explanation.

Given the apparent absurdity of its consequence, this argument requires closer inspection. First of all, the kind of robustness is different from the technique Kuorikoski et al. (2010) identify with economic theory. According to their account, the desideratum is robustness against variations in merely mathematical tractability assumptions, not the core causal mechanism or substantial idealizations. This actually establishes the relevant difference between the two cases: increasing the number of mechanisms that could bring about a phenomenon increases the likelihood of the phenomenon, given that the mechanisms themselves are not impossible and do not counteract each other.

[15] I suspect that the argument is extensible to include at least some of the standard examples from geographical economics (Kuorikoski et al., 2010) and population biology (cf. Levins, 1968, cited from Odenbaugh and Alexandrova (2011)), but I am not defending such a claim here.

[16] Following a common shorthand, I use capital letters both to denote an entity or process in the world and the proposition asserting its existence in the actual world.

[17] The number of possible alternatives is assumed to be finite. For a defense of this assumption, cf. Dawid et al. (2015).

[18] See Nachbar (1998) for a discussion.

Tractability assumptions, on the other hand, are known not to be true for the target system and supposed to not contribute anything to the explanation. Therefore, unlike robustness against variation in the core mechanism, replacing the former type of assumptions cannot provide confirmation the same way showing robustness against substantial assumpations can.

Another crucial assumption is constituted by the two conditions mentioned in passing, namely that the mechanism is actually not known not to be instantiated and that the various mechanisms do not counteract each other. Let me start with the former.

In the case of ABM, it is often plausible that in different social systems in the space of potential target systems, different mechanisms are at work. There is, to the best of contemporary knowledge, no universal law precisely governing the social exchange of opinion. In fact, there is a more general skepticism about universality in social systems (Gallegati et al., 2006), blocking arguments common to discourse in physics (Dardashti et al., 2015).

Therefore, the assumption that a certain social group is operating under a certain mechanism usually cannot be known to be false in advance of inspecting *this particular group*. Therefore, it is plausible that numerous mechanisms should be attributed positive probability.

The assumption of non-counteracting processes is more difficult to establish. To be precise, the assumption is that the mechanisms do not counteract each other such as to render the phenomenon R at most as likely as by one of the mechanisms operating alone. Put simple, adding another operating process has to make the outcome to be confirmed more likely. This assumption can be circumvented by additional assumptions on the likelihood of the processes to operate, but obtaining such likelihoods plausibly requires empirical means.

If this is actually a sound argument, why doesn't non-empirical confirmation play a bigger role in science? There are several possible answers. First, the probabilities to be gained may very well be tiny compared to the gains by empirical investigation. Actually inspecting the opinions of the villagers and finding a pattern of polarization seems to provide a lot more confirmation than enumerating mechanisms; the pragmatic scientific rationale of such an enumeration is hypothesis generation, not confirmation.

Second, there simply might not be a substantial number of possible mechanisms known. Theoretical ABM are, due to the ease of manipulation and simulation, particularly capable of inflating the number of possible explanations. However, the results of this practice point to a third, and the most important reason that non-empirical confirmation plays little to no role in the acquisition of scientific knowledge.

In ABM, for each result establishing a certain phenomenon using a particular model, one can usually find at least one response, be that another model, a variation of the original, or merely a further exploration of the full parameters space, showing that the result is only valid for a narrow parameter range or that mechanism M^* actually generates quite different results R^*. As any known mechanism M_i generating R increases the likelihood that R actually occurs, every mechanism M_j generating $R^* \implies \neg R$ decreases this likelihood. Therefore, the actual degree of confirmation is not only diminished substantially, it also waxes and wanes dramatically in an area where new models are constructed as easily as

in ABM.

It is also helpful to contrast these considerations with a different, but related argument for non-empirical confirmation, the so-called no-alternatives argument (Dawid et al., 2015). If certain conditions on the number of possible alternatives, the difficulty of the problem and a few technical requirements are satisfied, according to the no-alternatives-argument, the fact that scientists are unable to identify equally empirically adequate alternative theories confirms the existing theories.

The argument shares the basic tenet that non-empirical facts can provide confirmation, but is structurally very different. The argument is set before the background of physics, which has a successful history of narrowing down the set of empirically adequate candidate theories. The argument from multirealizability or robustness instead has to be understood in the light of scientific practice in the social sciences, where theories rarely are rejected, and even basic empirical facts can be controversial.[19] What is actually to be confirmed is a proposition on the prevalence of a phenomenon.

Furthermore, the no-alternative-argument departs from a point of highly satisfactory theories, while the baseline of the argument from robustness is actually not to have any mechanism at hand explaining a phenomenon, and therefore rejecting its existence when empirical evidence is lacking or ambiguous. But both arguments are built on the idea that the existence of an alternative theory or mechanism may impact the degree of confirmation. In the case of the no-alternative-argument, the absence of alternatives to explain an *established* phenomenon supports existing theory; in the case of robust realizability, the independence from a particular theoretical mechanism confirms the existence of an uncertain phenomenon.

5.3.2 The Need for Fragility

The debate on robustness is generally centered around the question which epistemic advantages robustness confers once it has been established. But at times, the flip side of this debate surfaces, as when Kuorikoski et al. (2010) point out that outcomes *should* be sensitive to changes in the Galilean idealizations. By Galilean idealizations, they refer to idealizations which remove the influence of causal factors that are believed to impact the systems behavior in the actual world.[20] The point of such idealizations generally is to investigate a particular causal mechanism. Disregarding friction in physics, migration and mutation in population biology or transaction cost in markets are common examples of Galilean idealizations.

Opposed to these are mere mathematical or technical assumptions that ensure tractability. Choosing a certain class of utility functions, assuming infinite populations, and deciding on a particular update schedule in ABM are common examples of such assumptions. They are, according to the prevailing view, insubstantial and should therefore not impact the results. Robustness against such assumptions is seen as evidence that the outcome is a

[19] Compare my previous references to the polarization debate in political science.

[20] For another classification of idealizations, cf. **?**.

product of the substantial assumptions at work.

This picture of the adequate extent of fragility is, however, too limited in the case of ABM.[21] There are two major problems: (1) Fragility against variation in or relaxation of Galilean idealizations is not always desirable. (2) In some cases of prima facie arbitrary technical assumptions, fragility *is* desirable. Let me start by elaborating the former claim.

Take, once again, the example of polarization of opinions. Empirical investigations support the hypothesis that polarization is a pervasive phenomenon.[22] But at the same time, it is very plausible that a variety of mechanisms is operating in opinion dynamics. Therefore, relaxing Galilean idealizations in models of opinion dynamics, such as the independence of beliefs in different propositions, should be expected *not* to threaten the prevalence of polarization in all cases,

By definition, relaxing a Galilean idealization should have an impact on system behavior; it is possible for different effects to cancel out, in particular when relaxing or varying multiple idealizations at once, but more commonly there should be an observable impact on system behavior. Adding friction to a frictionless plane should change the behavior of the ball rolling down. Up to this point, the received view that model behavior should not be robust against changes in Galilean idealizations holds up.

However, when switching from the exact quantitative picture to a more qualitative one, there is no compelling support for the same claim. First of all, while there are quantitative measures of polarization that could be evaluated precisely as to what a certain variation in idealizations effectuates, polarization is not uniquely defined by one particular measure. If different measures agree that a social group is somewhat polarized in opinion, but disagree about the precise degree, a change in idealizations can end up in the same measures still pointing to polarization, just to a different degree. In particular, one measure may judge polarization to be more severe under idealization I, another measure under the alternative set of idealizing assumptions I^*.

This points to a second reason why phenomenological fragility against variations in Galilean idealizations is not generally to be expected. Phenomena are often operationalizations of fairly abstract concepts such as, in the running example, opinion polarization; but the same holds true for the division into center and periphery in geographical economics – the example Kuorikoski et al. (2010) utilitze. The reason that one should expect such higher-level phenomena to be robust not only against technical, but also substantial assumptions, is their multirealizability in the actual world.

The social world is, just as agent-based models representing it, highly varied in the details of its organizations. Nevertheless, social scientists and philosophers observe certain relatively universal phenomena: Social norms are established and maintained, certain patterns of opinion in a communicating group repeat across conditions, market failures occur across time and geographical location as long as something resembling a market exists, and so forth. But the underlying organizational forms and therefore the implementing

[21] Once again, I believe that a similar argument could be put forward in other areas, but since I am not providing such an argument here, I will only commit to the more modest claim about ABM.

[22] Cf. McCright and Dunlap (2011); Fiorina and Abrams (2008); Evans (2003); Poole and Rosenthal (1984) for a differentiated discussion.

mechanisms to be modeled change drastically.

There are two plausible general forms to explain this fact. First, there are actually certain universal human features that allow the replication of said phenomena under a variety of circumstances. For the time being, I dismiss this approach. The second one claims that the same macroscopical phenomena are realizable by means of various different mechanisms. If this view is accurate, there is a correspondence to the variation of non-technical assumptions in agent-based models. If the phenomena are robust across actual mechanisms in the real world, they certainly should be against variation in the representations of those mechanisms.

But it is also not clear that model behavior should always be invariant under changes in technical assumptions. Consider the following two examples:[23] in the case of the core-periphery model of geographical economics, as mentioned before, it is necessary to assume some class of utility functions to fully specify the model. In the original version of the model, these are so-called iceberg utilities. As Kuorikoski et al. argue, economists have since provided results supporting substantial robustness against variation in this assumption, which they take to strengthen the model.

Similarly, Hegselmann and Krause (2002) point out that they tested their model to be robust against replacing a synchronous update schedule for their agents by sequential update in random order. Again, they treat this as increasing the robustness of their results in some relevant sense.

However, while the assumptions are considered technical, they could just as well be framed as Galilean idealizations. Instead of using an update ordering of any real group of agents – which is highly unlikely to be ever known – opinion dynamics models assume an update order in their algorithm. But effects such as anchoring, where agents adapt their estimates according to an initially presented anchor (Jacowitz and Kahneman, 1995), point to the actual impact of ordering on process outcomes.

Similarly, while it is a technical requirement to choose one particular class of utility functions, it might still be assumed that there is a utility function that actually models accurately human behavior in spatially differentiated markets. Hence, these assumptions could be analyzed as Galilean idealizations, too, and as a consequence, the analyst can legitimately expect outcomes generated by their model to be fragile against variation in these assumptions.

The simplest response to keep up the idea that results should be fragile only against Galilean idealizations in the light of these problems would be to argue that these assumptions have been misclassified as technical assumptions. The class of technical assumptions would then shrink drastically; the discretization of continuous equations in physics is one of the remaining examples, so the distinction would still not break down entirely. This response seems largely satisfactory, but leaves one significant point of concern.

There is a clear difference between the omission of friction on a plane and the assumption of a specific update schedule in an ABM: without friction, the physicist can still write down a completely specified model, in the sense that it can be solved – or at least attempted

[23] The example of the core-periphery model and its analysis is adopted from Kuorikoski et al. (2010).

to be solved. An ABM without an update schedule, on the other hand, is incomplete; it cannot be solved. It is not clear how fundamental this difference actually is, but at least at the surface, it provides a distinction between these assumptions.

To summarize, the claim that models should be robust against variation in merley technical assumptions can be upheld, if at some cost. However, there is no general fact of the matter whether the behavior of ABM – and potentially other types of model – should be robust or fragile against variation and relaxation in Galilean idealizations. Deciding towards what an ABM's outcomes should be fragile has, unfortunately for the philosopher, to be decided partly on empirical grounds.

5.4 ABM in the Methodological Landscape

Since the early days of agent-based modeling and simulation, its practitioners tried to situate it on the methodological map. From the early days on, agent-based models have been suggested as a great tool in the theoretician's toolbox. As Schelling (1971) famously argued, his model of residential segregation offers only a potential explanation of the phenomenon.[24] Agent-based models can undeniably support claims to possibility, which are a crucial part of the enterprise of social theory.

Beyond this basic capability, ABM like the Schelling model allow the theoretician to block certain inferences; in the case of residential segregation, an inference from the phenomenon to underlying racist sentiments. The mere existence of an alternative explanation invalidates the inference from the macroscopical observation to a certain set of individual level motives. This picture also fits very well with accounts within the social sciences: Axelrod (1997) argued that simulations are theoretical in the sense that they are built from a set of formalized assumptions; the analysis of simulations results resembles the analysis of empirical data, but it actually is an application of inferential statistics to a distribution of theorems.

Putting aside their theoretical capabilities, it has been a long standing controversy in the philosophy of science to which degree computer simulations are also epistemically similar – more precisely, similarly powerful in their ability to support inferences – to controlled experiments. The arguments often fail to transfer from the general case of simulations to that of ABM, so I will skip large parts of the debate and turn immediately to this particular case.[25]

The appropriate category for comparison in the case of ABM are human subject laboratory experiments in the social sciences. Such experiments are most often employed in psychology and economics, but there is also an increasing utilization in sociology, and certainly on sociological issues.[26] There are, at the surface, obvious differences between

[24]He actually claims explicitly that legal and economic inequality are indispensable for an actual explanation of residential segregation in the United States of his day (cf. Schelling, 1971, p. 144–145).

[25]I refer the reader to (Saam, 2015) for an overview of the debate from the viewpoint of social science.

[26]Paradigmatic cases of lab experiments are reported, to give just a minimal, but quite representative sample from various disciplines, in the studies by Milgram (1963), Tversky and Kahneman (1981) or

agent-based simulation and experiment.[27]

But on closer inspection, most of these differences vanish and only one main epistemic difference remains: experiments, unlike ABM, give immediate rise to the claim that the observed behavior *could* also be observed outside the laboratory, that the investigated mechanism *could* also impact behavior in uncontrolled environments. This inference can be called an *inference to implementability*.

My argument will proceed from an account of the epistemically irrelevant differences between simulation and experiment to establishing and investigating the remaining distinction and its impact on the interpretation of results generated from ABM.

5.4.1 Narrowing the Gap

Since there is no uncontroversial account of the epistemic power of computer simulations (CS), it is difficult to establish an unassailable argument on its relationship to experiments. To arrive at an account that is robust against even strong variations in basic accounts of the epistemology of CS, I shall sketch two accounts which at the surface lead to contradicting conclusions on the relationship between simulation and experiment. The focus will be the exact remainder in difference between those two accounts, since this difference, as it turns out, can be identified with the difference between human subject experiment and ABM.

The most thoroughgoing argument to establish a fundamental epistemic difference between laboratory experiment and computer simulation is Beisbart's argument view (Beisbart, 2012). According to Beisbart, every CS can be reconstructed as a deductive argument, and its epistemic power reduces to that of the argument. Another way to put the difference is that relative to an experiment, CS are "overcontrolled" (cf. Beisbart, 2011, Sec. 4.5.1), meaning that the outcomes depend solely on the assumptions and implementation details going into the simulation model and its implementation as a computational model.

In terms of causation, experiments in the social sciences assume at least one causal pathway relevant to the experimental outcome running through a human agent. In a computer simulation representing the same process, everything is internal to the model's implementation, and could, at least in principle, be derived without simulation. This is an accurate observation in principle, but misses the point when transferred from the original application to rather simple physical systems to human subject experiments and ABM. I remain agnostic as to whether it is actually useful for more complicated physical systems.

First, the argument view fails to account for the opacity of computer simulations [28]. The complexity of computer simulations hinders the scientist from effectively overcontrolling the simulation; this is why results obtained from agent-based simulation are so often surprising

Diekmann et al. (2015).

[27] When I refer to experiments in this paper, it always means human subject controlled experiment. I am aware that there are other experimental methods employed in the social sciences, such as field experiments, and I believe that their epistemology differs more substantially from that of ABM.

[28] Cf. the discussion in Section 5.2. Note that the argument there was not to dismiss opacity, but only its impact on understanding.

to the modelers themselves. As a consequence, the overcontrol argument, though plausible in the abstract, fails to delimit simulations from experiments.

Second, the object of investigation actually can exist in a digital computer in agent-based models. There is no pendulum, no acceleration or angle to be found in the machine running Beisbart's pendulum model. However, the mental states of human agents – a crucial object of interest in ABM and human subject experiments – can be reproduced by a digital computer. As human agents maintain certain beliefs by virtue of certain brain states, artificial agents can maintain the same states of belief by virtue of hardware states.[29]

Beisbart is still right that the actual processes running within the simulation are entirely determined by the modeler, but the particular difference between physical models and ABM diminish the distance between experiments and this particular subtype of simulation.

On the opposite end of the debate, Morrison (2009) utilizes the shared need for mediating models in experiments and computer simulations to narrow the gap between these two methodologies. The experimental measurement of particle spin serves as one of her main examples: it requires not only heavy experimental machinery, but it remains indirect with physical necessity. Therefore, the models connecting spin with measurable quantities are indispensable to the measurement procedure. According to Morrison, this creates a fundamental similarity to computer simulations, which are also interpreted as measurement devices in the same fashion as the models employed in the experimental measurement of spin.

This account can be labeled "assimilatory", since it narrows the apparent epistemic gap between computer and laboratory experiments, without generally closing it entirely. Morrison concludes that simulation results are measurements in the same sense as the outcome of an experiment. Note that, even if the conclusion is valid, it does not yet imply that simulations and experiments are epistemically on a par. As Guala (2002) and Winsberg (2003, cited from Frigg and Reiss (2009)) point out, the power of computational methods crucially depends on background knowledge, and so do experiments.

Morrison appears to believe that this leaves both methods equally epistemically powerful; the mere fact that experiments are executed on "the same stuff" the target system consists of cannot establish an epistemic difference, and background knowledge can be as good as or better in the case of simulations than experiments.

It is important to point out that Morrison's account is logically compatible with Beisbart's argument view. With sufficient background knowledge and technical capability, it should be possible to exert enough control in a human subject experiment to reconstruct it as an argument, too: If the scientist can set the relevant initial brain state and has an approximately formal theory of its transitions under experimental manipulation, it seems plausible that this activity is also reconstructible as an argument, employing the coupled system of scientist and human subject instead of a digital computer.

The scientist does not even have to be able to choose the initial brain state: she could

[29]Beisbart is, I believe, himself committed to the view that machines are able to entertain beliefs, since his reconstruction thesis claims that the argument reconstructing the simulation is performed by a coupled system of scientist and machine. (cf. Beisbart, 2012, pp. 419–428)

instead measure it with high precision; just as in many computer simulation studies, where the initial state is often not directly controlled but chosen at random and recorded.

So as it turns out, the difference between plausible assimilatory views and accounts emphasizing the fundamental epistemic difference between simulation and experiment is way more subtle than it appears at first. I conjecture that the main difference actually concerns the background knowledge necessary to transfer simulation and experimental results to knowledge about the target system.

Is there a general account of the relevant background knowledge, regardless of the type of computer simulation and subject matter discipline? While the above-mentioned accounts of simulation draw heavily on examples taken from physics, the implicit assumption is that they are transferable. Before turning to the special case of ABM and controlled experiments in the social sciences, I shall provide an indirect argument against this hypothesis.

Returning to Morrison's spin example, there appears to be an important difference between running an experiment and simulating the same process, even if either results in measurement: In the experiment, according to physical background knowledge, a feature of the same physical kind (spin) as in other physical systems is present. The justification of external validity, i.e. the transfer of outcomes to spin systems in general, depends on this physical similarity. Note that it does not rely on material similarity unqualified, but according to a specific physical background theory that attributes spin to many physical systems.

Computer simulation does not rely on an actual instantiation of the physical property of spin, and therefore establishing its external validity needs additional inferential steps that establish the relationship of representation between the model and any physical system.[30]

The same is not necessarily true for human subject experiments. As argued above, artificial agents actually are able to hold belief, i.e. the object under investigation is of the same kind – namely a multirealizable mental state. While there is still additional work to be done to establish external validity, it is much less demanding than in the physical case. Between the physical spin system and the computer model, there is a difference in kind.

I leave it to physicists and philosophers of physics to spell out the exact requirements on inferences to establish the model-target relation necessary to conclude external validity. But mental states, the fundamental building block of an ABM's target in the social sciences, are of the same kind as the representations possible in a digital computer. Therefore, the epistemic difference between the two can be narrowed down to one very particular inference relying on background knowledge.

5.4.2 Inferences to Implementability

First, consider which inferences are generally supported by laboratory experiments in the social sciences. A typical example can be found in framing experiments (Tversky and Kahneman, 1981). Subjects are split into multiple groups, with the most basic version being

[30]Parker (2009) makes a very similar point to support the idea that materiality itself is not important, but relevant similarity, which can, but does not have to be, ensured by material similarity.

two. Then, they are presented with a choice between a risky and a deterministic option to choose. Scenario descriptions differ for the experimental groups, but the respective options across the frames are equivalent in terms of expected outcome; they only differ, for example, with respect to whether the deterministic outcome is letting people die or killing them.[31] Such experiments regularly result in significant framing effects, i.e. the proportion of agents choosing the risky or the deterministic option – in general probabilistically – depends on the framing in terms of killing or letting die.

What claims can be established by such experiments and their results? For starters, it does not follow that human agents are subject to framing effects regardless of context. The laboratory necessarily creates a highly artificial environment to control for the impact of anything but the investigated phenomenon, in our case framing. Standard techniques to control for interfering influences include randomization of conditions, recording certain features of the subjects and applying inference statistics category-wise and most importantly, isolating the laboratory environment from the innumerable set of events accompanying and impacting real-world choices.

However, laboratory experiments support certain claims about the mechanisms under investigation. Framing experiments establish the existence of framing effects, therefore providing potential explanations for various phenomena observable in the wild. Potential explanations, though, are precisely what agent-based models and simulations are supposed to deliver. Schelling's famous model of residential segregation according to his own account provides a potential explanation, whereas the real-world phenomenon of ethnic segregation requires a complex, multi-factorial explanation.

The shared artificiality of controlled experiments and ABM lets neither of them provide immediate actual explanations of any phenomenon found in the real world of human behavior. Both are suitable to establish possible explanations, but laboratory experiments are also able to establish certain existence claims. Precisely in those existence claims lies the remaining difference between experiment and ABM.

The inference that any laboratory experiment supports, but needs to be established separately for any ABM, is the one to implementability in the target system. The inference is in fact tautological for the experimental setting, since to perform an experiment successfully implies that whatever the relationship between the resulting phenomenon and the experimental manipulation turns out to be, can be realized in actual human agents. The same does not hold true for the computer simulation of an ABM, and in some cases, the inference to implementability is quite difficult.

In general, ABM are designed on the basis of simple agentic rules that enable an easy, though merely inductively valid, inference to implementability. Consider, once again, the Schelling model. The only behavioral rule agents follow requires them to count the frequency with which members of different ethnicities occur in their immediate neighborhood, compare the result with a certain threshold and, conditional on the result of the comparison, move to a different neighborhood. Every step in the algorithm can be easily performed

[31] This is not the place to discuss whether there is an important difference in the propositional content of the two vignettes, such as an important moral difference.

by a human agent with normal perceptual capabilities, the ability to count and to compare natural numbers.

Some real human agents lack some of these abilities, and furthermore might be limited in their choice to move or impacted by competing causal pathways determining their settling behavior, but it is difficult to deny implementability in human agents. But this fallible argument has to be provided, since the inference does not come with the construction and execution of the simulation model itself.

Not all agent-based models make it that easy to argue for implementability. Olsson and Vallinder (2013) suggest an ABM for epistemic communities engaging in individual inquiry and social exchange of reports, simultaneously updating their beliefs in a given hypothesis and the reliability of both themselves as an inquirer and their interaction partners as sources of information. The degree of belief is represented as a simple probability, but the reliability estimation requires entertaining and updating a continuous probability density.

Here, it needs to be established that the numerical representation of the continuous function is sufficiently good with respect to the model's purpose; numerical adequacy is a generic problem for discretization that could just as well arise in a physics simulation. The more interesting point is that the most common architecture for digital computers, the von-Neumann machine, is structurally very different from the human brain. Thus, the claim that the von-Neumann machine and the human brain are capable of implementing the same function or a sufficiently similar one efficiently, is less than obvious.

In summary, laboratory experiments provide the experimenter with the inference to implementability for free, whereas the simulationist has to put in additional work to draw it. How much additional work is required depends on the ABM in question. On a side note, this gives at least one argument to keep agent behavior simple to ensure a robust inductive inference to implementability.

As important as this difference is, it seems to be less than is commonly assumed about the epistemic difference between computer simulation and laboratory experiments. This assimilation is both attributing more epistemic power to agent-based simulations, and reducing the assumed power of laboratory experiments in the human sciences. However, the remaining difference seems to leave laboratory experiments as the epistemically superior method, all else being equal; but all else is very much *not* equal.

Even if simulations were epistemically strictly inferior, they would not be rendered obsolete, since there are prudential and ethical limitations to human subject experiments. Not only do they bear the potential to harm subjects, they also tend to consume immense resources. Standard ABM of opinion formation can easily be run on a modern desktop computer simulating thousands of agents, which is for all practical purposes impossible in the laboratory. In the age of large-scale online social networking platforms, field experiments on such scales have actually become possible. The degree of control necessary for a controlled experiment in the strict sense cannot be exerted in such experiments, leaving them in a different methodological category.

But the advantages of agent-based models and their simulation go beyond enabling otherwise impossible studies. Such models, unlike an experiment, explicitly state the mechanism(s) governing agent behavior. Thus, an ABM does not only establish an input-

output relationship between independent and dependent variables, it gives an immediate theoretical interpretation as long as the behavioral rules have been designed providing a thoroughgoing interpretation. In the case of an experiment, a theoretical model has to be imposed on the processes unfolding within the actual human agents. Therefore, the ABM can support a more immediate claim to *understanding*.

The experimenter may – and probably should – supply the additional inference to a theoretical mechanism, just as the simulationist should offer an argument for implementability. Unlike the inference to implementability in ABM, this kind of inference is commonly included explicitly in experimental studies, since experiments are usually set up as operationalizations of certain theory-inspired hypotheses.

Nevertheless, it is worth pointing out that any well-designed ABM, meaning in this context one with a thoroughgoing interpretation of any variable and rule included, provides the connection to theory for free. This fact is plausibly a driving reason for social scientists to employ ABM in theory exploration rather than in support of existence claims, based on the free inference preferred for the purpose.

To summarize, the epistemic capabilities of simulations of ABM versus laboratory experiments in the social sciences differ, but only in a narrow sense: agent-based simulations require background knowledge that supports inferences to implementability that come for free with laboratory experiments. On the other hand, agent-based models easily provide at least one theoretical interpretation due to their formal nature, which has to be supplanted in experimental studies. Either method requires more extensive inferential work if it is to be generalized beyond claims to existence and possibility, but the only fundamental difference is to be found in establishing these very basic types of claims.

5.5 The Adequacy of Normative Models

The philosophical use of agent-based models differs in one important respect from its application in science: Normative conclusions are to be derived by its utilization. Titelbaum (2017) explores a variety of the advantages of model building in a normative enterprise, but throughout his discussion assumes that there is a normative equivalent to descriptive facts that model outcomes fit or fail to fit. If that assumption is necessary to make sense of normative modeling, the modeler seems to be committed to a realist view of the normative realm.[32]

But on closer inspection, it turns out that normative modelers should commit to a constructivist view of the normative realm, even though they are not logically forced to adopt such a theory. Let me first state what I take to be a realist and a constructivist view of the normative:

Definition 1 (Realism) *On a realist account of normativity, normative statements are true or false, regardless of our knowledge of or even access to the normative facts thereby*

[32]Note that my claim is not itself ontological, but rather concerns the ontological commitments of a certain epistemology.

expressed.

Definition 2 (Constructivism) *On a constructivist account of normativity, the truth or falsity (or undecidability) of normative statements depends on an agent's – or population's – attitude towards them. In that sense, normative facts are constructed, usually following a procedure such as reflective equilibrium (Rawls, 1975). Hence, the construction is restricted by requirements such as rationality, and results in propositions, i.e. implies that normative assertions can be true, false or undecidable.*

It is worth pointing out that, while certainly not uniformly accepted, in particular among philosophers, a realist account of descriptive facts and their relationship to models is appealing. There, Titelbaum's description of comparing models with respect to their fit to data offers an adequate account of modeling practice. This is not to reject the claim of theory-laddeness for the realm of descriptive modeling. Which descriptive facts are considered relevant or are sought after in the first place may very well depend on the available models. As Morrison (2009) points out, models can be irreplaceable tools in accessing empirical facts. But this does not threaten the idea that there are facts regardless of our attitude towards them, or whether we even existed with our particular cognitive capacities.

The same is problematic in the normative realm for a number of reasons that metaethicists have been developing for quite some time. I focus on two points particularly relevant to normative modeling: the possibility of epistemic access to data, and pragmatics of the model construction process.

Titelbaum suggests that the relevant data for normative models could be provided, for example, by intuitions (Titelbaum, 2017).[33] But intuitions fall notoriously short of two important conditions we expect from data: interpersonal and intrapersonal persistence.

Examples of both intra- and interpersonally ephemeral intuitions are easy to be found. Nozick (1969) discusses his observation that audiences confronted with the Newcomb problem have strong, but divergent intuitions of what is rationally required.[34] Philosophical thought experiments eliciting intuitions are often faced with this challenge, and whatever constitutes their utility, it cannot rely on it producing interpersonally consistent intuitions.

For the intrapersonal part, it is helpful to look at framing effects in the so-called trolley problem. Trolley problems are a type of thought experiment, commonly analyzed in normative ethics. In one standard variant, the decision maker faces a choice between killing one person to save multiple others in one case, and letting one die to achieve the same in another (Thomson, 1976). There are many further variations of the problem, but it appears that people facing this decision problem fail to be consistent in their action-guiding

[33] I discuss only intuitions as a possible source of normative facts, since Titelbaums only alternative suggestion – well-constructed arguments – have to be built on assumptions themselves and generally cannot serve as a fixed point for model evaluation as is required by the realist view.

[34] The problem creates diverging guidance from the decision principles of dominance and expected utility maximization, and elicits intuitions on which to follow. One way of reading it may also be that neither rule tells the whole story, but I am not concerned so much with the interpretation of this particular thought experiment than the diverging intuitions it evokes.

intuitions. A common explanation to such inconsistencies of intuition are the aforementioned framing effects (Tversky and Kahneman, 1981): when essentially the same decision problem is rephrased, human agents often respond to the re-framing and switch judgment. If that is the correct explanation of the observed behavior, intuitions also fail to be intrapersonally persistent.

But to fit a model, it cannot be that one philosopher has "data" another cannot replicate; even worse, the fit of a model may depend largely on the way the modeler chose to frame the problem to himself before formalizing it. Therefore, intuitions fail to play the role of data for normative modeling, and a compelling alternative is lacking.

Note, that this argument need not even deny the possibility that intuitions might be a somewhat unreliable device of accessing normative facts. The argument is built entirely on the features of these "measurements" that make them unfit to play the role that empirical data plays for descriptive modeling.

Titelbaum touches the pragmatics of normative modeling several times, in particular when discussing the AGM framework. The aspect I am interested in concerns the way new normative principles arise in modeling; sometimes they are built-in, and serve the purpose of satisfying certain desiderata, such as logical consistency.

But there are many cases where normative principles are better understood as a consequence of the model, or introduced within the context of the model and later re-applied to the target *as already normative*. Weisberg and Muldoon (2009), for example, introduce multiple alternative success measures along the way of analyzing their agent-based model, realizing that their initial normative standard did not capture everything they found relevant.

For another example, Douven (2010) switches from one measure of accuracy to another to accommodate the possibility that one of his variables represents a probability instead of a simple scalar value. That step changes the normative standard employed within the model, and is based purely on the technical requirements of one interpretation of the model instead of another. This is not to say that anything is wrong with that practice, quite the opposite is true.

The reason normative claims within models often change during the iterative process of model construction is that the normatively valid is constructed along the way, too; or, at least, a suggestion to what should be recognized as normatively valid is made. Maybe the model is built on latent inconsistent assumptions, or it fails to satisfy other criteria of acceptance. But what would make its claims true is not adequacy to any kind of normative data, but their being part of a sound normative model.

This also captures Titelbaum's claim to model holism much better than a realist view. I agree when he says on the role of Normality in Bayesian modeling: "To a modeler, Normality was never meant to stand alone as a rational norm. Its significance, and its advantages, can be seen only in light of the other Bayesian rules, and in light of the effects it produces in concert with those rules." (cf. Titelbaum, 2017, p. 13 in the manuscript) His reasoning seems to be based on a variant of Quinean holism (Quine, 1980), arguing that only the whole of the model, not any single assumption such as Normality, can be confronted with the facts.

But as I argued before, there are no such facts available at all; the only reason to reject a normative principle based on normative modeling is the principle's failure to fit into a sound model with all the other desirable principles. Therefore, a constructivist point of view is more adequate to normative modeling. To avoid confusion, this does not imply that a modeler should be a constructivist with respect to the descriptive features of her model; the same model can be descriptively adequate – since it fits the relevant data – and normatively inadequate – since it relies on inconsistent normative assumptions – or vice versa. Being a realist about the descriptive is still possible under the presented view.

The work of normative modeling and simulation is not incompatible with a realist perspective, but it poses certain challenges to the realist in particularly sharp contour that the constructivist need not face. Let me conclude with but a brief remark on antirealism, by which I understand either the position that all normative statements are false or that they do not carry any truth value at all (the latter view is sometimes also called non-cognitiviesm).[35] In general, it seems irrational for a normative modeler to take an antirealist stance to his work.

Why engage in the process of constructing deductive arguments, if they are neither true nor false or made true by ex falso quodlibet? However, there is one interesting point I cannot currently answer: Modeling requires to idealize away many features of the world, and make all statements within the models technically false. To solve this problem, we need a more thorough understanding of what it even means to impose idealizations on normative facts than is currently available.

5.6 Conclusion

The epistemology of philosophical agent-based models, as it turns out, provides enough specific details to resolve some of debates that cannot be settled on the general level of the philosophy of computer simulations, or at least have not yet been settled.

Agent-based models, as I have shown, are defended much easier against opacity challenges than simulation models in general do. This is partly due to the simplicity of theoretical ABM, but even more importantly due to the fact that the very process of building an ABM, often as the initial formalization of a phenomenon or theoretical account, forces the modeler to ensure a thorough interpretation. As a consequence, theoretical agent-based models are an effective tool to enable scientific understanding.

Furthermore, robustness and fragility form important criteria of adequacy for agent-based models – as they do for other kinds of models, even beyond the realm of computational models. The main reason to restrict attention to ABM on these complementary concepts is to avoid certain atypical cases, such as purely mathematical simulation models, for which these criteria plausibly differ substantially.

A more substantial reason to restrict attention to agent-based models concerns the discussion on the relationship between ABM and experiments. Here, focusing on ABM allows

[35] Mackie's error theory (Mackie, 1997) and emotivist theories (Stevenson, 1996) provide typical examples from metaethics.

to restrict the comparison class of empirical research methods. Human subject laboratory experiments, as I argued, are epistemologically equivalent to agent-based simulations except for the inference to implementability. An analog argument for additional empirical methods, such as physical experiments or survey studies, seems impossible. Some simulations can be related to some empirical methods, but nothing more general seems to hold.

Finally, and not restricted to agent-based models though clearly extending to them, I touched on the issue of normativity. As I argue, an agent-based modeler deriving normative consequences has strong reasons, but is not logically forced to, take a constructivist stance on normative facts. However, the discussion of normativity in models is still in a more exploratory phase, and there is no canonical understanding of what it means, for example, to apply idealizations to normative states of affairs.

To summarize, agent-based models are able to support the kinds of arguments I made in Chapters 3 and 4: they enable an audience to understand the modeled process, their outcomes are sufficiently, but not universally robust and the necessary simplicity of agents enables inferences to implementability.

Chapter 6

Conclusion

6.1 Summary

This essay started by posing the question how to analyze an instance of social behavior suspect of collective irrationality. Inspecting environmental conditions, the social process of the group in question, and the set of applicable evaluative standards deriving from the group's ends allows an observer in principle to attribute collective irrationality.

The normative perspective provides an important extension of the empirical research available on apparently irrational collective behavior, since that line of work tends to retain a purely positivist stance. The college drinking norms discussed in Chapter 3 exemplify the problem arising from leaving normative standards implicit: the existence of an unpopular social norm is inferred from the recorded preference structure of the population. But in this account, "unpopular social norm" is reduced to a descriptive label, since no sufficient argument is given to support an actual normative inference.

Moving beyond this shortcoming is a genuinely philosophical contribution. Using a general argument scheme, the analyst need not impose their own norms implicitly to judge rationality, but can explicitly state assumptions and develop a transparent argument for collective irrationality – which may or may not be compelling to the target audience.

However, as the case studies in Chapter 2 showcase and Chapter 4 develops in more detail, there is often insufficient reason to choose a unique goal structure to judge a social process. Instead, any analysis of collective rationality has to take into account multiple criteria, and ideally, show them coinciding for the social phenomenon under scrutiny. In the case of social information aggregation, many variations of accuracy norms support the same conclusions, leaving us with a robust though not universal result.

My investigation of unpopular social norms also supports a simple, yet important insight on collective rationality: generally effective and efficient processes sometimes yield poor outcomes – e.g. unpopular social norms – and the observer needs to be careful on which level of description collective irrationality can legitimately be attributed. If the behavior in question is the norm-guided behavior of bullying school children, attributing collective irrationality seems warranted. But the underlying process of norm formation is, according

to the presented analysis, fairly effective and often leads to a high degree of successful coordination on reasonable social norms.

Furthermore, the argument for a trade-off between robust strategies and optimal strategies under ideal conditions used the argument scheme in slight variation: if two social processes (referring to the strategy mix employed) are optimally effective under plausible variation in the environment, neither of them can be regarded rational or irrational unconditionally. There is a substantial open space of social processes for which it is currently undecidable whether they are rational or irrational.

Given this account, the rational status of a process may change when new processes are suggested and investigated. If there is a process superior to the previously known ones under all circumstances considered, the previously undecidable cases are now classified as irrational. This conclusion may seem strange, but it is a consequence of relativizing the validity of an argument for irrationality to known processes, which I argued for in Chapter 2.

Finally, the discussion of the epistemology of agent-based models serves a dual purpose: for the more obvious one, it justifies the extensive utilization of ABM to construct arguments for – or against – the collective rationality of various social processes, and arguments on such processes more generally.

A bit more subtly, it provides further reasons to take an engineering approach to philosophy: if ABM (and other problem specific models) are epistemically as powerful as I argued, philosophers can gain a great deal of insight from constructing and studying such models, even before they are transferred to the sciences and empirically investigated. Building models itself provides valuable philosophical insight.

To conclude this essay, I want to discuss the issue of interventions based on stylized models, which I largely set aside during the previous inquiry, but which seems to be a chief concern for the application of theoretical models (Section 6.2). Furthermore, Section 6.3 elaborates on a relationship briefly mentioned at several points, namely between collective rationality and morality. Finally, Section 6.4 provides an outlook on the work still ahead in resolving theoretical and practical problems of collective rationality using computational methods.

6.2 Interventions

From the central influence in norm learning to the choice of strategies in belief aggregation, I discussed a variety of interventions on social processes. This should not come as a surprise, since the structure of the argument scheme set up in Chapter 2 relies on reference to alternative processes; and interventions allow the analyst of collective rationality to construct alternative processes that are, due to their similarity, prima facie plausibly implementable.

I use the term "intervention" in a non-technical way[1] to denote any change in a process that impacts the outcome and appears realizable from the current process. This under-

[1] As opposed to, for example, Woodward and Hitchcock (2003).

standing of intervention is context-sensitive: when designing communication processes in a company, changing the entire network layout may be realizable, but for most normal social networks it is not. But this is to be expected given my original discussion of splitting up any model of the target system into environment and process.

The possible objection to be addressed here concerns the justification of actual interventions on the basis of arguments relying on heavily stylized theoretical models – namely, a significant part of the arguments presented in Chapters 3 and 4. The objection may be phrased roughly thusly: stylized theoretical models contain numerous false assumptions, and generally fail any test of quantitative fit. When deciding on interventions in real world systems, however, quantitative fit is necessary, since the decision maker needs to compare costs and benefits of at least the status quo to that of the intervention condition. More often, the status quo also needs to be weighed against alternative interventions. Since stylized theoretical models fail to enable such quantitative comparisons, they cannot justify the choice of an intervention.

There are other potential objections, but the lack of quantitative fit seems to be the clearest and most compelling. Defeating the aforementioned argument – or at least weakening its force – requires to spell out some of its implicit assumptions.

The objection's main assumption can be stated in the following way: there is a viable, that is more reliable, alternative to grounding one's decision for or against a particular intervention on a highly stylized model. It is helpful to pull this assumption apart into two components. For the assumption to hold, either of the following claims has to be true: (1) the status quo is an acceptable alternative to any risky intervention with an unreliably predictable outcome. (2) there is another method of assessing both interventions and the status quo that is, all things considered, more reliable than a stylized theoretical model.

Let me start on (1): it is very difficult to ascertain the truth of this claim without also making the second true. If the status quo is well enough understood to allow for this judgment, any method to assess the status quo may seem a way to go to assess interventions. This line of reasoning is incorrect. First of all, many methodologies, such as survey studies, allow a fairly precise investigation of the actual, but are hardly useful to support counterfactual claims, such as the argument for the superiority of a possible intervention.

Furthermore, the status quo could also be clearly and obviously bad – in our case, produce collectively irrational outcomes. A social norm of excessive violence in the classroom may, for example, warrant an intervention that is less well understood due to the urgency of the situation than the suspicion that astronomers are influenced too much by their theoretical preferences instead of available data.

Hence, scenarios are imaginable where the performance of a group, though in some respects suboptimal, is good enough to shield it from dubiously justified interventions, even if the analyst is unable to precisely quantify everything. Many fields of science could be taken as examples: they work – for example in terms of technical application – fairly well, and therefore social epistemologists should be hesitant to risk the outcomes currently provided by the process by implementing sweeping interventions based solely on stylized models. But in many cases, the status quo is not clearly good enough – in a qualitative

sense – to support refuting arguments from highly stylized models.

So let me turn to the second clause of the disjunction: there is an alternative method to justify – or reject – suggested interventions. Presumably, such a justificatory procedure would rely on some kind of empirical method, like experimentation, survey studies, observational studies and so on. But all these methods have their own shortcomings: survey studies are poor at supporting counterfactual inferences. Experiments, if the argument developed in Chapter 5 is valid, are only in a very narrow sense superior to simulation models with respect to external validity.

Furthermore, ethical considerations, resource constraints or time pressure can cut alternative methods off. In some cases, a policy maker should not wait until the results of a longitudinal survey study come in, since too much harm is done in the meantime. Or an intervention on schoolyard bullying should – in the ethical sense – not be tested experimentally unless its efficacy is already supported by another method, to avoid the unnecessary harm of a failing intervention – at least in the domain of human social interaction.

All this is necessarily vague. If there was no vagueness in the normative assessment of the status quo, one could probably dismiss a merely qualitative model. But generally, assessments of collective rationality are just not that precise. This is not to say that stylized theoretical models and computer simulations could replace other methods; neither is the implication that nothing can be justified anyway. The point is merely that in an environment where precise quantification itself is problematic, highly idealized models *can* support arguments for interventions, at the very least by complementing the empirical methods by virtue of their differential pragmatic disadvantages.

6.3 Rationality and Ethics

Even though this is not an essay on ethics nor even metaethics, it is closely related to ethical questions. In particular, an attribution of collective irrationality could serve as a motivational tool for ethical behavior: if an ethically inadmissible pattern of behavior also turns out to be irrational, it should be easier to convince agents to change their behavior. Similarly, a group realizing their collective irrationality should be motivated to switch over to a different process.

Of course, there are some additional difficulties to this motivational consideration in the case of groups, since the relationship between anyone's motivation and the group's ends is indirect, and the group cannot act without its members. Nevertheless, discovering a shortcoming in collective rationality is plausibly a reason for members of the group either to leave, to change its behavior, or to get rid of the agents who block behavioral change; to name a simple example, members of an NGO with a certain purpose failing to realize its end will, on recognizing this failure, either switch to a different NGO or try to change the processes within their organization.

But this connection is not particularly tight, and it may also work in the opposite direction, i.e. drive a collective agent to unethical behavior, if it happens to be a more effective means to their ends. This leads to a second question on the relationship between

normative ethics and collective rationality: is there actually a logical relationship between the collectively rational and the morally good on the level not of motivation, but of normative content? More precisely, can certain moral notions be reduced to rationality to sidestep the difficulties of ethical and metaethical discourse?

There is no simple response to this question, partly because it depends on one's stance in both metaethics and normative ethics. Let me explain the two points in turn.

The main metaethical problem is the relativism ensuing from identifying even a minimal ethical system with a collectively rational social process. The problem, as metaethicists are well aware, is that constraints provided by rationality requirements are generally unable to single out one ethical system, but leave us with entirely incompatible but rationally admissible options.[2]

Hence, to maintain an identity of minimal morality with collective rationality its proponent has to bite the bullet of relativism. Maybe there are societies organized efficiently and effectively but maintain a system of slavery. No matter how unappealing such a position is, it seems consistent. The argument, therefore, is rather that once our imaginary proponent buys into relativism, they have to defend their position against the usual *metaethical* objections to this position and hence, a debate purely on collective rationality cannot replace the metaethical debate.

Can at least normative ethics be taken entirely out of the picture if one accepts a logical connection between the rational and the right and good? The answer is also negative. To see this, consider just two competing accounts of utilitarianism, simplified for the purpose of this discussion. On the one hand, there is rules utilitarianism, which claims that the adequate subject of moral evaluation are sets of rules; act utilitarianism, on the other hand, argues that instead individual actions should be judged. They agree, or at least that will be assumed here, on the goal of utility maximization, but they differ on the entities in the world to judge by that standard.[3]

But the analyst of collective rationality is also forced to take a stance on the same kind of question. In the discussion on unpopular norms, the focal object of analysis is a process generating norms, and while individual actions of course take place in the course of this process, they are very much not the object of normative attribution. When inquiring aggregation of socially available information, however, the focus was very much on individual strategies, and hence their actions.

Thus, if deciding between act and rules utilitarianism is a serious problem of normative ethics, the choices an analyst of collective rationality makes for structuring her inquiry can have consequences for her stance in normative ethics. If that is correct, criticism of this choice or another translates back from normative ethics. Independent reasons developed in normative ethics to refuse judging individual acts may invalidate the ethical conclusions drawn from an account of collective rationality assuming the focality of action.

Where does this leave us on the relationship between collective rationality and ethics?

[2]A discussion of the possible coherence of Caligula's ethics provides a compelling example of the problem in Street (2010).

[3]For an introductory discussion of these views in normative ethics, cf. Sinnott-Armstrong (2015).

Collective rationality, as it turns out, cannot provide a way around having debates in moral philosophy. As appealing as the idea is that there might be some baseline ethical canon the realization of which in most societies is guaranteed by pressures to collectively rational behavior, it is mistaken for two reasons: the first is that, whether one accepts these minimal standards is still dependent on their stance in normative ethics, as I just argued. Second, as this essay's models and the large body of empirical literature they are built on suggest, collective *irrationality* can persist indefinitely, and therefore failure in rationality fails to create sufficient selective pressure to ensure even minimal morality.

Henceforth the sober conclusion has to be that the investigation of collective rationality may support ethical improvement by, for example, dissolving ethically wrong unpopular social norms; but the connection is contingent, not one of logical or conceptual necessity, and no a priori argument carries from the collectively rational to the ethically good.

6.4 Outlook

Following an engineering approach to philosophy, one naturally expects an inquiry to leave any number of open problems. If, as assumed in this essay, arguments on collective rationality need to reference the particulars of the group processes under consideration, a myriad of additional problems and phenomena remain to be analyzed.

Yet I still want to provide an outlook on how to progress from here. There are at least three projects deriving from the present inquiry: analyzing further case studies with the suggested models to extend and test their reach as theoretical models, systematizing collective goal structures and – the most general and far-reaching project – building up a modeling toolbox for social philosophy in general and social epistemology in particular. Let me expand on these projects to conclude.

The discussion of legal epistemology in Chapter 4 briefly touched on an important aspect of theoretical agent-based models: they are often to some degree generic and adaptable to further purposes. While there are clearly important differences between jury deliberation and scientific knowledge aggregation, a good model can build from their commonalities. In this sense, further case studies could strengthen understanding of the relationship between model and target as well as of the target systems themselves.

Similarly, simulating the evolution of unpopular norms is not the only plausible application of the ABM suggested in Chapter 3. Voting behavior in a context of partial observability would be another interesting application; how is the parliamentary process impacted by partial visibility of other parliamentarians' voting behavior, be it by knowledge of their political allegiance, direct communication or any other process of limited communication? Which important pieces of such voting procedures are missing in the model?

The various arguments I constructed along this essay relied on a motley crew of collective end structures.[4] From Pareto-optimality to accuracy, from maxi-min norms to

[4] I avoid the term "norms" here to avoid inappropriate associations and confusion.

utility summation, there is a wide variety of evaluative standards to judge the rationality of groups. The difficulty of bringing order to this diverse assortment has several distinct sources, two of which I want to mention.

First, part of these end structures are referencing beliefs, while others are defined on utilities, preferences and the like. Second, as I argued, many goals can be molded into different specific forms to fit a purpose. Accuracy, for example, is not a single norm, but can take a number of different shapes for a group, some of which result in incompatible judgments. As a result, it is difficult to put these goal structures in a uniform conceptual framework.

The key to the first problem may lie in an observation made in Chapter 2: for collective rationality, there is no strict differentiation between theoretical and practical rationality, between belief and action. Therefore, I conjecture that all collective goal structures are expressible in terms of actions, or more precisely, their outcomes. This conjecture parallels the analog controversial position in the case of individual rationality.[5] The second problem is rather one of structuring a wide variety of aggregation mechanisms, which plausibly can be done by a more technical than philosophical analysis of their mathematical properties.

Finally, the engineering approach to philosophy suggests building a more complete formal toolbox for social epistemology. Many of the tools are already available from a technical standpoint: network theory, formal learning theory, game theory and a number of other relevant formal frameworks are all well-developed for themselves. However, when it comes to ABM in social epistemology in particular, there is little in terms of standards, reliable conventions and methodological guidelines.[6] This might seem like a minor technical point of little philosophical impact, but the impression is deceptive.

Before social philosophers can succeed in establishing a more systematic way of building their agent-based models, they run the permanent risk of misunderstanding each other, failing to fully analyze their model, or leave any number of theoretically important features out. Building on standard models in opinion dynamics in one case, and at least on a pre-existing model with decision theoretic background as I do in Chapters 3 and 4, is a small step towards unifying philosophical ABM, but hopefully a useful precedent.

[5]For example, Dutch book arguments rely on the assumption that the rationality of beliefs has to be fundamentally grounded in the actions they entice.

[6]Note that several authors have already suggested quasi-formal protocols to reign in ABM, though at least in philosophy, so far with little success. For examples, see Grimm et al. (2010) and Richiardi et al. (2006).

Bibliography

Albert, R. and A.-L. Barabási (2002). Statistical Mechanics of Complex Networks. *Reviews of Modern Physics 74*(1), 47–97.

Alexander, J. M., J. Himmelreich, and C. Thompson (2015). Epistemic Landscapes, Optimal Search, and the Division of Cognitive Labor. *Philosophy of Science 82*(3), 424–453.

Ariely, D. (2008). *Predictably Irrational*. HarperCollins New York.

Arrow, K. J. (1950). A difficulty in the Concept of Social Welfare. *Journal of Political Economy 58*(4), 328–346.

Auspurg, K., T. Hinz, and A. Schneck (2014). Ausmaß und Risikofaktoren des *Publication Bias* in der Deutschen Soziologie. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie 66*(4), 549–573.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Axelrod, R. (1986). An Evolutionary Approach to Norms. *American Political Science Review 80*(04), 1095–1111.

Axelrod, R. (1997). Advancing the Art of Simulation in the Social Sciences. In *Simulating Social Phenomena*, pp. 21–40. Springer.

Banerjee, A. V. (1992). A Simple Model of Herd Behavior. *The Quarterly Journal of Economics 107*(3), 797–817.

Barabási, A.-L. and R. Albert (1999). Emergence of Scaling in Random Networks. *Science 286*(5439), 509–512.

Beisbart, C. (2011). *A Transformation of Normal Science. Computer Simulations from a Philosophical Perspective*. Unpublished Habilitation Thesis Technical University Dortmund.

Beisbart, C. (2012). How Can Computer Simulations Produce New Knowledge? *European Journal for Philosophy of Science 2*(3), 395–434.

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bicchieri, C. and Y. Fukui (1999). The Great Illusion: Ignorance, Informational Cascades, and the Persistence of Unpopular Norms. *Business Ethics Quarterly 9*(01), 127–155.

Bicchieri, C. and H. Mercier (2014). Norms and Beliefs: How Change Occurs. In *The Complexity of Social Norms*, pp. 37–54. Springer.

Black, D. (1948). On the Rationale of Group Decision-Making. *Journal of Political Economy 56*(1), 23–34.

Bovens, L. and S. Hartmann (2003). *Bayesian Epistemology*. Oxford University Press on Demand.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthey Weather Review 78*(1), 1–3.

Buskens, V., R. Corten, and W. Raub (2015). Social Networks. In N. Braun and N. Saam (Eds.), *Handbuch Modellbildung und Simulation in den Sozialwissenschaften*, pp. 663–687. Springer.

Centola, D., R. Willer, and M. Macy (2005). The Emperor's Dilemma: A Computational Model of Self-Enforcing Norms. *American Journal of Sociology 110*(4), 1009–1040.

Dardashti, R., S. Hartmann, K. P. Y. Thebault, and E. Winsberg (2015, December). Hawking Radiation and Analogue Experiments: A Bayesian Analysis.

Darley, J. M. and B. Latane (1968). Bystander Intervention in Emergencies: Diffusion of Responsibility. *Journal of Personality and Social Psychology 8*(4p1), 377–383.

Davidson, D., J. C. C. McKinsey, and P. Suppes (1955). Outlines of a Formal Theory of Value. *Philosophy of science 22*(2), 140–160.

Dawid, R., S. Hartmann, and J. Sprenger (2015). The No Alternatives Argument. *The British Journal for the Philosophy of Science 66*(1), 213–234.

De Millo, R. A., R. J. Upton, and A. J. Perlis (1980). Social Processes and Proofs of Theorems and Programs. *The Mathematical Intelligencer 3*(1), 31–40.

De Regt, H. W. and D. Dieks (2005). A Contextual Approach to Scientific Understanding. *Synthese 144*(1), 137–170.

Dennett, D. (2009). Intentional Systems Theory. In B. McLaughlin, A. Beckermann, and S. Walter (Eds.), *The Oxford Handbook of Philosophy of Mind*, pp. 339–350. Oxford University Press: Oxford, UK.

Diekmann, A., W. Przepiorka, and H. Rauhut (2015). Lifting the Veil of Ignorance: An Experiment on the Contagiousness of Norm Violations. *Rationality and Society 27*(3), 309–333.

Doehne, M., M. von Grundherr, and M. Schäfer (2018). Peer Influence in Bullying: The Autonomy-Enhancing Effect of Moral Competence. *Aggressive Behavior 44*(6), 591–600.

Donato, D., L. Laura, S. Leonardi, and S. Millozzi (2004). Large Scale Properties of the Webgraph. *The European Physical Journal B-Condensed Matter and Complex Systems 38*(2), 239–243.

Douven, I. (2010). Simulating Peer Disagreements. *Studies in History and Philosophy of Science Part A 41*(2), 148–157.

Eger, S. (2016). Opinion Dynamics and Wisdom Under Out-Group Discrimination. *Mathematical Social Sciences 80*, 97–107.

Elkin, L. and G. Wheeler (2018). Resolving Peer Disagreements Through Imprecise Probabilities. *Noûs 52*(2), 260–278.

Elster, J. (1990). Norms of Revenge. *Ethics 100*(4), 862–885.

Elster, J. (2000). Social Norms and Economic Theory. In *Culture and Politics*, pp. 363–380. Springer.

Epstein, J. M. (1999). Agent-Based Computational Models and Generative Social Science. *Complexity 4*(5), 41–60.

Erdös, P. and A. Rényi (1960). On the Evolution of Random Graphs. *Publ. Math. Inst. Hungar. Acad. Sci 5*, 17–61.

Evans, J. H. (2003). Have Americans' Attitudes Become More Polarized? – An Update. *Social Science Quarterly 84*(1), 71–90.

Fiorina, M. P. and S. J. Abrams (2008). Political Polarization in the American Public. *Annu. Rev. Polit. Sci. 11*, 563–588.

Frigg, R. and J. Reiss (2009). The Philosophy of Simulation: Hot New Issues or Same Old Stew? *Synthese 169*(3), 593–613.

Gallegati, M., S. Keen, T. Lux, and P. Ormerod (2006). Worrying Trends in Econophysics. *Physica A: Statistical Mechanics and its Applications 370*(1), 1–6.

Gërxhani, K. and J. Bruggeman (2015). Time Lag and Communication in Changing Unpopular Norms. *PloS One 10*(4), 1–17.

Giebel, M. E. (2015). *Livius: Ab Urbe Condita, Liber I-V*. Stuttgart: Reclam.

Gigerenzer, G. (2015). On the Supposed Evidence for Libertarian Paternalism. *Review of Philosophy and Psychology 6*(3), 361–383.

Gintis, H. (2009). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton University Press.

Grimm, V., U. Berger, D. L. DeAngelis, J. G. Polhill, J. Giske, and S. F. Railsback (2010). The ODD Protocol: A Review and First Update. *Ecological Modelling 221*(23), 2760–2768.

Guala, F. (2002). Models, Simulations, and Experiments. In L. Magnani and N. Nersesseian (Eds.), *Model-based reasoning*, pp. 59–74. Springer.

Hahn, U., C. Merdes, and M. von Sydow (2018). How good is your evidence and how would you know? *Topics in cognitive science 10*(4), 660–678.

Hegselmann, R. (2017). Thomas C. Schelling and James M. Sakoda: The Intellectual, Technical, and Social History of a Model. *Journal of Artificial Societies and Social Simulation 20*(3).

Hegselmann, R. and A. Flache (1998). Understanding Complex Social Dynamics: A Plea for Cellular Automata Based Modelling. *Journal of Artificial Societies and Social Simulation 1*(3).

Hegselmann, R., S. König, S. Kurz, C. Niemann, and J. Rambau (n.d.). Optimal Opinion Control: The Campaign Problem. *arXiv preprint arXiv:1410.8419*.

Hegselmann, R. and U. Krause (2002). Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation. *Journal of Artificial Societies and Social Simulation 5*(3).

Hegselmann, R. and U. Krause (2006). Truth and Cognitive Division of Labour: First Steps Towards a Computer Aided Social Epistemology. *Journal of Artificial Societies and Social Simulation 9*(3).

Hegselmann, R. and U. Krause (2015). Opinion Dynamics Under the Influence of Radical Groups, Charismatic Leaders, and Other Constant Signals: A Simple Unifying Model. *Networks & Heterogeneous Media 10*(3), 477–509.

Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Hobbes, T. (1984[1651]). *Leviathan, oder Stoff, Form und Gewalt eines kirchlichen und bürgerlichen Staates*. Suhrkamp.

Holman, B. and J. P. Bruner (2015). The Problem of Intransigently Biased Agents. *Philosophy of Science 82*(5), 956–968.

Humphreys, P. (2009). The Philosophical Novelty of Computer Simulation Methods. *Synthese 169*(3), 615–626.

Jacowitz, K. E. and D. Kahneman (1995). Measures of Anchoring in Estimation Tasks. *Personality and Social Psychology Bulletin 21*(11), 1161–1166.

Java, A., X. Song, T. Finin, and B. Tseng (2007). Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65. ACM.

Jeong, H., Z. Néda, and A.-L. Barabási (2003). Measuring Preferential Attachment in Evolving Networks. *EPL (Europhysics Letters) 61*(4), 567–572.

Joas, H. and W. Knöbl (2004). *Sozialtheorie*. Frankfurt am Main: Suhrkamp.

Joas, H. and W. Knöbl (2013). *Sozialtheorie: Zwanzig einführende Vorlesungen*. Suhrkamp Verlag.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

Kahneman, D. and A. Tversky (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica 47*(2), 263–291.

Kauffman, K. (1981). Prison Officers' Attitudes and Perceptions of Attitudes: A Case of Pluralistic Ignorance. *Journal of Research in Crime and Delinquency 18*(2), 272–294.

Kitcher, P. (1990). The Division of Cognitive Labor. *The Journal of Philosophy 87*(1), 5–22.

Kroneberg, C. (2014). *Frames, Scripts, and Variable Rationality: An Integrative Theory of Action*, pp. 95–123. John Wiley & Sons, Ltd.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions, 2nd enl. ed.* University of Chicago Press.

Kuhn, T. S. (1977). Objectivity, Value Judgment, and Theory Choice. In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, pp. 320–339. University of Chicago Press.

Kuorikoski, J., A. Lehtinen, and C. Marchionni (2010). Economic Modelling as Robustness Analysis. *The British Journal for the Philosophy of Science 61*(3), 541–567.

Kuran, T. and C. R. Sunstein (1999). Availability Cascades and Risk Regulation. *Stanford Law Review 51*(4), 683–768.

Lackey, J. (2010). Testimony: Acquiring Knowledge from Others. In A. Goldman and D. Whitcomb (Eds.), *Social Epistemology: An Anthology*. Oxford University Press.

Lakatos, I. (1968). Criticism and the Methodology of Scientific Research Programmes. In *Proceedings of the Aristotelian society*, Volume 69, pp. 149–186. JSTOR.

Lambert, T. A., A. S. Kahn, and K. J. Apple (2003). Pluralistic Ignorance and Hooking Up. *Journal of Sex Research 40*(2), 129–133.

Laudan, L. (2006). *Truth, Error, and Criminal Law: An Essay in Legal Epistemology*. Cambridge University Press.

Lehrer, K. and C. Wagner (1981). *Rational Consensus in Science and Society. A Philosophical and Mathematical Study*. Dordrecht: D. Reidel Publ. Co.

Lehrer, K. and C. Wagner (2012). *Rational Consensus in Science and Society: A Philosophical and Mathematical Study*, Volume 24. Springer Science & Business Media.

Leitgeb, H. and R. Pettigrew (2010a). An Objective Justification of Bayesianism I: Measuring Inaccuracy. *Philosophy of Science 77*(2), 201–235.

Leitgeb, H. and R. Pettigrew (2010b). An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy. *Philosophy of Science 77*(2), 236–272.

Levins, R. (1968). *Evolution in Changing Environments: Some Theoretical Explorations*. Number 2 in Monographs in Population Biology. Princeton University Press.

Levinstein, B. A. (2012). Leitgeb and Pettigrew On Accuracy and Updating. *Philosophy of Science 79*(3), 413–424.

List, C. and P. Pettit (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press.

Mackie, J. (1997). From: Ethics: Inventing Right and Wrong. In S. Darwall, A. Gibbard, and P. Railton (Eds.), *Moral Discourse and Practice. Some Philosophical Approaches*. Oxford University Press.

Mäs, M. (2015). Modelle sozialer Beeinflussung. In N. Braun and N. Saam (Eds.), *Handbuch Modellbildung und Simulation in den Sozialwissenschaften*, pp. 971–997. Springer.

Mäs, M. and A. Flache (2013). Differentiation Without Distancing. Explaining Bi-Polarization of Opinions without Negative Influence. *PloS one 8*(11), e74516.

Mayo-Wilson, C. (2014). Reliability of Testimonial Norms in Scientific Communities. *Synthese 191*(1), 55–78.

McCright, A. M. and R. E. Dunlap (2011). The Politicization of Climate Change and Polarization in the American Public's Views of Global Warming, 2001–2010. *The Sociological Quarterly 52*(2), 155–194.

Merdes, C. (2017). Growing Unpopular Norms. *Journal of Artificial Societies and Social Simulation 20*(3), 5.

Merdes, C. (2018, Nov). Strategy and the pursuit of truth. *Synthese*.

Milgram, S. (1963). Behavioral Study of Obedience. *The Journal of Abnormal and Social Psychology 67*(4), 371–378.

Mongin, P. and M. Pivato (2016). Social Evaluation Under Risk and Uncertainty. In M. D. Adler and M. Fleurbaey (Eds.), *The Oxford Handbook of Well-Being and Public Policy*. Oxford Handbooks Online.

Morreau, M. (2016). Arrow's Theorem. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University.

Morrison, M. (2009). Models, Measurement and Computer Simulation: The Changing Face of Experimentation. *Philosophical Studies 143*(1), 33–57.

Nachbar, J. H. (1998). The Last Word on Giffen Goods? *Economic Theory 11*(2), 403–412.

Nozick, R. (1969). Newcomb's Problem and Two Principles of Choice. In *Essays in honor of Carl G. Hempel*, pp. 114–146. Springer.

Nozick, R. (1974). *Anarchy, State and Utopia.* Basic Books.

Nussbaum, M. C. (1997). Capabilities and Human Rights. *Fordham L. Rev. 66*, 273.

Odenbaugh, J. and A. Alexandrova (2011). Buyer Beware: Robustness Analyses in Economics and Biology. *Biology & Philosophy 26*(5), 757–771.

O'Gorman, H. J. (1975). Pluralistic Ignorance and White Estimates of White Support for Racial Segregation. *Public Opinion Quarterly 39*(3), 313–330.

Olsson, E. J. and A. Vallinder (2013). Norms of Assertion and Communication in Social Networks. *Synthese 190*(13), 2557–2571.

Oreskes, N. and E. M. Conway (2011). *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming.* Bloomsbury Publishing USA.

O'Connor, C. and J. O. Weatherall (2018). Scientific polarization. *European Journal for Philosophy of Science 8*(3), 855–875.

Parker, W. S. (2009). Does Matter Really Matter? Computer Simulations, Experiments, and Materiality. *Synthese 169*(3), 483–496.

Poole, K. T. and H. Rosenthal (1984). The Polarization of American Politics. *The Journal of Politics 46*(04), 1061–1079.

Prentice, D. A. and D. T. Miller (1993). Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm. *Journal of Personality and Social Psychology 64*(2), 243–256.

Quine, W. (1980). Two dogmas of empiricism. In *From a Logical Point of View: 9 Logico-Philosophical Essays*, pp. 20–46. Harvard University Press.

Raub, W., V. Buskens, and R. Corten (2015). Social Dilemmas and Cooperation. In N. Braun and N. Saam (Eds.), *Handbuch Modellbildung und Simulation in den Sozialwissenschaften*, pp. 597–626. Springer.

Rawls, J. (1975). *Eine Theorie der Gerechtigkeit.* Suhrkamp.

Regan, H. M., M. Colyvan, and L. Markovchick-Nicholls (2006). A Formal Model for Consensus and Negotiation in Environmental Management. *Journal of Environmental Management 80*(2), 167–176.

Reutlinger, A., D. Hangleiter, and S. Hartmann (2017). Understanding (with) Toy Models. *The British Journal for the Philosophy of Science 69*(4), 1069–1099.

Richiardi, M., R. Leombruni, N. J. Saam, and M. Sonnessa (2006). A Common protocol for Agent-Based Social Simulation. *Journal of Artificial Societies and Social Simulation 9*(1), 15.

Rose, Reginald Budjuhn, H. (1997). *Die 12 Geschworenen*. Stuttgart: Reclam.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir (1991). Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. *The American Economic Review 81*(5), 1068–1095.

Saam, N. J. (2015). Simulation in den Sozialwissenschaften. In N. Braun and N. Saam (Eds.), *Handbuch Modellbildung und Simulation in den Sozialwissenschaften*, pp. 61–95. Springer.

Sakoda, J. M. (1971). The Checkerboard Model of Social Interaction. *The Journal of Mathematical Sociology 1*(1), 119–132.

Salmivalli, C., K. Lagerspetz, K. Björkqvist, K. Österman, and A. Kaukiainen (1996). Bullying as a Group Process: Participant Roles and Their Relations to Social Status Within the Group. *Aggressive Behavior 22*(1), 1–15.

Savage, L. J. (1972). *The Foundations of Statistics*. Dover Publications.

Schelling, T. C. (1971). Dynamic Models of Segregation. *Journal of Mathematical Sociology 1*(2), 143–186.

Sen, A. (1986). Social Choice Theory. *Handbook of Mathematical Economics 3*, 1073–1181.

Shamir, J. and M. Shamir (1997). Pluralistic Ignorance Across Issues and Over Time: Information Cues and Biases. *The Public Opinion Quarterly 61*(2), 227–260.

Sinnott-Armstrong, W. (2015). Consequentialism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2015 ed.). Metaphysics Research Lab, Stanford University.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.

Skyrms, B. (2014). *Evolution of the Social Contract*. Cambridge University Press.

Stevenson, C. (1996). The Emotive Meaning of Ethical Terms. In S. Darwall, A. Gibbard, and P. Railton (Eds.), *Moral Discourse and Practice*, pp. 72–82. Oxford University Press.

Street, S. (2010). What is Constructivism in Ethics and Metaethics? *Philosophy Compass 5*(5), 363–384.

Strevens, M. (2003). The Role of the Priority Rule in Science. *The Journal of Philosophy 100*(2), 55–79.

Thaler, R. and C. Sunstein (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness.* New Haven, CT: Yale University Press.

Thoma, J. (2015). The Epistemic Division of Labor Revisited. *Philosophy of Science 82*(3), 454– 2.

Thomson, J. J. (1976). Killing, Letting Die, and the Trolley Problem. *The Monist 59*, 204–217.

Titelbaum, M. (2017). Normative Modeling. In J. Horvath (Ed.), *Methods in Analytic Philosophy: A Contemporary Reader.* Bloomsbury Academic Press.

Trout, J. (2002). Scientific Explanation and the Sense of Understanding. *Philosophy of Science 69*(2), 212–233.

Tversky, A. and D. Kahneman (1981). The Framing of Decisions and the Psychology of Choice. *Science 211*(4481), 453–458.

Volterra, V. (1926). Fluctuations in the Abundance of a Species Considered Mathematically. *Nature 118*, 558–560.

Watts, D. J. and S. H. Strogatz (1998). Collective Dynamics of Small-World Networks. *Nature 393*(6684), 440–442.

Weibull, J. W. (1997). *Evolutionary Game Theory*. MIT Press.

Weisberg, M. and R. Muldoon (2009). Epistemic Landscapes and the Division of Cognitive Labor. *Philosophy of Science 76*(2), 225–252.

Willer, R., K. Kuwabara, and M. W. Macy (2009). The False Enforcement of Unpopular Norms. *American Journal of Sociology 115*(2), 451–490.

Winsberg, E. (2003). Simulated Experiments: Methodology for a Virtual World. *Philosophy of Science 70*(1), 105–125.

Winsberg, E. (2010). *Science in the Age of Computer Simulation*. University of Chicago Press.

Woodward, J. and C. Hitchcock (2003). Explanatory Generalizations, Part I: A Counterfactual Account. *Noûs 37*(1), 1–24.

Zollman, K. J. (2007). The Communication Structure of Epistemic Communities. *Philosophy of Science 74*(5), 574–587.

Acknowledgment

Many people have influenced my thoughts while writing this essay, and I can only try to mention the most important sources of advice, philosophical insight and support. First for the obvious, I owe a great deal to the supervision of Stephan Hartmann, who not only provided advice and the opportunity to discuss my ideas, but also provided me with numerous links in the social network of academic philosophy. I also want to thank my second supervisor, Kevin Zollman. Visiting him in Pittsburgh provided me with the opportunity to discuss large parts of Chapter 4 and some of the ideas in Chapter 5, which, at the time, were still developing. This personal influence accompanies the important influence of his work in simulation in philosophy and social epistemology.

Besides my supervisors, many of the great people at the Munich Center for Mathematical Philosophy discussed my ideas and offered critical comments. In particular, Gregory Wheeler helped me straighten out my understanding on Bayesian modeling and Alexander Reutlinger greatly improved my understanding of the understanding provided by simulation models in philosophy. I also benefited greatly from the conversations with Momme von Sydow, Ulrike Hahn, Malte Döhne, Jean Baccelli Simon Scheller and Rush Stewart.

Furthermore, I want to thank Nicole J. Saam and the participants of her doctoral seminar, who provided me with the opportunity to present my ideas to a sociological audience. Similarly, I want to thank the participans of the meetings of the SPP "New Frameworks of Rationality" for the opportunity to speak to a multidisciplinary audience, and in particular for asking me time and again about the attribution of collective ends.

Index