Jonas Stein

Reconciling Epistemic and Identity Diversity

Identifying pathways to better decision-making in social groups

Reconciling epistemic and identity diversity

Identifying pathways to better decision-making in social groups

Jonas Stein

Funding acknowledgment: This study is part of the research program Sustainable Cooperation – Roadmaps to Resilient Societies (SCOOP). The author is grateful to the Netherlands Organization for Scientific Research (NW) and the Dutch Ministry of Education, Culture and Science (OCW) for generously funding this research in the context of its 2027 Gravitation Program (grant number 024.003.025).

Print:	Ridderprint | www.ridderprint.nl
Cover design:	Liese Schmidt | www.lieseschmidt.online
Cover illustration:	The cover depicts an illustration of the parable of the 'blind men and the elephant': A group of blind men comes across an elephant, which none of them have seen or heard of before. Each of them touches a different part of the animal and makes an observation that is partly right but still incorrect: One touches a leg and becomes convinced the elephant is like a tree, another touches an ear and thinks the elephant is like a fan, another touches the tail and believes the elephant to be a snake. They dispute and, in some versions of the tale, part in disagreement, without finding out what the elephant truly is. In other versions, they share their knowledge and collaborate, allowing them to 'see' the whole animal. The parable reminds us of how challenging it can be to bring together the different perspectives of individuals in groups, but also the potential these perspectives have if combined well.

university of groningen

Reconciling epistemic and identity diversity

Identifying pathways to better decision-making in social groups

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. J.M.A. Scherpen
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Monday 26 January 2026 at 16.15 hours

by

Jonas David Stein

born on 21 April 1995

Supervisors
Prof. dr. A. Flache
Prof. dr. J.W. Romeijn
Prof. dr. M. Maes

Co-supervisor
Dr. V.C. Frey

Assessment Committee
Prof. dr. L. Henderson
Prof. dr. S.E. Page
Prof. dr. H. Rauhut

It was six men of Indostan, to learning much inclined,
who went to see the elephant (Though all of them were blind),
that each by observation, might satisfy his mind.

The first approached the elephant, and, happening to fall,
against his broad and sturdy side, at once began to bawl:
"God bless me! but the elephant, is nothing but a wall!"

The second feeling of the tusk, cried: "Ho! what have we here,
so very round and smooth and sharp? To me ‘tis mighty clear,
this wonder of an elephant, is very like a spear!"

The third approached the animal, and, happening to take,
the squirming trunk within his hands, "I see," quoth he,
the elephant is very like a snake!"

The fourth reached out his eager hand, and felt about the knee:
"What most this wondrous beast is like, is mighty plain," quoth he;
"Tis clear enough the elephant is very like a tree."

[…]

And so these men of Indostan disputed loud and long,
each in his own opinion, exceeding stiff and strong,
Though each was partly in the right, and all were in the wrong!

John Godfrey Saxe (1872): The Blind Men and the Elephant, I-V, VIII.

Contents

1 Synthesis **1**
1.1 Introduction 2
1.2 Definition of concepts 8
1.3 Homophilous interactions 10
1.4 Information perception 13
1.5 Methods 16
1.6 Conclusions 18
1.7 Limitations and future research 20

2 Realtime user ratings as a strategy for combatting misinformation **25**
2.1 Introduction 27
2.2 Theoretical model and expectations 29
2.3 Results 35
2.4 Discussion 37
2.5 Methods 39
2.6 Appendix 41

3 How homophily can improve collective decision-making in diverse teams **49**
3.1 Introduction 51
3.2 Model description 54
3.3 Setup of simulation experiments 60
3.4 Results 62
3.5 Discussion 69
3.6 Appendix 73

4 Perceived cognitive differences facilitate complex social learning **77**
4.1 Introduction 79
4.2 Experimental design 82
4.3 Results 88
4.4 Discussion 92
4.5 Appendix 95

5 How argumentation styles and preference perceptions affect deliberation outcomes in groups with conflicting stakes 99
5.1 Introduction 101
5.2 Motivating the model 102
5.3 Model description 106
5.4 Setup of simulation experiments 111
5.5 Results 112
5.6 Discussion 117
5.7 Conclusion 121
5.8 Appendix 122
6 Propositions 127
7 English summary 129
8 Nederlandse samenvatting 133
9 References 137
10 Acknowledgements 153
11 ICS dissertation series 157
12 Biographical information 177

Chapter 1

Synthesis

This chapter benefitted from discussions with and feedback by Andreas Flache, Vincenz Frey, Michael Mäs, and Jan-Willem Romeijn.

1.1 Introduction

A fundamental aspect of human life is that people see themselves as members of groups or collectivities, which provide a shared sense of identity (Ellemers, 2012; Trepte & Loy, 2017). Perspectives on much of recent political developments – debates around migration, women's reproductive rights, or the surge of nationalist movements – would be incomplete if they disregarded that people identify with certain social groups. Social identities shape individuals' sense of self, provide a sense of belonging, and guide their beliefs, behaviors, and roles (Hogg, 2016).

Identities are commonly associated with underlying epistemic characteristics, such as the information people possess, the beliefs they hold, and the perspectives they bring (Zmigrod, 2022). People with similar socio-demographic characteristics, such as their gender, age, or sexual identity, often share similar experiences (Peters, 2021). Research on opinion dynamics and polarization demonstrates how our ideological leaning may inform our stance on political issues (DellaPosta et al., 2015), and studies on organizations outline how our professional identity guides us on how to tackle and solve certain problems (Northcraft et al., 1995; Van Dijk et al., 2012).

Socio-cultural and technological changes of the past decades have increased the frequency and intensity of interactions between people with different identities. Migration contributes to the growing diversity of local populations, and a surge in female labor force participation has transformed the demographic composition of many workplaces. International organizations facilitate collaboration between individuals from different national contexts, and, since the dawn of the 21st century, digital communication platforms enable people from different regions to easily connect across geographical boundaries.

1.1.1 Group diversity as a double-edged sword

Scholarly research has long regarded increased interaction between people of different identities as a source of epistemic potential (Page, 2019). Research on team decision-making highlights that a variety of skills and perspectives can enhance a group's ability to generate creative and effective solutions (Carter & Phillips, 2017). When individuals with different ways of thinking come together, they are more likely to challenge assumptions and approach problems from multiple angles (Aminpour et al., 2021; Levine et al., 2014; Phillips, 2003). Similarly, findings from studies on the wisdom of crowds suggest that groups composed of diverse problem-solvers—who may not be very accurate individually but provide a wide range of assessments—outperform groups of experts (Hong & Page, 2004; Keuschnigg & Ganser, 2017). Recently, the idea that diverse perspectives can foster more accurate collective decisions has inspired the design of fact-checking features on several online social media platforms: only if users with an otherwise diverse array of viewpoints agree in their judgement, a veracity assessment is published (Meta, 2025).

However, diversity also comes with challenges. Interpersonal differences can make it challenging to build mutual trust, shared norms, integrated networks and open communication necessary to exploit diversity in the process of finding better solutions (Northcraft et al., 1995; Schimmelpfennig et al., 2022). In the worst case, divergent identities can lead to the rejection of information that could improve decisions because the source of the information is evaluated negatively (Balietti et al., 2021; Guilbeault et al., 2018). Indeed, some of the different identities discussed in public discourse bear such signs: Political identity in the US, for example, is characterized by 'affective polarization' (Iyengar et al., 2019; West & Iyengar, 2022), i.e., negative attitudes and sentiment towards opposite-partisan peers.

It hence appears that diversity is a 'double-edged sword' (Milliken & Martins, 1996; Phillips & O'Reilly, 1998). Epistemic diversity – i.e., a multitude of different perspectives, skills, and knowledge – can lead to better solutions. But when these differences are tied to group-based identities, it can become difficult for groups to fully utilize their epistemic potential. This has led researchers to try to identify ways to reap the benefits of epistemic diversity while avoiding pitfalls of identity diversity. Work on demographic faultlines in teams, for example, pointed to the pivotal role of actors who share demographic attributes with multiple subgroups and can, thus, act as bridges over the subgroup divides (Flache & Mäs, 2008; Lau & Murnighan, 1998; Mäs et al., 2013). Experimental studies in small groups suggest that demographically diverse teams should first identify underlying 'deep-level' similarities, such as shared norms and interests, to circumvent that dissimilar group members reject each other's accounts (Phillips et al., 2006; Phillips & Loyd, 2006). Finally, in larger groups such as online communities, it may simply be most effective to obfuscate political and demographic differences if members are to listen to unfamiliar stances (Guilbeault et al., 2018, 2024; van der Does et al., 2022).

1.1.2 Disciplinary approaches to studying group diversity

Different scholarly fields investigate decision-making in diverse groups from multiple angles. Organizational research, for example, has gathered much observational data on the demographic composition of work teams and their performance (Bell et al., 2011; Horwitz & Horwitz, 2007; Joshi & Roh, 2009). Sociological agent-based models of opinion dynamics have made advances in identifying how micro-behavioral interaction patterns can breed macro-level phenomena such as polarization (Deffuant et al., 2000; Degroot, 1974; Hegselmann & Krause, 2002). And formal theoretical work from the field of social epistemology closely studies the efforts of agent populations in seeking epistemically justified, truthful decisions (Assaad et al., 2023; O'Connor & Weatherall, 2018; Zollman, 2010).

While specialization of different approaches comes with the advantage of breaking down the research problem into more manageable elements, the possible

disadvantage is that insights obtained in one field of research may not hold up once neglected factors studied by the other approaches are considered. For example, by drawing mostly on observational data, organizational research focuses on factor-based explanations in which group characteristics (task diversity, socio-demographic composition) are associated with group-level outcomes (economic productivity, time spent on tasks, etc.). However, by mere statistical association alone, it can become challenging to pin down precisely which micro-level interaction patterns give rise to group-level performance (DellaPosta et al., 2015).

Sociological agent-based models of opinion dynamics, on the other hand, pay close attention to micro-level interaction patterns and the theoretical mechanisms that produce them. They excel at capturing the social dynamics that precede individual belief formation but typically study only opinion distributions within populations (Flache et al., 2017). Models of this kind often carry the implicit connotation that stable polarization is worse than convergence around a single opinion. However, they lack criteria on which opinions are preferable over others regarding a specified goal, making it difficult to draw normative conclusions about the quality of the opinion distributions at which these populations arrive.

In contrast, formal philosophical models of social epistemology prioritize normative evaluation (O'Connor et al., 2024; Solomon, 2006). They pay close attention to agents' performance with respect to pre-defined epistemic criteria and can clarify the conditions under which group deliberation promotes truth-tracking or rational belief formation. Yet, this focus may overlook the influence of non-epistemic factors – such as deviations from 'rational' deliberation and the use of cognitive biases and heuristics – that are central to human behavior.

1.1.3 An integrative agenda

In this research, I therefore take an approach that connects the theoretical perspectives of social epistemology (Assaad et al., 2023; Zollman, 2010) and opinion dynamics (Flache et al., 2017) with empirical findings identified by organizational and social psychological research (Carter & Phillips, 2017; Page, 2019). I consider information exchange in diverse communities as a complex phenomenon with linkages between group-level features, interaction dynamics, and resulting decision outcomes, within a given setting under study. To illustrate this point, I sketch a schematic overview of preliminary considerations guiding my investigations, which is presented in Figure 1.1.

First, it is important to establish the setting in which interactions take place. Outlining what type of problem groups are trying to solve provides context and sets a perimeter for the decision problem we want to study. It can also be used to derive normative criteria that later enable the researcher to evaluate the quality of the decisions groups make. For instance, the criterion for a setting in which individuals are trying to identify misinformation in online networks would be accuracy, i.e., the

degree to which users are able to correctly identify a given piece of misinformation as false. Next to the type of problem, it is worthwhile to examine what kind of communication architecture is present for individuals to exchange information. Research on opinion dynamics, for example, shows that the communication of arguments versus simple beliefs can decide over whether groups find consensus or polarize (Feliciani, 2025). In this sense, settings can help explain competing findings and outline boundary conditions of the theories being tested.

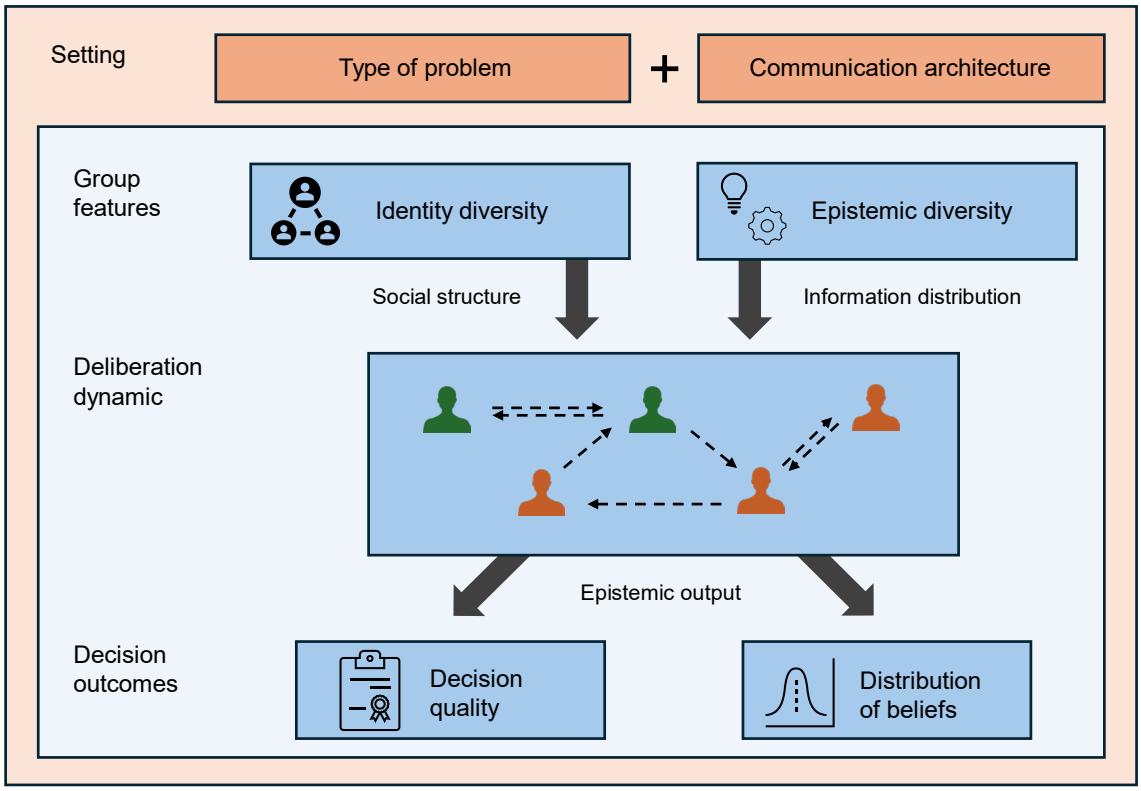


Figure 1.1 Schematic overview of an integrative approach to studying decision-making in diverse groups

Next to settings, one must consider the group-level features that impact the exchange of information. Examining what (subgroup) identities are present in the group can provide guidance in hypothesizing what type of interactions we might observe. Identities – such as those that are associated with different socio-demographic characteristics – inform the social structure of the population under study, and thereby influence how individuals interact with one another (McPherson et al., 2001). Similarly, gaining clarity on the distribution of information among group members can help theorizing how information needs to be exchanged for an effective solution: For instance, information could be distributed according to a 'hidden profile' (Stasser & Titus, 1985) such that crucial arguments are distributed across the group, or alternatively, these crucial arguments may simply reside with single actors. Here, it is likely that the former distribution would necessitate much

more intricate exchange processes than the latter, which would call for short and powerful influence processes.

Third, by keeping a close eye on the interaction dynamics within a group, we can study how micro-level mechanisms of interaction govern information exchanges between individuals. Do group members preferentially interact with others of the same identity, are they perhaps even structurally confined to do so? How do individuals select information they want to disclose, and how do they perceive and process the information they receive? And how does information spread over many encounters?

Lastly, repeated interactions can produce unanticipated epistemic output, both in terms of the quality of decisions groups make, but also in terms of the distribution of individual beliefs in the group. Self-reinforcing dynamics of social influence between decision-makers can, for example, decide whether populations develop 'swarm intelligence' or fall victim to falsehoods (Bikhchandani et al., 1992; V. Frey & van de Rijt, 2021). Similarly, repeated interactional processes of opinion exchanges and a preference to interact with like-minded others can produce irresolvable disagreement or result in consensus, depending on a number of factors, such as the role of opinion leaders, media outlets, or initial network configurations (Lipatov et al., 2025). Therefore, outcomes can be hard to predict, calling for intricate analyses of the dynamics leading up to them.

Of course, trying to cover all possible settings, group features, interaction dynamics, and macro-level consequences is beyond the scope of a dissertation project. Instead, this dissertation zooms in on specific contexts, examines selected aspects of identity diversity and epistemic diversity present in these contexts, and observes the interactional dynamics that arise from them. A last step is then to explore possible outcomes in terms of individual and collective decision-making.

To give an illustration of my approach, Chapter 2 studies bipartisan online crowds in their attempts to identify falsehoods (setting) and considers their partisan identities and affiliated ideological biases (group features). The study then examines how users influence each other as they make choices about the veracity of the informational messages they are exposed to (deliberation dynamic) and explores the extent to which users can correctly identify misinformation (decision outcomes).

Together, the dissertation represents a collection of studies analyzing information exchange in diverse populations. Its aim is to study how and under which conditions epistemic diversity in groups with different identities can lead to better decision making. What constitutes 'better' decision-making is context-dependent and defined by the evaluative criteria relevant to each problem setting under study.

1.1.4 Nuanced insights into a complex research problem

Through its focus on settings, group-level features, behavioral mechanisms, and macro-level consequences, my work follows a tradition of analytical sociology that

provides rigorous accounts of the phenomena it studies (Hedström & Bearman, 2011). It does not try to propose a holistic account or a universal theory but rather focuses on 'theories of the middle range' (Merton, 1949) that aim to explain how diversity shapes human interaction, in a given setting and with an eye on particular outcomes of interest.

Aside from following an analytical sociological agenda, I integrate perspectives from multiple fields. Through the agent-based models introduced by the following chapters, I aim to take a step towards combining the normative focus of social epistemological models with the social complexity present in models of opinion dynamics. I complement this approach with two experimental studies which draw on well-established behavioral mechanisms from social psychology - social influence, ingroup favoritism, and cognitive heuristics – and then demonstrate how these mechanisms, over repeated interactions, manifest in the decision-making of individuals and groups.

Perhaps, the most important insight generated by my studies is that diversity may not always manifest as a double-edged sword, in the sense that different identities make it challenging to integrate otherwise valuable epistemic potential. Instead, and as I will show, the picture is much more nuanced. Empirical findings from Chapter 4 show that exposure to others with a different identity can improve learning from dissimilar others. This suggests that visible diversity does not necessarily challenge the integration of unfamiliar information but can instead facilitate it. The simulation study presented in Chapter 3 leads to a similar conclusion. It shows how identity diversity in work teams, can enable better collective decisions.

However, identity diversity can also 'backfire' in larger environments of online social media users: Chapter 2 demonstrates how ideological segregation can hamper users' ability to spot falsehoods. Lastly, Chapter 5 examines how groups can reach decisions when members have identity-based, conflicting stakes. By using different argumentation styles, members can learn about their own interests and those of others, and either arrive at a consensus based on incomplete information, or part in the truthful conclusion that some interests are irreconcilable.

1.1.5 Societal relevance

The societal relevance of reconciling epistemic and identity diversity for better decision-making cannot be understated. Whether we attempt to spot falsehoods on online social media, collaborate on a project as scientists, or aim to make a shared decision in a committee we may be part of: We constantly rely on the input of others to make informed decisions in the complex and challenging environments that surround us. Different identities can make it harder to integrate this input, but they can also facilitate better decisions.

In a world where interactions between individuals of different identities are becoming ever more frequent, outlining ways to overcome these challenges and

identifying settings where diversity can be an asset rather than a hurdle is essential. Western societies are increasingly marked by affective polarization (Iyengar et al., 2019), making it more difficult to communicate across ideological boundaries. This polarization goes beyond mere disagreement—it reflects deep emotional divisions that hinder mutual understanding and trust. Public discourse is grappling with what some describe as a 'post-truth' era (Lewandowsky et al., 2017), in which different segments of society no longer agree on what counts as a fact, a valid argument, or a trustworthy source of information. For instance, the COVID-19 pandemic demonstrated how the same scientific claims were accepted by some and rejected by others, depending not just on prior beliefs but also on broader identity-related commitments (Druckman et al., 2021). Disagreements about evidence and truth are shaped by who people are and the groups they belong to. In such a climate, understanding how identity and epistemic diversity interact is more urgent than ever.

The remainder of this chapter is structured as follows. The next section clarifies two key concepts, namely, identity diversity and epistemic diversity. It is followed by two sections on central theoretical perspectives. The first theory section explains how homophily, the tendency to interact more frequently with similar others, represents a crucial factor that influences the deliberation dynamic in diverse groups. Two of my dissertation chapters examine homophilous interactions closely, which is why a summary of these chapters will be presented there. The second theory section deals with the way individuals perceive and process information from (dis-)similar others, along with an overview that locates my other two dissertation chapters in this context. A subsequent section on methods explains my choice of agent-based simulation models and experiments. I close the chapter with a section summarizing my conclusions and another section that provides an outlook for future research.

1.2 Definition of concepts

Two central concepts guide my studies, namely, identity diversity and epistemic diversity. As mentioned earlier, identity can take on many different forms and has been an influential concept for much of the social sciences (Hogg, 2006). Many of these works are related to Henri Tajfel's seminal social identity theory. Tajfel classically defined social identity as an 'individual's knowledge that he belongs to certain social groups together with some emotional and value significance to him of this group membership' (Tajfel, 1972, p. 292). I will mostly follow this definition, particularly because it is inextricably tied to a form of social group membership; but use a more specific notion: I here look at identities as stable traits that influence interactions between individuals, are plausibly correlated with individuals' knowledge, skills, and perspectives, and both salient in and relevant to the context where individuals interact. Identities satisfying these conditions could be, for example, ideological leaning in a political context, disciplinary background in a shared scientific project, or organizational membership in an inter-firm collaboration.

Identity diversity, by extension, is a group characteristic of any collectivity in which the sum of different identities is greater than one. Of course, one can also think of identity diversity as a multiplicity of traits within each individual, giving rise to questions such as their relationship to someone's knowledge, when specific self-identities become salient in certain contexts, and how they influence interactions (Ellemers et al., 2002). Such questions are relevant to study, but they are not in the focus of this dissertation. Mine is an investigation of interactions within groups, and for this reason, I will focus on settings where a group-related identity is salient. In addition, and as the term 'group' implies, I assume that in the settings I study, there are always at least two group members with the same identity. To be able to make precise theoretical predictions, and to trace these predictions in my simulations and empirical studies, I mostly limit the number of different identities within each group to two such that any group member has a clearly defined ingroup and outgroup.

I refer to epistemic characteristics, on the other hand, as those individual traits that relate to someone's knowledge, perspectives and skills (Solomon, 2006). In contrast to identities, I assume epistemic traits to be less visible and obvious, and I argue that they are in principle transferable. Say that, drawing on the example from above, epistemic characteristics represent specific arguments in favor of or against certain political decisions, field-related insights of different scientists, or entrepreneurial expertise in markets for different products. Arguments, insights, and expertise can be communicated and shared. They are, however, not visible to others from the onset. In social epistemology, epistemic characteristics are often defined as any traits that are relevant to agent's endeavors of truth-seeking (O'Connor et al., 2024). My conceptualization includes this definition, but I assume a wider concept that also encompasses searching for optimal decisions in professional contexts (Chapter 3), learning how to navigate unknown environments (Chapter 4), ideological biases (Chapter 2), and group-based preferences for different decision options (Chapter 5).

Epistemic diversity, then, refers to the diversity of knowledge, skills, and perspectives within a group. There are many ways epistemic diversity can be distributed in a population. One can, for example, imagine that a few individuals have most of the relevant knowledge, while others have very little. Or, similarly, that some group members have disproportionately less accurate evidence than others. Here, a question that would be interesting to study is how individuals should interact with each other such that relevant information is shared most effectively, and that accurate evidence crowds out inaccurate evidence. My investigations relate to this question, but assume a slightly more specific setting: namely, that subgroups within a population have different epistemic characteristics, each with their own advantages and disadvantages. The question I study then is how groups can integrate their knowledge and bring together different perspectives such that they arrive at decisions that are more accurate, truthful, or innovative.

1.3 Homophilous interactions

Identities influence individual interactions, and, as this section outlines, there are many ways in which this can facilitate or hinder the utilization of epistemic diversity. In much of the social sciences, it is a well-established finding that people are 'homophilous': People with similar socio-demographic characteristics tend to interact with each other more frequently (McPherson et al., 2001), whether this concerns friendship ties among same-gender students (Kretschmer et al., 2024; Rambaran et al., 2015), racial segregation in the US (Sander et al., 2018), ideological clustering in online social networks (Barberá et al., 2015; Levy & Razin, 2019), or sharing hobbies and interests with people of a similar age (Thomas, 2019).

A natural explanation for the widespread prevalence of homophily in human interaction is that people may simply have a preference for similar others. However, highly homophilous population states can also arise from weak individual preferences to interact with similar others. Early computational models of residential segregation show how homophily towards ethnic ingroup members can lead to sequences of individual movements that, in the long term, cause states where intergroup encounters are much less frequent than anyone would have wanted (Sakoda, 1971; Schelling, 1971). This conjecture is not limited to racial segregation alone, but extends to phenomena such as status (Fossett, 2006) and school segregation (Stoica & Flache, 2014). Similarly, DellaPosta's study on 'why liberals drink lattes' (2015) suggests that the reason why people with similar socio-demographics also have similar tastes and hobbies is not because they have latent characteristics that determine their preferences. Instead, similarities emerge endogenously through a process of homophilous interactions, weakly correlated preferences, and individuals influencing each other.

Schelling's model on residential segregation and DellaPosta's study on correlated tastes show how homophily can give rise to emergent population-level states where dissimilar individuals interact with each other muss less frequently than intuitively anticipated. Paired with the fact that homophily is a well-documented empirical regularity and a strong force in humans, this makes it relevant and interesting to study in the context of information exchange in diverse groups. If homophily can produce cultural rifts and opinion polarization to an extent where antagonistic factions barely interact with each other, can the same process undermine that subgroups with different identities share information with one another? Second, what will be the consequences for the quality of the decisions in groups where this is the case?

Potential answers can be found in organizational literature on the exploration-exploitation tradeoff (Hills et al., 2015; Lazer & Friedman, 2007; March, 1991). This literature assumes that actors – such as organizations, but also individuals – often face a choice between looking for new solutions on their own (explore) or copy the solution of someone else (exploit). A central tenet of the exploration-exploitation

canon is that on a population level, too much exploitation is often suboptimal: If everyone copies the best-performing actor, populations stop exploring the full spectrum of possible solutions and get stuck on what seems best at the moment. From this arises a tension between individual rationality – which is to cheaply copy others – and population welfare.

The exploration-exploitation literature suggests that because of the tension between individual rationality and collective welfare, excessive influence between actors must sometimes be reduced so that sufficiently many solutions are being considered. Connecting literature on the exploration-exploitation canon and research on homophily, I here explore the possibility that homophily can be a way to reduce influence. My dissertation presents two studies that explore this possibility, and I will briefly use the following paragraphs to summarize them here. Unfortunately, and as I will show, too much homophily can be detrimental to crowd wisdom in networked contexts with binary true-or-false choices. But when groups consider more complex problems with multiple decision alternatives, homophily can indeed provide the beneficial boundaries to excessive influence suggested by previous literature.

1.3.1 Chapter 2: Veracity assessments in ideologically homophilous online networks

Chapter 2 approaches homophily from a network perspective (Kossinets & Watts, 2009). It investigates how verdicts about the veracity of informational messages evolve in a setting of larger online communities in which identities are represented by ideological orientation. The chapter does not look at homophily as an individual preference. Instead, it investigates the consequences of an ideologically segregated, homophilous network structure, which can be found to varying degrees in many social networks online (Conover et al., 2011; González-Bailón et al., 2023; Stein, Keuschnigg, et al., 2023). It presents individuals with a relatively simple choice: They receive short political messages containing a factual statement. Messages are either true or false; and individuals must make a binary rating choice about the messages' veracity. Ratings are added to an aggregate veracity score. Users are exposed to the score consisting of the rating choices of others before them and add their own rating to the score shown to later users.

My simulation-based analysis and a subsequent large-scale empirical experiment reveal that segregated networks undermine users' ability to make accurate choices. This happens because in segregated communities, users who share the same ideological orientation are the first to rate messages originating within their group. Because of their ideological bias, these users tend to evaluate ideologically aligned falsehoods as true. This initiates a dynamic where early biased ratings influence subsequent users' judgments, resulting in a cascade of inaccurate evaluations. I then compare this finding to the dynamics in mixed environments, where users of different ideological camps take turns at contributing to the rating: Here, turn-taking

leads ideological biases to cancel each other out, and mixing causes a positive feedback loop that improves the overall accuracy of the rating.

The study outlines a setting in which interactions between similar individuals have detrimental epistemic consequences. It connects to research on the wisdom of crowds, which shows that under certain conditions, collective judgements can be surprisingly accurate although individual assessments are not (Galton, 1907). It also shows, however, how ideological segregation can undermine the wisdom of crowds, resonating with earlier studies suggesting that social influence paired with a non-random rating order is usually bad for the accuracy of collective verdicts (Becker et al., 2017, 2022; V. Frey & van de Rijt, 2021).

The relatively simple information exchange process in this study enabled me to closely investigate the dynamics behind the verdicts of online crowds on binary true/false choices but did not account for the complexity that can arise from discussions where groups exchange arguments in an attempt to make better decisions. Chapter 3 takes a step towards such an approach.

1.3.2 Chapter 3: Argument exchange and decision-making in homophilous teams

Chapter 3 abstracts from a networked environment and instead assumes a setting in which I simulate discussions in small teams with subgroups of different professional identities. Unlike in Chapter 2, where only aggregate rating scores were passed on from one individual to the next, I here model argument exchanges where group members learn about the viability of decision options with different quality. Arguments are distributed according to the 'hidden profile paradigm' (Stasser & Titus, 1985) such that subgroups initially favor different yet inferior decision options. This captures the idea that identity diversity − represented by subgroup membership − often comes along with epistemic diversity, i.e., different knowledge and preferences. I then study the interaction dynamics in simulated teams with varying levels of homophily, i.e., increased chances of interaction between members with the same professional identity.

In doing so, Chapter 3 investigates competing theoretical expectations concerning homophilous interactions in group decision-making settings. Sociological models of opinion dynamics show how homophilous interactions can induce irresolvable disagreement in groups and lead to stable states of bi-polarization, which can impair their capacity to make consensual collective decisions (Deffuant et al., 2000; Flache et al., 2017; Hegselmann & Krause, 2002). Epistemological models (J. Wu & O'Connor, 2021; Zollman, 2010) and research on the exploration-exploitation tradeoff (Hills et al., 2015), on the other hand, oppose this conjecture and suggest that boundaries in communication ensure that individuals explore the full spectrum of possible decisions before converging around suboptimal solutions.

Results show that homophilous environments improve the quality of collective decisions, supporting research pointing to the epistemic benefits of bounded communication (Hills et al., 2015; Zollman, 2010). In my model, this happened because preferential interactions with similar others first induced disagreement between subgroups but led to consensus on the best decision alternative in the long run. Teams with no homophily, on the other hand, often fell prey to states of suboptimal consensus in which one subgroup convinced the other of an inferior decision alternative before enough arguments in favor of the best decision were shared.

To summarize, Chapter 2 and Chapter 3 examine how homophily impacts information exchange and decision-making quality in diverse communities. They show that homophilous interactions can result in very different epistemic outcomes: In larger groups with binary true-or-false choices, interactions that happen predominantly between users of the same ideological affiliation produce inferior collective judgments. In smaller teams with many decision alternatives and more elaborate ways of exchanging arguments, homophily can foster optimal decision-making.

1.4 Information perception

Different identities can influence how often individuals interact with one another, but they can also affect how likely we are to be influenced by someone else (Feliciani et al., 2021). This section provides an overview of the latter perspective, with a particular focus on two aspects: The first aspect relates to how receptive individuals are to learning from someone else, depending on whether they share the same identity with the sender. The second aspect relates to how individuals perceive and process information, depending on their social group membership and the interests that come along with it. Chapter 4 examines the former aspect more closely, and Chapter 5 the latter. A summary of each respective aspect is included with the chapter overviews in the subsections below.

1.4.1 Chapter 4: Identity (dis-)similarity and learning

Ample research shows that social learning is subject to a similarity bias: People are more likely to be influenced by those who resemble themselves (Smaldino & Velilla, 2025). Across childhood and adulthood, people tend to copy individuals who share salient characteristics with them—such as age, gender, or ethnicity—on the assumption that similarity makes others' behavior more relevant to them (Chartrand & Lakin, 2013; Kinzler, 2021).

However, a small number of studies on group decision-making also indicate that being exposed to someone with a different identity can sometimes foster learning. In these studies, demographic diversity increased individuals' openness to unfamiliar information (Phillips, 2003; Phillips et al., 2006) because people expected that dissimilar demographics could be a reason why others have different

perspectives. Conversely, unfamiliar information from similar individuals was ignored more frequently because it was regarded as a challenge of the dominant group opinion.

Research on similarity-biased social learning suggests that shared identities enhance influence, while studies on group discussions point out that this is more so the case for different identities. In Chapter 4, I examine these conjectures, aiming to provide explanations for their apparently competing accounts. The chapter presents a social learning experiment where participants develop suboptimal approaches to a problem-solving task and are subsequently exposed to the decisions of a demonstrator. The demonstrator pursues an alternative suboptimal approach; and participants can learn to optimize their own by combining it with what they observe.

Testing competing accounts of similarity-biased learning versus research on group discussions, a 2x2 experimental design varies whether the demonstrator has shared or dissimilar characteristics, and whether (dis-)similarity relates to ideological leaning or the outcome of a fictional cognitive style test. I expect that political ideology, known for its strong connection to affective devaluation of opposite-partisan peers (Iyengar et al., 2019; L. Mason, 2018; West & Iyengar, 2022; Westfall et al., 2015), will lead to similarity-biased learning. Different 'cognitive styles,' on the other hand, should be associated with epistemic potential, increase attention paid to the demonstrator, and result in more frequent optimization.

Results show that learning and optimization increased when different identities pointed towards cognitive differences, whereas ideological leaning did not affect whether participants optimized. My findings identify dissimilarity to foster attention and integration of unfamiliar behavior, but only if this dissimilarity leads to expectations learning potential. Results challenge the common assumption that people learn best from those who are similar to them (Chartrand & Lakin, 2013; Guilbeault et al., 2018; Reyes-García et al., 2016), suggesting that perceived cognitive differences can foster learning.

1.4.2 Chapter 5: Deliberation in groups with conflicting stakes

Chapters two to four examine how identities influence how often individuals interact with one another and how (dis)similarity affects their willingness to pay attention to someone else. The final chapter shifts focus to how social group membership influences information perception after it has been received.

More precisely, in Chapter 5, I model collective deliberation in a setting where individuals perceive and evaluate arguments based on subgroup-specific preferences for different decision options. I assume that these different preferences are rooted in the divergent interests of different subgroups. However, as each member only has access to a subset of available arguments that inform their initial

views, group members are not fully aware of them from the beginning. Groups then engage in a discussion process in which, by exchanging arguments, members can develop more accurate perceptions of the decision option they prefer.

The question I study then is whether discussions lead members to develop accurate perceptions, in which case their preferences for different decision options mean that groups part in truthful disagreement. Alternatively, a deliberation dynamic emerges where enough arguments in favor of a single decision option spread such that all members become convinced that this is the option they prefer. I call this outcome an ill-informed consensus because individuals hold perceptions that run contrary to their actual preferences.

In my simulations, I vary two parameters: first, the extent to which preferences diverge. Second, I compare groups in which individuals differ in the argumentation styles by which they select arguments to share with one another. 'Advocates' represent an idealized style of agonistic deliberation (Mouffe, 1999) and preferentially select arguments that strongly support the decision option they perceive to prefer at a given point in time. 'Diplomats' on the other hand, are inspired by a Habermasian ideal of reasonable debate (Habermas, 1985a) and avoid disagreement with their interaction partner, in that they will not raise arguments that support what their interaction partner prefers the least.[1]

Results show that convergence around ill-informed consensus versus parting in truthful disagreement depends not only on a group's argumentation style, but also on how strongly preferences diverge. Low preference divergence implied that group members were less aware of their actual preferences at the start, which meant that groups of advocates - whose argumentation style was based on convincing others of what they deem best – succeeded more often in steering populations towards consensus. At high divergence, initial perceptions were much more closely aligned with the divergent preferences of opposing social groups. This facilitated consensus among groups of diplomats who, when interacting with someone holding opposing viewpoints, brought up arguments in favor of a second-best alternative rather than insisting on their preferred option.

In sum, my model reveals insights into the complex relationship between discussion outcomes and argumentation styles in situations where epistemic diversity is modeled along group-based stakes. It represents a redevelopment of previous models in social epistemology, which usually assume that complete information will result in convergence around a single decision outcome that is best

[1] Another interpretation of these two argumentation styles is perhaps that discussions in groups of diplomats represent instances of 'intergroup dialogue', where members of different identity groups strive to resolve conflict through a process of mutual understanding and relating (Dessel & Rogge, 2008). Advocate-style argumentation, on the other hand, could be seen to resemble 'red teaming', a security practice that aims to improve systems by subgroups trying to exploit weak spots in another groups' approach (Kirvan, 2025). I thank the reading committee member Scott Page for this insightful connection.

for everyone. In my model, full information resulted in the accurate conclusion that different identities come with different interests, making individuals agree that it is best to disagree. Whether such truthful disagreement is normatively better than an ill-informed consensus depends on the context in which deliberation takes place: When individuals urgently must make a decision together, any consensus may be better than the status quo. Parting in disagreement, on the other hand, may be preferable when more fitting interaction partners can be easily found elsewhere.

1.5 Methods

Two methods guide my investigations of information exchange and decision-making in diverse groups: agent-based simulation models (ABMs) and empirical experiments. I use ABMs to build a formalized account of the theoretical assumptions underlying individual exchanges. In modeling the interactions between agents that follow from these assumptions, I can then study the decisions these simulated groups make. Empirical experiments, on the other hand, provide tests of micro-level assumptions made in the models and of group-level decision outcomes.

Unlike most theory building, which is either inductive or deductive, ABMs offer a third approach: theory construction using generative models (J. M. Epstein, 2012). Researchers build theories inductively by focusing on the micro-level mechanisms that produce a certain macro-level regularity, and formulate hypotheses to falsify or distinguish between models deductively (Keijzer, 2022). Unlike formal analytical (e.g., game-theoretical) tools, which often operate with more restrictive requirements, ABMs are more flexible in that they allow for a richer set of behavioral assumptions. Through their focus on simulated individuals (e.g., 'agents') that are responsive to other individuals within a given environment, ABMs are inherently actor-based. In ABMs, agents both have a social and a cognitive structure (Gilbert & Troitzsch, 2005), which is well suited for the problem I want to study: the social structure is given by the identities of simulated actors, and their epistemic characteristics define the cognitive structure. This enables me to build environments in which I can model interactions between agents as a function of their identity, and study how this influences the information exchange process.

ABMs offer another advantage, namely, that they are fitting and well-established theoretical tools for studying complex systems (Flache et al., 2022). Repeated influence between actors can give rise to macro-level phenomena that can be hard to anticipate through intuitive reasoning alone. In simulating these processes, ABMs can provide alternatives to factor-based explanations (DellaPosta et al., 2015). Through their ability to simulate emergent phenomena 'from the bottom up', ABMs can also be used to compare competing intuitions. Different theoretical accounts sometimes assume similar mechanisms of interaction on the micro-level but propose divergent consequences on the macro-level (Bianchi & Squazzoni, 2015; Feliciani, 2025; Marchi & Page, 2014). With ABMs, researchers can build a

formalized account of micro-level mechanisms and shed light on the settings and scope conditions under which different outcomes occur. In Chapter 3, I take such an approach and investigate how homophily impacts decision-making quality in small teams. Decisions benefitted from homophily in settings where argument distributions made the problem harder to solve, supporting social epistemological accounts of transient diversity (Zollman, 2010). Additional analyses presented in the same chapter then show that solving simpler problems, on the other hand, was hindered by homophily, which is closer to propositions that follow from models of opinion dynamics (Flache et al., 2017).

It is important to make clear that agent-based models are not empirical data, they are a theoretical tool (Macy & Willer, 2002). ABMs do not capture what is but provide explanations of how something can come about. Therefore, I complement my theoretical inquiries with two types of empirical experiments, each with their own reasons for choosing them: First, I employ a micro-level experiment in which I measure participants' responses to stimuli provided by the experimenter (M. Jackson & Cox, 2013). This type of experiment enables me to test theoretical assumptions about individual reactions to other individuals with different or similar identities. To this end, Chapter 4 compares how receptive individuals are to learning novel behavior from someone else, depending on whether they are exposed to someone with the same or a different identity. Importantly, learning from someone else did not involve a real person but a fictional demonstrator, which provided a clean experimental stimulus: Observed behavior was kept consistent, and only the demonstrator's identity-related characteristics were varied.

Experiments can be used to generate empirical insights not only into micro-interactional patterns, but also the macro-level consequences that result from them (Davis & Holt, 1993). This approach is prevalent in 'macro-sociological experiments' (Hedström, 2006; Hedström & Bearman, 2011).[2] Instead of only studying the reactions of individual participants to pre-determined stimuli, such experiments often investigate the interactional dynamics in groups. Macro-experiments are a powerful tool because they can capture emergent phenomena that arise from repeated interaction processes (Gërxhani & Miller, 2022), but they are also challenging to conduct: Due to the interactions between participants, individuals cannot be treated as independent units during statistical analysis. This means that comparisons must be made across groups, rendering implementation costly. Chapter 2 takes on the approach of a macro-sociological experiment and combines it with a simulation model. This made it possible to empirically validate every part of the simulation: First, I recreated the simulated setting, i.e., ideologically integrated

[2] Of course, sociology is not the first or only discipline to apply macro experiments. A non-exhaustive list includes behavioral experiments in economics (Snijder et al., 2024), group experiments in social psychology (Phillips & Loyd, 2006), or experiments on the evolution of culture and technology (Derex & Boyd, 2016).

versus segregated groups of online social media users assessing the veracity of informational messages. Second, I tested assumptions about how individuals are influenced by the assessments of previous users. Third, I observed the macro-level consequences that resulted from this process, namely, how ideological segregation could lead to self-reinforcing dynamics that undermined a groups' ability to identify falsehoods.

In both types of experiments, hypotheses can be tested with high internal validity because the experimenter has ample control over setting and input. Observational data and field experiments do not offer this possibility but are characterized by higher external validity (Tubergen, 2020): They capture the behavior of humans in much more natural environments. In this dissertation, I follow an experimental approach because of the complexity of human information exchange. Especially when communication is characterized by natural language, individuals may interpret arguments differently and researchers have limited insight into these interpretative processes. Natural language may give room for additional biases in information communication and processing (Wood, 2004) but also poses the danger that these biases become hard to disentangle. With experiments, I can focus on the theoretical mechanisms I am interested in and investigate information exchange processes with more clarity. Another drawback of observational data, and survey data in particular, is that statistical associations are vulnerable to unobserved covariates (Diekmann, 2023). In the context of my dissertation, such covariates could, for example, involve unobserved identity traits among group members, or unknown distributions of epistemic diversity, making causal claims difficult. Nevertheless, using observational data can be a valuable approach, which is why the last section of this chapter discusses possible ways to integrate my agenda with non-experimental methods.

1.6 Conclusions

The goal of this thesis is to reconcile identity diversity and epistemic diversity for better decisions in groups. As the dissertation shows, identity diversity makes group deliberation complex because it shapes opportunities for interaction and influences how individuals perceive others and the information they receive from them. This impacts the quality of the decisions groups and individuals make, oftentimes in ways that are hard to anticipate intuitively.

To keep track of different factors that influence decision-making in diverse groups, I propose an integrative research agenda that considers interactions between group-level features, deliberation dynamics, and resulting decision outcomes. The studies presented here zoom in on specific ways of interaction that are shaped by group member's identities – such as homophilous interactions between similar individuals, or different ways of perceiving information based on one's own group membership – examine their influence on the deliberation dynamic

and investigate decision outcomes. Making general statements about how identity and epistemic diversity can be 'reconciled' for better group decisions is challenging, but nevertheless, I do find a few factors that may help reap the benefits of epistemic diversity in groups, and I will outline these in the paragraphs below.

First, I demonstrate that the type of identity that is salient in a specific context matters. To this end, Chapter 4 elucidates that cognitive differences between individuals can enhance openness and influence. Exposure to someone with a different political ideology, on the other hand, deteriorated perceptions of competence and frequently resulted in ignorance. This finding connects to broader concerns about the 'post-truth era', where identity-driven cognition shapes what counts as credible information and whose claims are dismissed. Results suggest that systems design concerning settings where people can learn from each other, emphasis should be placed on characteristics that harness curiosity, rather than differences that trigger strong us-versus-them divides.

Second, Chapter 3 shows how even in situations where different identities do not necessarily come along with specific expectations about someone's epistemic qualities, identity diversity can nevertheless be beneficial to group decision quality. This is so because homophily – a tendency to interact with individuals with the same identity –can lengthen discussions in the short run but improves decision quality in the long term. Results support the notion that different identities can provide boundaries in communication that enable better decisions (Lazer & Friedman, 2007; March, 1991; Zollman, 2010). A tentative suggestion following from these results is that deliberative environments with complex decision problems should give individuals space to weigh different decision options in subgroups first instead of striving for rapid consensus.

However, caution is advised before concluding that less communication between individuals with different identities is always better. Chapter 2 demonstrates how ideological segregation in online social networks can deteriorate collective accuracy regarding binary true-or-false problems. The study outlines a setting in which ideological segregation has detrimental epistemic consequences. Importantly, however, it does not argue that ideological differences are epistemically disadvantageous per se. Rather, it shows that partisan biases can also improve collective accuracy, provided that the people holding similar biases do not form rigid clusters. This insight offers a constructive perspective for debates on misinformation: Ideological mixing can support fact-checking and error detection in online communities. To facilitate such mixing, social media algorithms could prioritize exposure to diverse perspectives rather than feeding content that aligns with pre-existing views.

Another central takeaway is that in settings where individuals' preferences correlate with their social group membership, different argumentation styles can decide over whether groups find consensus, or part in disagreement. Chapter 5

provides two important insights regarding this: Cautious argumentation styles where members avoid disagreements can foster consensus when subgroup interests strongly diverge. Persuasive argumentation that prioritizes convincing others of what one thinks is best, on the other hand, produces consensus when interests are less conflicting. These results produce a more nuanced understanding of the complex relationship between discussion outcomes and argumentation styles. They also suggest that in politicized contexts – where subgroup interests are often at loggerheads- trying to avoid disagreement can be an effective way to steer populations towards consensus.

Together, my findings underscore that there are no one-size-fits-all prescriptions for leveraging diversity in group deliberation. Instead, the consequences of identity and epistemic diversity depend on the identities at stake, the interactions they cause, and the setting in which these interactions take place. Effective system design must be sensitive to these factors. Rather than striving to eliminate identity-based differences or treat them as obstacles, designers of deliberative environments should try to create conditions under which such differences can facilitate the utilization of the epistemic resources that are often present in diverse groups. I believe that findings from this dissertation can offer guidance for designing environments that achieve this goal.

1.7 Limitations and future research

The findings of this dissertation must be interpreted with limitations in mind. While the general goal of this research is to study how identity diversity and epistemic diversity shape decision-making in groups, the breadth and complexity of this topic necessitate to single out specific aspects of group diversity and investigate their consequences for relevant decision outcomes. The approach chosen here follows a method referred to as KISS in the agent-based simulation literature (from the US Navy slogan “keep it simple, stupid”; Axelrod, 1997), in that it starts with as few theoretical ingredients as possible. I then add complexity along the way, showing how the interplay of even a limited set of factors can produce complex interactional dynamics that influence the quality of decisions individuals and groups make.

The different dissertation chapters focused on specific micro-level mechanisms – such as homophilous interactions in Chapter 2 and Chapter 3, or different ways of perceiving and processing information in Chapters 4 and Chapter 5 – but abstracted from the fact that in real-life interactions, a co-occurrence of these mechanisms may be present. Hence, a natural next step for future research would be to add complexity by investigating how a co-occurrence of different interaction-based and perception-related factors influences decision-making in groups. I undertake a step in this direction in Section 5.8 of Chapter 5, in which I add group-based homophily to the otherwise perception-centered agenda of the chapter.

Another limitation is that in my dissertation, I abstracted from vertical (or hierarchical) differences between groups and the individuals within them. Status and power differences can challenge the integration of epistemic diversity if higher-status individuals maintain disproportionate levels of influence without a corresponding epistemic justification for it (V. Frey et al., 2024). However, status hierarchies can also make discussions more efficient if status attributions correspond to competence. Research on Matthew effects investigates how stochastic processes can decouple individual epistemic performance from the rewards these individuals receive for it (Bol et al., 2018), but they have not been studied in contexts where individuals make a decision together. The stochasticity of these processes can make status allocations hard to predict. Investigating how groups make decisions in contexts where collective deliberation is coupled with social rewards is therefore an interesting future topic to study.

Alternatively, status differences can also be treated as exogenous factors rather than endogenous products of interactional dynamics. This could, for example, involve that some subgroups of individuals occupy more central positions in a network, or that arguments raised by these individuals exert a disproportionate amount of influence (cf. Grow & Flache, 2019). In my dissertation, I instead focus on settings where subgroups are different in terms of their epistemic attributes but similar in terms of their position in the social strata. My theoretical research then shows that nominal (as opposed to hierarchical) identity differences alone can give rise to deliberation dynamics that are hard to anticipate intuitively. Including status and hierarchy differences in my investigations would perhaps increase their realism. However, it would also add another layer of complexity, reduce tractability of my simulation models and involve additional assumptions about the interplay of nominal identity differences and vertical status hierarchies. Such assumptions must be made carefully and necessitate theorizing beyond the scope of this dissertation.

A general limitation of the agent-based models I present are their relatively simple numerical representation of arguments. Because agents form preferences by summing over the numerical arguments they have access to, it is easy to intuitively follow and interpret how agents arrive at their convictions. Since Bayesian models of preference updating (Assaad et al., 2023; Madsen et al., 2018) are considered more 'rational' from the agent's perspective, incorporating such approaches represents a promising direction for future modeling efforts. Likewise, recent developments in large language models (LLMs) open up possibilities for representing argument exchange using realistic human language, rather than the abstract argument structures used here (Betz, 2022; Du et al., 2023). However, despite their ability to produce coherent text, it remains uncertain whether LLMs can accurately capture human behavior in complex deliberative settings where individuals have different identities but also different knowledge that must be brought together for better collective decisions.

It is important to mention again that the theoretical agent-based models presented in this dissertation are not empirical data but formal tools. They rest on plausible assumptions about human behavior, which are grounded in empirical findings from the social and cognitive sciences. As such, their strength lies not in full descriptive accuracy but in their capacity to systematically explore the implications of these assumptions for decision-making in diverse groups. While this allows for novel theoretical insights, it also means that the validity of their conclusions depends on the realism of their underlying assumptions. Therefore, the predictions derived from these models should be tested through empirical experimentation, which Chapter 2 does. The simulation models in Chapter 3 and Chapter 5, however, do not make a link between theoretical exploration and empirical testing yet. A next step would therefore be to translate their theoretical predictions into controlled laboratory experiments. The fact that my models rest on a framework derived from experimental research, namely the 'hidden profile' paradigm (Lu et al., 2012; Stasser & Titus, 1985) should facilitate in bridging this gap.

Another way to lend external validity to my simulation models and empirical experiments stems from using observational data. While this can be a valuable effort in some contexts, the scope of my research makes it challenging to do so. Survey data on work teams (Bell et al., 2011; Horwitz & Horwitz, 2007), for instance, can reveal correlations between team diversity and productivity, but rarely includes information about the underlying interpersonal interactions that drive those outcomes. An alternative is digital trace data, such as communication records from online social media platforms (Keusch & Kreuter, 2021; Stier et al., 2020). These data sources offer detailed, time-stamped records of micro-level interactions, making them particularly useful for studying processes like opinion polarization in networked contexts. However, leveraging such data to examine how identities affect information exchange between users is often difficult: user demographics typically must be inferred algorithmically and are frequently subject to privacy restrictions. Moreover, while there is much data about users' opinions from platforms like X (formerly Twitter; Colleoni et al., 2014; Conover et al., 2011) or Facebook (Bakshy et al., 2015; Guess et al., 2019), it is far more difficult to identify contexts where these users make collective decisions – let alone evaluate the quality of those decisions.

One promising alternative to address the limitations of both survey and social media data is the use of data from novel digital platforms for participatory democracy (Grossi et al., 2024). For example, platforms such as LiquidFeedback enable structured deliberation and collective decision-making among diverse stakeholder groups – including public institutions, citizens, and private organizations – on predefined policy issues. These platforms offer a promising opportunity to study decision-making in diverse groups with both observable interactions and measurable outcomes.

In sum, and notwithstanding the limitations listed above, this dissertation advances understanding of how identity and epistemic diversity interact in shaping group decision-making. It sheds light on the mechanisms through which identity-related dynamics can either hinder or enhance the use of diverse knowledge and offers a context-sensitive account of deliberation in heterogeneous groups. Through a combination of theoretical modeling and empirical experiments, I demonstrate that identity diversity is not inherently detrimental to epistemic outcomes – on the contrary, under the right conditions, it can act as a catalyst for improved group decisions. These achievements lay the groundwork for a promising research agenda that integrates social epistemological, sociological, and social psychological work.

Chapter 2

Realtime user ratings as a strategy for combatting misinformation

This chapter is published as: Stein, J., Frey, V., & van de Rijt, A. (2023). Realtime user ratings as a strategy for combatting misinformation: an experimental study. *Scientific reports*, *13*(1), 1626.

Abstract

Because fact-checking takes time, verdicts are usually reached after a message has gone viral and interventions can have only limited effect. A new approach recently proposed in scholarship and piloted on online platforms is to harness the wisdom of the crowd by enabling recipients of an online message to attach veracity assessments to it. The intention is to allow poor initial crowd reception to temper belief in and further spread of misinformation. We study this approach by letting 4,000 subjects in 80 experimental bipartisan communities sequentially rate the veracity of informational messages. We find that in well-mixed communities, the public display of earlier veracity ratings indeed enhances the correct classification of true and false messages by subsequent users. However, crowd intelligence backfires when false information is sequentially rated in ideologically segregated communities. This happens because early raters' ideological bias, which is aligned with a message, influences later raters' assessments away from the truth. These results suggest that network segregation poses an important problem for community misinformation detection systems that must be accounted for in the design of such systems.

2.1 Introduction

Twenty percent of the time users spend consuming news on the four largest social media sites, they are looking at content linking to one of 98 websites that researchers, professional fact-checkers and journalists agree produce fake, deceptive, low-quality or hyperpartisan news (Allen et al., 2020). This figure excludes misinformation that is produced less systematically, e.g. by news sites that only occasionally err or deceive, or by users themselves. So how can the propagation of such information on social media be mitigated? Extant approaches include algorithmically aided misinformation detection (Del Vicario et al., 2019; B. Guo et al., 2019; Tacchini et al., 2017), professional fact-checking with subsequent flagging or retraction (Ecker et al., 2010; Lewandowsky et al., 2012), and crowdsourced veracity assessments (Allen et al., 2021; Kim et al., 2019; Pennycook & Rand, 2019a). An issue with these approaches is speed. Professional and crowd-based fact-checking takes time and many algorithms cannot act instantly as they must first gather behavioral data. Until a verdict is reached so that a piece of false information can be flagged or retracted, potentially false information can spread unchecked.

An emergent approach in the scientific literature on misinformation (Allen et al., 2022; Kim et al., 2019; Pretus et al., 2022; Pröllochs, 2021) as well as in practice is to harness the wisdom of the crowd by enabling recipients of an online message on a social media platform to attach veracity assessments to it. This may allow poor initial crowd reception to temper belief in and further spread of misinformation. For example, on Twitter's Birdwatch users can write notes and attach these to a Tweet, explaining why they believe it is or is not misleading. Other users can rate these notes or write additional notes in response.

The main challenge this approach must overcome is to somehow function in online environments where truth seeking is not the dominant driver of behavior, but rather personal convictions, e.g. of a political or ideological nature. Previous studies on misinformation have shown that sharing decisions regarding messages with a clear political leaning are primarily guided by users' ideological congruence with the message and only little by perceived veracity (Pennycook & Rand, 2019b; Scheufele & Krause, 2019; Vosoughi et al., 2018). Nonetheless, previous work on the wisdom of the crowd shows that also when individuals have strong individual biases of an ideological or other nature, as long as the average individual's assessment is better than random, the aggregation of judgments produces an accurate collective assessment. This work assumes that individuals in a crowd cast independent votes (Galton, 1907; Surowiecki, 2004), and it suggests that while individual judgements may not be very accurate, their average often closely approximates the truth (Allen et al., 2021; Baker, 1975; Becker et al., 2022; Condorcet, 1785). Recent experimental studies further show that when individuals do not make true-or-false decisions independently but are influenced by the decisions of those who came before them – as they would be on social media – individuals' accuracy further improves (Becker

et al., 2017; V. Frey & van de Rijt, 2021; Friedkin & Bullo, 2017; Goeree et al., 2007; Van de Rijt, 2019). This is so because as long as the average decision-maker is more accurate than random, prior decisions of others will tend to nudge decision-makers towards the truth. If a developing rating starts off with the majority of decisions being correct, this will influence subsequent users towards making the correct decision and thereby further improve the rating. It can of course happen that the same social influence dynamics facilitate the spread of a false belief, namely when the initial decisions are incorrect. Subsequent users may then be influenced to make incorrect decisions themselves, further solidifying the incorrect rating.

In previous studies on social influence and the wisdom of crowds (Becker et al., 2022; Da & Xing Huang, 2020; V. Frey & van de Rijt, 2021; Guilbeault et al., 2018; Lorenz et al., 2011), only chance could generate a large early majority favoring the wrong veracity verdict, because in these studies either all subjects first cast an independent vote and then could revise based on the first round results, or the order according to which subjects made decisions was random. In online social networks contexts, however, the order according to which subjects would cast veracity verdicts occurs along the path through which information disseminates. And herein lies the problem: Such an order is far from random. Online social networks, like most social networks, are homophilous (Bakshy et al., 2015; Barberá et al., 2015), comprising communities of predominantly like-minded peers (Boutyline & Willer, 2017; Cinelli et al., 2021; Del Vicario et al., 2016). The level of segregation rarely reaches the extremity implied by the terms 'echo chambers' or 'filter bubbles' but is nonetheless substantial (Conover et al., 2011; Eady et al., 2019; Flaxman et al., 2016; Muise et al., 2022). Different groups of online users have different levels of ability to identify misinformation (Borah, 2022; Pennycook & Rand, 2020; Scheufele & Krause, 2019; Tacchini et al., 2017), and this ability correlates with their ideological identity (Del Vicario et al., 2019) and demographic characteristics (Guess et al., 2019). Misinformation is often politically or ideologically charged, or intentionally designed to mislead only a specific part of the population (Lazer et al., 2018; Vosoughi et al., 2018), and it usually appears among and targets those clusters of users who are most susceptible to it. Hence it would then appear that ratings would have to be able to cope with misinformation initially being rated in communities of individuals who all tend to have the same biases and likely believe the misinformation or in bad faith misclassify it as true.

Our study explores real-time user ratings under such circumstances in a large-scale experiment with 2,000 liberal and 2,000 conservative subjects in 80 bipartisan groups (Figure 2.1). We implement two scenarios in which ratings are broadcast immediately after launch: First, a scenario mimicking the development of a real-time rating in an ideologically *integrated* network marked by many cross-partisan ties and no clustering according to ideology. Second, we implement a scenario representing the typical rating sequence in an ideologically *segregated* network,

with individuals whose ideological identity is aligned with the content of a message rating the message first and more critical individuals rating the message later. These scenarios represent ideal-types and maximize our treatment as extreme cases of a continuum along which more and less segregated real-world online communities are positioned (Barberá et al., 2015; Cinelli et al., 2021). We further compare these scenarios with a control condition resembling the setup in which crowd-based ratings have been studied previously (Allen et al., 2021; Pennycook & Rand, 2019a): namely, a scenario in which subjects rate messages independently and without information about the rating decisions of others.

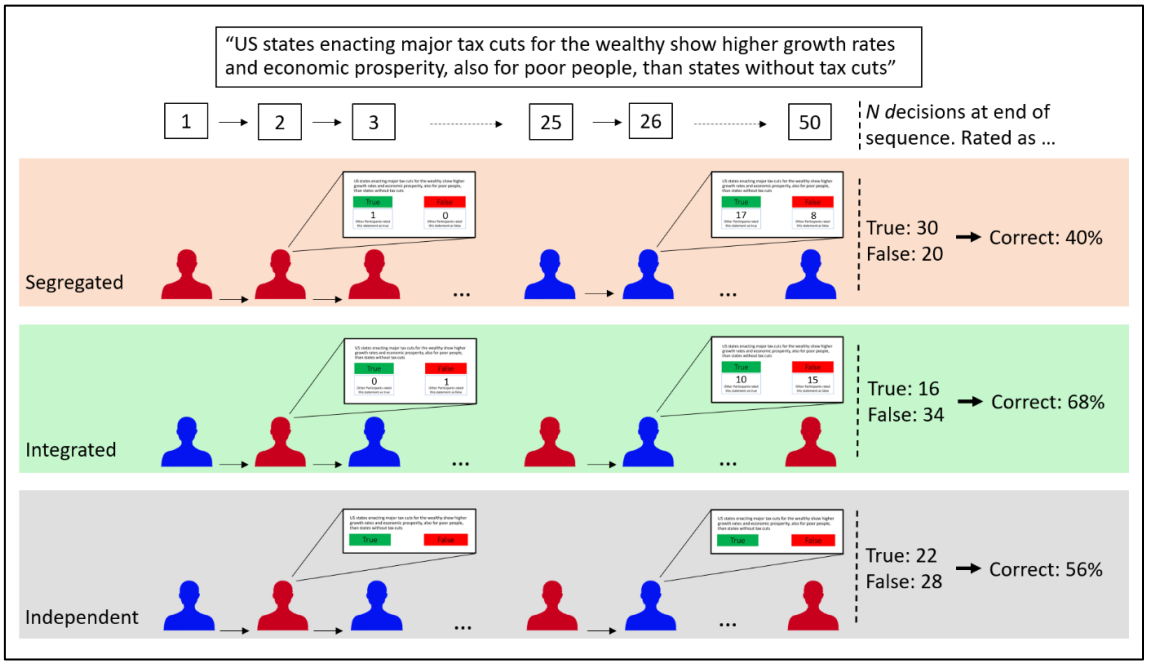


Figure 2.1 Study setup. Subjects were randomly assigned to ideologically segregated, ideologically integrated, or independent groups and instructed to rate 20 informational messages as true or false. Subjects rated messages sequentially. In segregated and integrated groups, subjects were presented a count of previous group members' rating decisions. Subjects in independent groups made rating decisions without being exposed to previous subjects' decisions. Red (blue) icons represent self-identified conservatives (liberals). The example message in the figure is false and has a conservative leaning. For each rating group, we measured how many correct and incorrect rating decisions were made (right side of the figure, data taken from exemplary groups observed in the experiment). We collected data from 80 rating groups with 50 subjects each, amounting to a total of 80,000 rating decisions.

2.2 Theoretical model and expectations

The simulation model we introduce in this section predicts that if groups are ideologically integrated, broadcasting the rating will trigger a positive feedback loop that improves individuals' capacity to differentiate between true and false messages. This happens despite strong ideological bias for or against such messages. Similarly,

if true information is rated in segregated groups, early ratings from individuals with an ideological bias in favor of the true message foster the development of a correct rating. However, broadcasting the rating backfires and reduces correct identification when false information is first rated exclusively by ideologically friendly users and only later by ideologically opposed individuals.

In our model, individuals $1 \leq i \leq n$ make a binary rating decision C_i with regard to an informational message m with veracity $v = 1$ if the message is true, or $v = -1$ if the message is false. Ratings are made sequentially. Individuals' propensity to make a correct rating decision $Prob(C_i = 1)$ is given by the following logistic function:

$$Prob(C_i = 1) = \left(1 + \frac{d_i}{1-d_i}\, e^{-s \times r_i}\right)^{-1} \tag{2.1}$$

The propensity to correctly classify is negatively impacted by how difficult it is to correctly classify a certain message. This difficulty, d_i, is the probability of incorrectly classifying a message when this is done independently, in the absence of information from others ($0 \leq d_i \leq 1$). d_i takes on the value of d_{align} for ideologically aligned individuals and d_{mis} for misaligned individuals. The difficulty terms d_{align} and d_{mis} capture ideological bias stemming from cognitive mechanisms such as motivated reasoning (Haidt, 2012; Mercier & Sperber, 2011) and confirmation bias (Nickerson, 1998): It is more difficult for aligned individuals to identify a false (aligned) message as false, but less difficult for misaligned individuals to identify a false (misaligned) message as false ($v = -1 \rightarrow d_{align} > d_{mis}$). Likewise, cognitive bias makes it less difficult for aligned individuals to find true information true, but more difficult for misaligned individuals ($v = 1 \rightarrow d_{align} < d_{mis}$). Formally, $d_{align} = \bar{d} - (b \times v)/2$ and $d_{mis} = \bar{d} + (b \times v)/2$, where $\bar{d}$ denotes the average level of difficulty in the population. As we use an equal number of aligned and misaligned individuals in each simulation as well as in the experiment we report on later, $\bar{d} = \frac{(d_{align} + d_{mis})}{2}$. The term b captures to what extent a message activates bias in individuals ($0 \leq b \leq 1$) and corresponds to the absolute difference in difficulty between aligned and unaligned individuals: $b = |\, d_{align} - d_{mis}\,|$. Individual i's propensity to correctly rate a message further depends on the previous classification decisions of others through the rating r_i, which is the average of previous decisions (Equation 2.2).

$$r_{i,\ i>1} = \frac{\sum_{j<i} c_j}{i-1} \tag{2.2}$$

r_i ranges from -1 (= all prior user classifications were incorrect) to +1 (= all prior classifications were correct). For the first individual, $i = 1$, r_i equals 0. s denotes the degree to which individuals are influenced by rating r_i. We assume positive susceptibility to the rating ($s > 0$), which implies that $Prob(C_i = 1)$ monotonically increases with r_i, and we assume that everyone is equally susceptible to social influence.

We derive hypotheses through simulation of this model. Each simulation run starts with the first individual $i = 1$, making a first rating decision with $Prob(C_i = 1) = 1 - d_i$ in the absence of prior ratings. The decision of i factors into the rating signal of the next individual $i + 1$, r_{i+1}, influencing i+1's rating decision. The simulation stops after $i = N$ has made their decision. We match population sizes of our simulations with those in the experiment ($N = 50$). Similar results are obtained for smaller and larger populations. Simulation runs are executed 10,000 times for each parameter combination of interest. The dependent variable is the fraction of correct rating decisions out of all rating decisions, computed as an average of fractions over many simulation runs. We choose a target value that reflects average performance rather than a group decision because real-time ratings do not intend to reflect a final verdict (such as a majority vote) but aim to improve raters' information detection capabilities.

We investigate the interplay of rating order, message veracity and cognitive biases in two real-time rating scenarios in which ratings are broadcast immediately: In the *segregated scenario*, a message originates and spreads in the aligned cluster so that aligned individuals sequentially rate first. The message then reaches the misaligned cluster and misaligned individuals rate it until everyone in the population has made their decision. In the *integrated scenario*, aligned and misaligned individuals alternate in making ratings. These scenarios are compared with an *independence scenario* in which choice order is alternating as well but in which the rating is not broadcast so that individuals make choices without knowledge of others' ratings (i.e., $s = 0$ implying $Prob(C_i = 1) = 1 - d_i \; \forall \; i$).

In the independence scenario, the fraction of correct rating decisions equals the inverse of the average level of difficulty in the population, i.e., $1 - \bar{d}$. In the integrated scenario, it is to be expected that more individuals will make correct rating decisions than in the independence scenario if $\bar{d} < 0.5$ and fewer if $\bar{d} > 0.5$ (Figure 2.2 A, left). Namely, if $\bar{d} < 0.5$, the first individual is more likely to make a correct rather than a false rating decision. If the first individual makes a correct rating decision, they influence the following individual to make a correct decision themselves, which enhances the accuracy of the rating for the next individual, and so forth. A real-time rating triggers a positive feedback loop for $\bar{d} < 0.5$, where each subsequent i^{th} rating has a higher probability to be correct than the previous one (compare Figure 2.2 B, left). Individual biases cancel each other out in the alternating ratings of aligned and misaligned individuals. These theoretical expectations hold for true as well as false messages equally since we assume no systematic differences in information difficulty between true and false information. A negative feedback loop, or 'backfiring', on the opposite, is expected to be triggered for $\bar{d} > 0.5$ since individuals are more likely to make incorrect rather than correct decisions. We accordingly formulate Hypothesis 1.

H1: When it is not too difficult to classify a message correctly ($\bar{d} < 0.5$), then individuals in integrated groups (with information about previous rating choices) classify true and false messages more often correctly than individuals in independent groups (without information about previous rating choices).

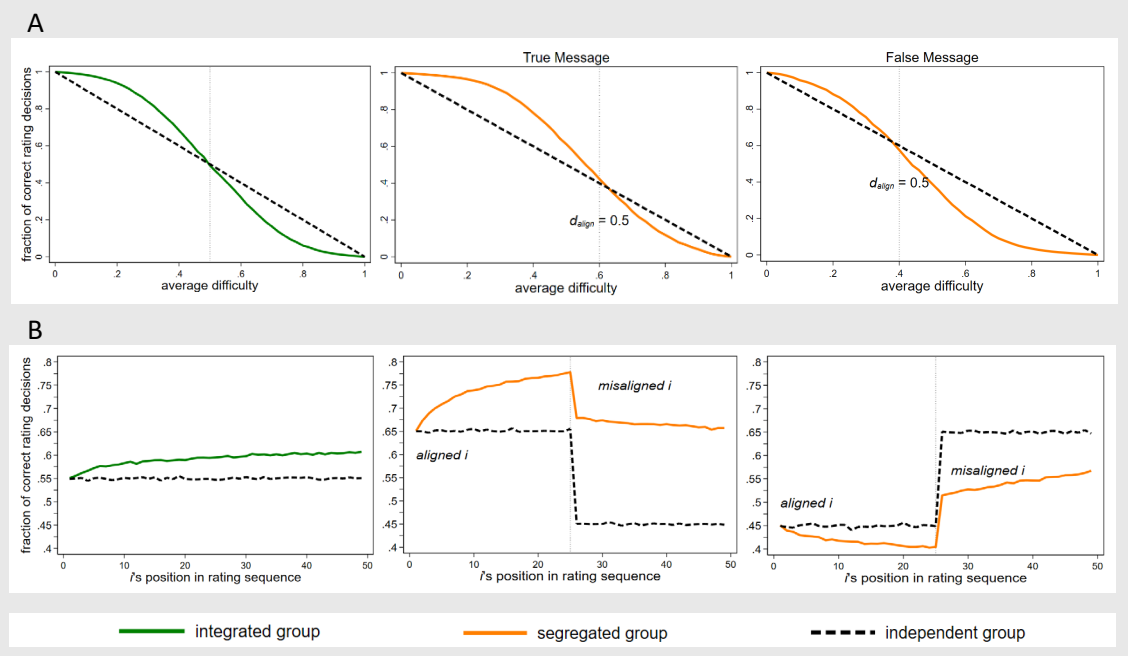


Figure 2.2 Theoretical expectations. We simulate sequences of rating decisions in 30,000 groups of bipartisan agents (25 ideologically aligned and 25 misaligned agents per group). The fraction of correct rating decisions is shown (**A**) as a function of difficulty and (**B**) as a function of an agents' position in the rating sequence. In integrated groups, agents classify messages more often correctly than those in independent groups when average difficulty $\bar{d} < 0.5$, irrespective of message veracity. In segregated groups, agents classify messages more often correctly than those in independent groups if true messages are being rated but classify messages less often correctly if false messages are rated. Parameters used in all panels: $| d_{align} - d_{mis} | = b = 0.2$, $v = [-1; 1]$, $s = [0; 2.35]$. Panel B: $\bar{d} = 0.45$.

In the segregated scenario, aligned individuals give ratings first. Since they align with the standpoint of a given message, they are more likely to correctly identify a true message as true. On the other hand, compared to misaligned individuals, they have greater difficulty identifying a false message as false. Since aligned individuals are the ones to rate first, their decisions will determine the early accuracy of the rating signal and influence later raters. If messages are true and difficulty among aligned individuals d_{align} is below 0.5, the rating is likely to enter a positive feedback loop. Later misaligned raters – although less likely to make correct rating decisions due to their bias – will make a correct decision more often than those raters without exposure to a rating signal (Figure 2.2 A, center). If messages are false and d_{align} is

instead above 0.5, early raters are likely to make incorrect rating decisions and the rating is expected to backfire, resulting in a lower fraction of correct ratings compared to independent groups (Figure 2.2 A, right). This happens even if the average difficulty across all individuals is below 0.5.

> *H2: When it is not too difficult for ideologically aligned individuals to classify a message correctly ($d_{align} < 0.5$), then individuals in segregated groups (with information about previous rating choices) classify true messages more often correctly than individuals in independent groups (without information about previous rating choices).*
>
> *H3: When it is difficult for ideologically aligned individuals to classify a message correctly ($d_{align} > 0.5$), then individuals in segregated groups (with information about previous rating choices) classify false messages less often correctly than individuals in independent groups (without information about previous rating choices).*

The center panel of Figure 2.2 B illustrates the evolution of the rating signal when a true message originates in an ideologically aligned cluster, showing how the fraction of correct decisions by an agent's position in a sequence (averaged over 10,000 simulation runs) is strictly higher in the segregated scenario than in the independent scenario. This can be attributed to the positive feedback loop that is likely to occur when a true message with low aligned difficulty (d_{align}) accumulates an increasingly accurate rating signal. The center plot of Figure 2.2 B also shows how the fraction of correct decisions among misaligned individuals decreases in *i*'s position. Because for true messages $d_{mis} > d_{align}$, rating accuracy will decrease to some extent among misaligned individuals. Conversely, if a false message originates in an ideologically aligned cluster (Figure 2.2 B, right panel), high aligned difficulty will trigger a negative feedback loop and the message accumulates an increasingly incorrect rating signal. Ratings will also recover to some extent once misaligned agents rate the false message because $d_{mis} < d_{align}$. Taken together, we expect the following dynamics:

> *H4: When it is not too difficult for ideologically aligned individuals to classify a message correctly ($d_{align} < 0.5$), then in segregated groups (with information about previous rating choices) classification accuracy first gradually improves among aligned individuals (H4a) and then gradually deteriorates among misaligned individuals (H4b).*
>
> *H5: When it is difficult for ideologically aligned individuals to classify a message correctly ($d_{align} > 0.5$), then in segregated groups (with information about previous rating choices) classification accuracy first gradually deteriorates among aligned individuals (H5a) and then gradually improves among misaligned individuals (H5b).*

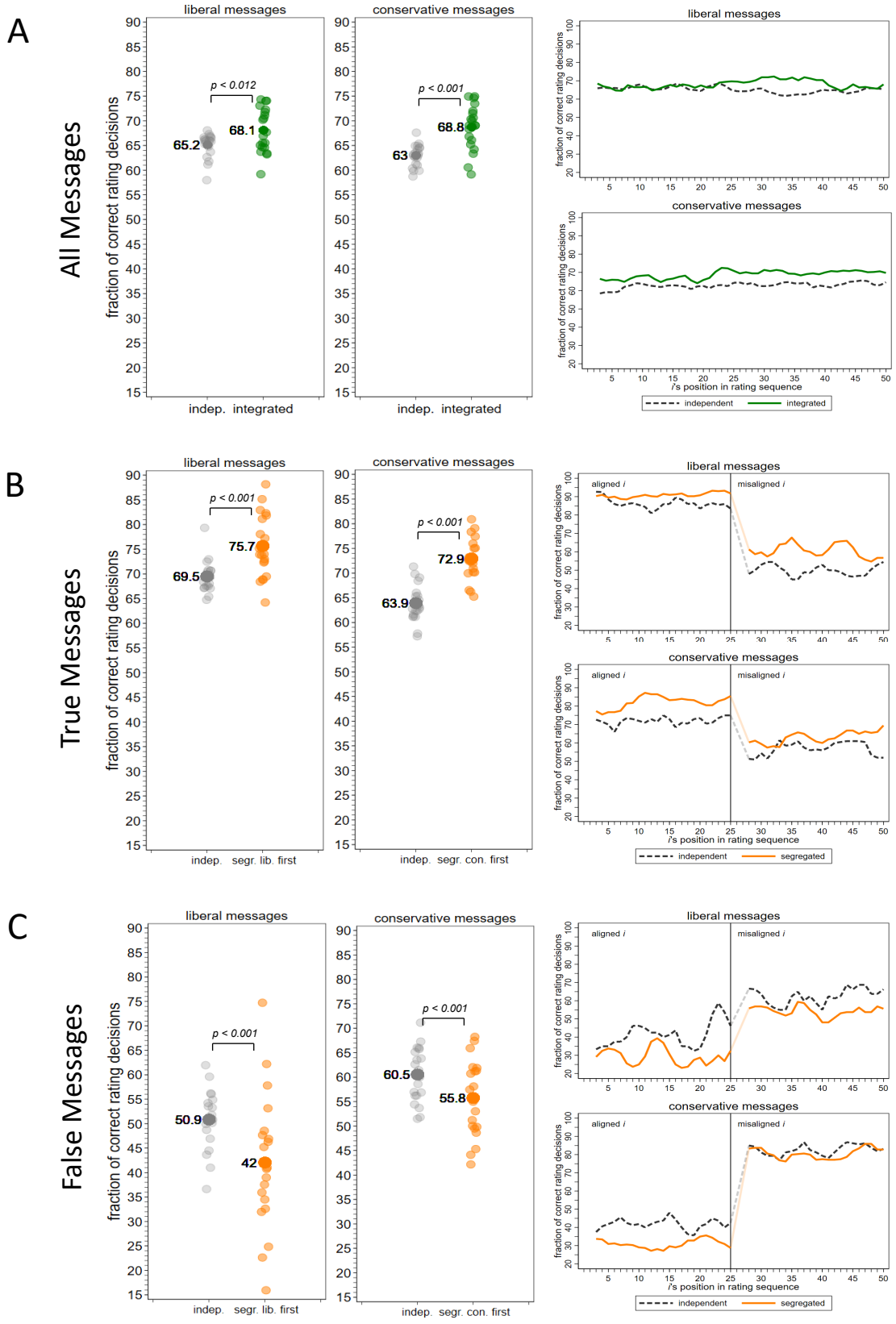


Figure 2.3 Broadcasting ratings enhances subjects' rating accuracy in ideologically integrated groups **(A)** and, when messages are true, also in segregated groups **(B).** Ratings backfire in ideologically segregated groups when messages are false **(C).** Left side of the figure: Each

small, shaded circle represents the number of correct rating decisions divided by all rating decisions in one group. Large circles represent means over groups. Right side of the figure: Lines represent moving averages over three subjects. $N_{subjects}$ = 4,000; N_{groups} = 80. All messages with $\bar{d} < 0.5$; Panel B: true messages with $d_{aligned} < 0.5$. Panel C: false messages with $d_{aligned} > 0.5$. See Methods.

2.3 Results

We tested our hypotheses by letting 80 groups of 50 participants sequentially rate true and false informational messages in an online experiment (N = 4,000 participants with a total of 80,000 decisions). An equal number of self-reported conservative or liberal subjects rated informational messages that clearly supported either of the two ideological viewpoints. In doing so, we ensured that participants had systematic cognitive biases in favor of or against certain messages. Subjects were recruited from Amazon Mechanical Turk and Prolific. We implemented the three conditions studied in the simulations: a segregated condition (20 groups starting with liberals and 20 groups starting with conservatives), an integrated condition (20 groups), and an independent condition (another 20 groups). Each rating group featured 25 liberal and 25 conservative subjects. In each group, only one subject was active at a time to ensure a continuous, sequential evolution of real-time ratings. Subjects were recruited in small batches according to the number of available slots in experimental groups. Each subject answered to the same set of 20 messages, totaling 1,000 rating decisions per group. See Figure 2.1, Methods, and Appendix for details. Unless indicated otherwise, test results are derived from two-sided randomization tests. We test hypotheses separately for liberal and for conservative messages as to ensure a homogenous message sample in each step of the analysis.

2.3.1 Integrated groups

Consistent with Hypothesis 1, broadcasting the rating in ideologically integrated groups improved classification accuracy: The fraction of correct rating decisions in integrated groups was higher than in independent groups, both when liberal messages were rated (integrated 68.1% vs. independent 65.2%; ATE = 2.9%, p < 0.011, N = 40) and when conservative messages were rated (integrated 68.8% vs. independent 63.0%; ATE = 5.7%, p < 0.001, N = 40). Figure 2.3 A illustrates this finding. Importantly, the fraction of correct rating decisions was higher irrespective of whether subjects rated true messages (integrated 71.6% vs. independent 66.7%; ATE = 4.9%, p < 0.001, N = 40) or false messages (integrated 65.0% vs. independent 61.1%; ATE = 3.9%, p = 0.049, N = 40). Aggregating over liberal and conservative messages further revealed a robust treatment effect, with 68.5% of rating decisions being accurate in integrated groups and 64.0% in independent groups (ATE = 4.4%, p < 0.001, N = 40).

2.3.2 Segregated groups

Supporting Hypothesis 2, subjects were more often accurate if a true message was being rated and those who aligned with the connotation of the message were to do ratings first (Figure 2.3 B). In segregated groups where liberals rated liberal-leaning true messages first, the overall fraction of correct rating decisions rose by 6.2 percentage points as compared to independent groups rating the same messages (independent 69.5% versus liberal-first 75.7.%; ATE = 6.2%, $p < 0.001$, $N = 40$). Similarly, when conservatives rated conservative-leaning true messages first, the fraction of correct rating decisions increased by 9.0 percentage points (independent 72.9% versus conservative-first 62.9%; ATE = 9.0%, $p < 0.001$, $N = 40$).

The right side of Figure 2.3 B illustrates how broadcasting the rating in segregated sequences enhances rating performance: Compared to subjects in independent groups, the average propensity of a subject to make a correct rating decision in segregated groups increases after the first few ratings have been made, and then consistently stays above the average accuracy of subjects in independent groups. The large decreases in accuracy around the 25th individual in Panel B reflect that aligned subjects are likely to rate aligned true messages as true while misaligned subjects are more likely to rate them as false.

Our third hypothesis postulates that ratings backfire when aligned subjects rate a *false* message first. Comparisons of independent and segregated groups in Figure 2.3 C show that this was indeed the case. In segregated groups where liberals rated liberal-leaning, false messages first, the fraction of correct rating decisions decreased from 50.1 percent in independent groups to 42.1 percent in liberal-first groups (ATE = 8.9%, $p = 0.013$, $N = 40$). In segregated groups where conservatives rated conservative-leaning, false messages first, the fraction of correct ratings sunk from 60.6 percent in independent groups to 55.7 percent in conservative-first groups (ATE = 4.8%, $p = 0.033$, $N = 40$). Backfiring becomes further visible in the right side of Figure 2.3 C: Real-time broadcasting of rating decisions decreases accuracy when aligned subjects make incorrect rating decisions in the beginning of a rating sequence, which influences subjects to make incorrect decisions later in the sequence.

Results of a multilevel logistic regression analysis show mixed evidence for Hypothesis 4 (see details in Section 2.6, Table 2.4). For true messages (H4a), aligned subjects' likelihood to make a correct rating decision did increase relative to their position in a rating sequence among conservative subjects ($\beta = .020$, $p = .02$). However, this was not the case for liberal subjects ($\beta = .015$, $p = .22$). We did not find decreasing tendencies for misaligned subjects to make correct rating decisions when messages were true (H4b). The tendency for liberal misaligned subjects was positive ($\beta = .019$, $p = .01$) and not significant for conservative misaligned subjects ($\beta = -.00$, $p = .44$). Similarly, no evidence for Hypothesis 5 was found (details in Section 2.6, Table 2.5). Aligned subjects' likelihood to make a correct rating decision did not

decrease relative to their position in a rating sequence (H5a), both among liberal aligned subjects (β = -.007, p = .53) and among conservative aligned subjects (β = -.001, p = .87). Neither did we find increasing tendencies for misaligned subjects to make correct rating decisions (H5b), irrespective whether they were liberal (β = - .002, p = .89) or conservative (β = .004, p = .70).

While patterns are much clearer on the macro-level, it is likely that individual idiosyncrasies contributed to the lack of clearly recognizable trends on the individual level: Our theoretical model assumed identical susceptibility to the rating signal and identical ability among subjects of the same ideological leaning to correctly classify information. In reality, subjects differ along those dimensions, contributing to noisier patterns that require more data for idiosyncrasies to cancel out. It is also thinkable that our participants acted more heuristically than the fine-grained interpretation of the rating signal used by agents in the simulation model. The rating signal may have only influenced subjects' rating choices when the discrepancy between the count of true versus false rating decisions was sufficiently large. If this were the case, trends on the individual position level would become more static and less continuous than those suggested by the theoretical simulations in Figure 2.2.

2.4 Discussion

Our findings identify a key condition for the viability of real-time user ratings as an intervention against misinformation in online social networks: the presence of a sufficient degree of ideological mixing. In bipartisan environments with only moderate homophily, the enabling of real-time user ratings may succeed at tempering belief and reducing spread at the crucial early stage of propagation. Existing approaches such as professional fact-checking have not been able to intervene in such a timely manner. The availability of information on the veracity perceptions of previous others then allows individuals to more often correctly classify both true and false messages than in the absence of such information. While partisanship is often thought to amplify users' belief in misinformation (Del Vicario et al., 2019; Lazer et al., 2018), this finding speaks to the resilience of well-mixed, balanced bipartisan communities (Shi et al., 2019). In ideologically segregated environments, by contrast, false information is more often incorrectly rated as true because systematically biased, early ratings mislead later decision-makers to make incorrect rating decisions as well. Such backfiring poses a challenge to the benefits of real-time ratings since many online social networks are marked by substantial ideological segregation (Bakshy et al., 2015; Boutyline & Willer, 2017; Cinelli et al., 2021; Del Vicario et al., 2016). Our study provides a systematic assessment of such backfiring effects in a controlled experimental setting.

In order to achieve this control, we had to build an artificial experimental environment that inevitably lacked contextual elements present on many real-world

platforms. An example of such an element is the felt presence of like-minded others in segregated online spaces, which is known to shift individual behavior towards sharing information according to partisan identity rather than information veracity (Barberá et al., 2015; Cinelli et al., 2021; Jun et al., 2017; Scheufele & Krause, 2019). By hiding other community members' ideological identities, our experiment excluded such corrupting effects on individual behavior. Another example is that we provided no incentives for correct veracity judgements. On the one hand that makes our environment match online social media that do not provide formal incentives either (Prior et al., 2015), but on the other hand there may be reputational incentives on many platforms that lead some categories of users to care about veracity more on such platforms than in our experiment. Future research examining backfiring effects in more ecological environments would complement the present study, for example by investigating ideological segregation and the temporal evolution of veracity assessments on Twitter's birdwatch or similar instruments on other platforms.

Given our study's finding that real-time user ratings backfire in ideologically segregated environments, another important avenue for future research is therefore to explore how such backfiring can be prevented. One approach toward achieving this is the weighting of ratings by user ideology, which would prevent them from becoming inaccurate when users' ideological biases are correlated with rating order or when populations are ideologically unbalanced. Alternatively, broadcasting a rating could be paused in highly homogenous environments until the rating is composed of a sufficiently diverse user base. Both these options, however, would require that the ideology of users be known or derived from earlier sharing and posting behavior. Moreover, initially paused broadcasting would come at the loss of potentially being able to warn users about false content early in the diffusion process.

The availability of a rating system can only limit the spread of misinformation if it influences sharing behavior. Our experiment only studied accuracy judgments, not the resulting sharing behavior. Another direction of future research is to investigate the consequences of rating systems for user behavior. One such consequence may be that users refrain from sharing information with a bad reputation because they do not want to risk misleading others (Kim et al., 2019). Broadcasting ratings along with a message will also make visible who shares information that is likely false. This would make it easier for both network neighbors and the online platform to put users under scrutiny who repeatedly share information with a bad reputation. As a consequence, users might consider carefully if they want to share such information. Recent research suggests that positive social cues facilitate sharing of information more when it is true rather than false (Z. Epstein et al., 2022). Future research may investigate if, conversely, users also avoid sharing information when social cues are negative, and whether this occurs out of fear of backlash, or out of intrinsic

hesitation to spread potential falsehoods. Platforms may also consider incentivizing such reputational considerations to improve the functioning of user ratings: On Twitter's birdwatch, users can only publish a rating if enough other users have identified their previous assessments as 'helpful'. Reputational efforts can also be incentivized by exposing distinguished users to fewer advertisements, granting higher visibility to those with better reputations, or by rewarding especially diligent community members with 'badge systems' as already implemented on Facebook.

Crowd-based rating systems require that the rating signal is informative. Research suggests that online users are reasonably able to discern true content from false content most of the time (Allen et al., 2021; Bail et al., 2020; Pennycook & Rand, 2019a, 2019b), and that they can be nudged to base their sharing decisions more on veracity, thus 'de-biasing' users' assessments (Maertens et al., 2021; Pennycook et al., 2020, 2021; Pennycook & Rand, 2019b). Of course, rating systems may be vulnerable to manipulation, e.g. by social bots or online trolls. This threat becomes particularly severe when malevolent behavior in one dominant ideological direction is concentrated among those who first rate a message, having similar adverse effects as the ratings of ideologically friendly users in segregated networks. These various limitations notwithstanding, we conclude that on ideologically integrated platforms, real-time user ratings can be a promising intervention for identifying misinformation early in its diffusion process and preventing users from believing in it. On highly segregated platforms, however, rating systems are likely to make things worse.

2.5 Methods

The experiment was pre-registered on the Open Science Framework and approved by the Ethics Committee of the European University Institute, Florence. Data were collected between August 18 and December 31, 2021.[3] All experiments and subsequent data handling were performed in accordance with relevant guidelines and regulations. We obtained informed consent from each participant prior to the experiment. The Appendix section 'Recruitment of experimental subjects' provides a detailed account of subject handling and consent procedures.

2.5.1 Informational messages

We used 20 true and false informational messages with either a liberal or a conservative ideological connotation to be rated by the subjects. Subjects were instructed to 'read 20 statements and click true if you believe a statement to be true and click false if you believe it to be false'. Within each experimental group, subject *i* had to complete rating all messages before subject *i*+1 could start their task. True messages summarized the main finding of a scientific article published in a social science or general science journal after 2015 as to provide a ground truth (example:

[3] Pre-registration and replication package available via https://osf.io/p5byq/.

"Gender diversity in student teams measurably improves their productivity"). False messages incorporated proven falsehoods by summarizing the inverted central finding of a published scientific article (example: "Human-induced CO2 levels in the air have no measurable impact on the likelihood of wildfires in California"). We ensured through pretesting that messages were indeed perceived as liberal or conservative leaning. Messages had the length of a tweet (< 280 characters) as to resemble pieces of information on online social media. We chose a balanced message set of 5 liberal and false, 5 liberal and true, 5 conservative and false, and 5 conservative and true messages. See Section 2.6.3 'Message selection' for a detailed account of the message set.

2.5.2 Analytical strategy

We calculate average difficulty $\bar{d}$ (for each message) as the fraction of correct rating decisions by the total of all decisions in the independence condition. d_{align} is calculated as the fraction of correct rating decisions among those who align with the ideological connotation of a message by all rating decisions from this group. Since decisions are completely independent, values for $\bar{d}$ and d_{align} do not have to be computed at the sequence level but are aggregated over all decisions in that condition. Throughout the analyses, we only selected messages with $\bar{d} < 0.5$. This was done to ensure that the average decision maker was better than random and to prevent a real-time rating from backfiring regardless of group composition. For each hypothesis, we then selected those messages that fell into the respective scope of the hypothesis, determined by the values of $\bar{d}$ and d_{align} (e.g., for the test of H2 we select only messages with $d_{align} < 0.5$). To test Hypotheses 1-3, we use a non-parametric permutation test with 100,000 permutations. Hypothesis 4, unlike the other hypotheses, concerns individual rather than group behavior and thus requires an individual-level test. We use a multilevel mixed-effects logit regression in which we regress individual rating decisions (correct vs. incorrect: 1 / 0) on the subjects' position in the sequence (see Section 2.6, Table 2.4 'Regression results Hypothesis 4'). Note that $\bar{d}$ and d_{align} are estimates rather than 'true values'. In the Appendix section 'Robustness of findings', we present analyses taking statistical uncertainty of $\bar{d}$ and d_{align} into account and obtain similar results for H1, H2, H4 and H5 at conventional significance levels.

2.6 Appendix

2.6.1 Recruitment of experimental subjects

Measuring performance at the level of 50-subject groups, our study required many participants. We therefore recruited from two online crowd-working platforms, Amazon Mechanical Turk (MTurk) and Prolific. 40 of our 80 rating groups were occupied by subjects recruited from MTurk; the other 40 groups were recruited from Prolific. Each set of 40 groups comprised 10 groups for each type of group / treatment condition. Hence, both datasets comprised 2,000 subjects (1,000 liberals and 1,000 conservatives). On MTurk, data were collected between August 18 and October 12, 2021 and between November 10 and December 31, 2021 on Prolific. Participation was restricted to US residents to ensure a homogenous participant pool responsive to our informational messages. On MTurk, we recruited subjects by posting a general advertisement for US 'workers'. Prolific allows for a more selective targeting of participants, and for this reason, we posted separate advertisements for liberals or conservatives. Because a large share of US Prolific workers identifies as liberal or conservative but does not indicate this in the platform's pre-screening information, we also recruited Democrats and Republicans (approximately 20% of the Prolific dataset), and Biden versus Trump voters from the 2020 election (another 20% of the dataset). On Prolific, we excluded active MTurkers to ensure independence of observations. Apart from these slightly different recruiting procedures, the studies were identical. Because only one subject per group could be active at a time, recruitment of subjects was sequential, and advertisements were continuously adapted according to the availability of liberal or conservative slots in our experimental groups.

After having clicked on our advertisement on MTurk or Prolific, subjects were routed to a screener study in which they were asked a standard question about ideological self-identification.[4] Subjects identifying as moderate were remunerated $0.15 and excluded from the study. Remaining subjects were instructed on their task, indicated their informed consent, and were asked the ideological self-identification question a second time. Subjects whose ideological leaning did not match their self-reported ideology in the initial screener study were remunerated $0.15 and excluded. The self-identification question, informed consent form, and experimental instructions are presented in Figure 2.4. Subjects who completed their task successfully earned $1.5. When subjects entered the experiment, we informed them that they were to do ratings in 'groups' and, depending on the condition, explicated whether they could see others' ratings or not. To mimic an online social media platform, where no financial incentives for a certain individual behavior are present, we chose to pay subjects flat fees instead of paying them for the accuracy with which they classified true and false messages.

[4] As implemented in the 2020 American National Election Studies Time Series Study: https://electionstudies.org/data-center/2020-time-series-study/

Task & payment

Thank you for accepting this HIT. This academic research study should not take more than 8 minutes.

1. You will see 23 short informational statements, which are either true or false.
2. Your task is to read the statements and then **click 'true' if you believe a statement to be true or click 'false' if you believe it to be false**.

Example statement:

"Black and Hispanic students admitted to elite US colleges perform more poorly than Asian students." - True or false?

When you finish the HIT, you will receive a completion code to get paid. You will receive **$0.15 + $1.5** upon completion of the study.

Disclaimer

There are statements to check that you are **paying attention** and are not a robot. If you answer these statements incorrectly, if you **make your decisions too fast**, or if you **fail to finish** your task in **10 minutes**, you are excluded from this study. You can participate only once. **Please do not close this tab or reload the page during the task.** If you leave the website during the task, you will not receive any earnings.

Terms and Agreements

The data collected in this study does not include any personally identifying information about you. By participating, you understand that the research data gathered during this study will be used by the researchers. A dataset that contains your fully anonymous data may be published. A record of your workerID will be deleted after this study.

The data for this study is collected and controlled by Arnout van de Rijt of the European University Institute (EUI) and processed by Jonas Stein of the University of Groningen. Your data is protected by EUIs data protection policy (PD10/2019). You may contact EUI's data protection officer through data_protection_officer@EUI.eu. You have the right to withdraw your consent for participating in this study at any time by closing this tab during the task. Upon withdrawal, your data will be deleted.

If you have questions concerning this study, please write to sociology.vanderijt@gmail.com

A copy of the consent form co-signed by the researcher can be obtained here.

Personal data

I have received sufficient information about this study and understand my role in it. The future processing of my personal data has been explained to me and is clear.

Terms of service

I have carefully read and understood the above information, agree to the terms for participation in this study, and am at least 18 years of age.

Before we start

Here is a 7-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative. Where would you place yourself on this scale?

- extremely liberal
- liberal
- slightly liberal
- moderate, middle of the road
- slightly conservative
- conservative
- extremely conservative

Figure 2.4 General instructions

2.6.2 Data quality

Apart from checking for consistent ideological leaning, we undertook further measures to ensure high data quality. First, we excluded subjects who did not read messages carefully. We excluded subjects from further participation if they made a rating decision after having seen a message for less than three seconds thrice. Second, if a subject took more than 10 minutes to finish their task, they were excluded from the study. Third, we excluded subjects that failed at least one of our three attention check messages (example: "Europe is in the southern hemisphere." True / False). Subjects who failed one of these quality checks were only paid the remuneration for our screener study, $0.15. Excluded subjects were replaced with new subjects and their rating choices were not considered in the counts that were displayed as rating signals to subsequent subjects.

Of the 9,512 subjects who reacted on our study advertisement and went through the screener tasks, 3,284 (34.5%) were excluded because they identified as moderates, reported inconsistent ideological leaning, or had an active MTurk account parallel to their participation on Prolific. 684 (7.1%) subjects chose to not proceed with the experiment after the pre-screener or refused consent. 568 subjects (5.37%) were unable to participate because they showed up at a time when no spots for subjects of their ideological leaning were available. 975 subjects (10.3%) failed at least one of our quality checks: 511 subjects (5.4%) had unreasonably short response times, 387 (4.1%) failed an attention check question and 77 subjects (0.8%) did not finish within 10 minutes. Overall, participation rates were similar across samples. On MTurk, 41.4% of all individuals who reacted to our study advertisement (2,000 out of 4,834) successfully finished the study; on Prolific, it was 42.8% (2,000 out of 4,678).

Table 2.1 Message set in the experiment

ID	true	lib.	text	d_{align}	$\bar{d}$	Hypoth.
1	0	0	Human-induced CO2 levels in the air have no measurable impact on the likelihood of wildfires in California	.48	.32	1
2	0	0	Black people have similar chances of receiving a job offer than white people, as long as they are similarly qualified	.70	.43	1, 3, 4
3	0	0	Children raised by homosexual parents are 10 percent more likely to experience mental health issues than children raised by heterosexual parents	.51	.32	1, 3, 4
4	0	0	Affirmative action has reduced the number of highly qualified applicants for specialized jobs by roughly 40 percent	.56	.37	1, 3, 4
5	0	0	US states enacting major tax cuts for the wealthy show higher growth rates and economic prosperity, also for poor people, than states without tax cuts	.48	.45	1, 3, 4
6	1	0	Cannabis use is associated with lasting damage to adolescents' cognitive functions	.36	.40	1, 2
7	1	0	Most gun control restrictions generally have had little effect on violent crime in US cities	.17	.37	1, 2
8	1	0	Children born to married parents have slightly better health at age 5 than children born to unmarried parents	.28	.32	1, 2
9	1	0	Unmarried couples are more likely to have additional sexual partners as compared to married couples	.27	.38	1, 2
10	1	0	The United States have won more medals at the Summer Olympics than any other nation	.30	.33	1, 2
11	0	1	Large influxes of migrants strongly increase the number of jobs available to native applicants of the host country	.37	.33	1
12	0	1	In countries run by left-wing political parties, immigrants integrate and learn the spoken language of the host country nine times faster than in right-wing countries	.67	.50	-
13	0	1	Men in same sex relationships are more likely to be in serious relationships than heterosexual men	.33	.30	1
14	0	1	Due to increasing inequalities in US healthcare provision, cancer patients in the US have 85 percent lower survival chances than patients in both eastern and western Europe	.61	.49	1, 3, 4
15	0	1	Israel's construction of a southern border wall between 2010 and 2013 was not effective, as the annual numbers of illegal crossings increased by 80 percent	.58	.49	1, 3, 4
16	1	1	Due to safe jobs and fixed wages, eastern Europeans felt economically more secure under socialism than under capitalism	.30	.41	1, 2
17	1	1	Police officers speak significantly less respectfully to black community members than to white ones in everyday traffic stops	.22	.33	1, 2
18	1	1	Gender diversity in student teams measurably improves their productivity	.10	.25	1, 2
19	1	1	When Latino immigrants move into an area, there is no measurable increase in homicide rates	.12	.26	1, 2
20	1	1	Germany's relative income equality has resulted in longer life expectancies for Germans as compared to Americans	.12	.28	1, 2

2.6.3 Message Selection

Prior to the experiment, messages were calibrated through pretesting: Independent evaluations of 350 conservative and 350 liberal subjects ensured that liberal messages were more likely to be perceived as true by liberal subjects, and conservative messages more likely to be perceived as true by conservative subjects. Second, we ensured during the pretest phase that subjects were more likely to make a correct rather than an incorrect evaluation of a messages' veracity: The median rating of a message always had to reflect the actual veracity of the message. Put differently, each message had to have an average difficulty below 0.5 and a bias greater zero. Both are scope conditions of this study: The wisdom of crowds requires that more than 50 percent of the population make a correct rating decision in independence (Baker, 1975; Condorcet, 1785), and there must be a difference in message difficulty among aligned versus misaligned subjects for a segregated rating orders to have any effect at all. Out of an original set of 144 messages used for pretesting, we chose a subset of 20 messages as compared to more messages to prevent subjects getting tired or inattentive after too many messages. Table 2.1 presents an overview of the message set used in the experiment.

Table 2.2 Subject behavior in the independence condition

Subject ideology	Message ideology	Message veracity	% thought true[1]
conservative	conservative	false	56.6
		true	72.4
	liberal	false	33.2
		true	54.3
liberal	conservative	false	19.1
		true	55.5
	liberal	false	51.3
		true	84.5

[1] per cell: $N_{subjects}$ = 500; $N_{decisions}$ = 2,500

2.6.4 Subjects' rating behavior

Consistent with the scope conditions of this study, subjects were reasonably able to tell true from false messages in the independent sequences of the experiment. On average, subjects in the independent condition thought 66.7% of true messages to be true, while this was the case for only 40.1% of false messages (paired t-test: $t = 34.5$, $p < 0.001$, $N = 500$). At the same time, subjects in the independent condition found 66.2% of messages that aligned with their ideology to be true, but only 40.5% of misaligned messages (paired t-test $t = 28.0$, $p < 0.001$, $N = 500$). This shows that indeed, cognitive biases played a role in such a manner that messages supporting one's own viewpoint are more often thought to be true – independent of the actual veracity of a message. It is noteworthy that liberals were especially inclined to find

liberal true messages true (85.7% of messages in this category), while this was rarely the case for false conservative messages (19.1%). Classifying false conservative messages as false and liberal true messages as true made liberals better at making correct rating decisions overall (average liberal subject: 67% correct vs. conservative subject: 59%; t-test $t = 11.3, p < 0.001, N = 500$). At the same time, liberal subjects were more biased than conservatives: The difference between finding aligned versus misaligned messages true was larger among liberals (30.1 pp.) as compared to conservatives (20.1 pp.; t-test $t = 5.5, p < 0.001, N = 500$). An overview of subject behavior by message veracity and message ideology is presented in Table 2.2. Overall, independent subjects rated 63.5% of their messages correctly, suggesting that ability in the population was indeed above 0.5 for at least a sizeable portion of all messages. Subjects in the independent condition of the Prolific sample were slightly better at making correct rating decisions than in the MTurk sample (64.8% versus 61.9%, t-test $t = 3.8, p < 0.001, N = 500$). Simultaneously, the difference in finding aligned versus misaligned messages to be true (bias) was higher among Prolific subjects as compared to MTurk subjects (Prolific: 31.5 pp., MTurk: 19.9 pp.; t-test $t = 6.4, p < 0.001, N = 500$). Both differences are consistent with prior studies finding subject quality higher on Prolific (Eyal et al., 2021; Peer et al., 2017) – subjects may have read questions more closely and thought of them more proactively. Indeed, a much lower number of quality check failures (94 subjects on Prolific versus 881 subjects on MTurk) suggests that fewer subjects showed satisficing behavior on Prolific. Because subject behavior was slightly different in the Prolific dataset as compared to the MTurk dataset, the following section reports additional analyses where underlying message parameters were constructed for each dataset separately.

2.6.5 Robustness of findings

We report two analyses of the robustness of our findings. First, since subjects from the MTurk dataset showed slightly different behavior than subjects from the Prolific dataset (see previous section), we treat subjects from each dataset as different populations and compute message parameters for each dataset separately. Second, we take into account that message parameters are estimates and not true values and only included messages whose 95 % confidence interval of average difficulty $\bar{d}$ did not overlap with the 0.5 threshold. For Hypotheses 2, we excluded all messages for which the upper confidence bound of d_{align} was below 0.5 and for Hypotheses 3 and 4, we excluded all messages for which the lower confidence bound was above 0.5. Note that since parameters were computed separately per dataset, sample sizes were smaller and standard errors larger. Table 2.3 presents an overview of how many messages were selected for each Hypothesis and dataset. We then computed the fraction of correct rating decisions per sequence and conducted the same analyses as in the main results section.

Consistent with Hypothesis 1, broadcasting ratings in integrated sequences led to an improvement in rating performance. In integrated groups, the fraction of correct rating decisions was higher than in independent sequences, both among liberal messages (72.7% versus 68.5%; *ATE* = 4.2%, $p < 0.001$, N = 40) and conservative messages (69.3% versus 63.7%; *ATE* = 5.6%, $p < 0.001$, N = 40). The additional results also support Hypothesis 2. Subjects were better at making correct rating decisions in groups where those who aligned with the connotation of a message were to do ratings first. In segregated groups where liberals rated first, the fraction of correct rating decisions rose by 5.7 percentage points as compared to independent groups (independent 71.4% versus liberal-first 77.0.%; two-sided randomization test: *ATE* = 5.7%, $p < 0.001$, N = 40). In segregated groups where conservatives rated first, average accuracy increased by 9.0 percentage points (independent 63.9% versus conservative-first 72.9%; two-sided randomization test: *ATE* = 9.0%, $p < 0.001$, N = 40).

Table 2.3 Number of selected messages per hypothesis

Hypothesis	Mturk	Prolific
1	16	18
2	10	10
3	2	3
4	2	3
Messages	20	20

Different from the results in the main text, the fraction of correct ratings did not decrease significantly in segregated groups when d_{align} was above 0.5. However, results indicated effects of identical direction and similar strength: In segregated groups where liberals rated first, the fraction of correct ratings decreased by 6.6 percentage points (independent 55.2% versus liberal-first 48.6.%; two-sided randomization test: *ATE* = 6.6%, $p = 0.126$, N = 20). In segregated groups where conservatives rated first, the fraction of correct ratings sunk by 3.4 percentage points as compared to independent sequences (independent 54.4% versus conservative-first 57.8.%; two-sided randomization test: *ATE* = 3.4%, $p = 0.108$, N = 40).

Consistent with the main results for Hypothesis 4, a multilevel logit regression for true messages did show significant increasing rating performance over conservative aligned individuals' positions ($\beta = .019$, $p = .01$), but no increasing performance for liberal aligned subjects. No decreasing performance among misaligned conservative or liberal subjects was found. Finally, in line with the main results for Hypothesis 5, no significant decreasing rating performance over aligned individuals' positions in

rating groups was present; or increasing performance for misaligned subjects. We conclude that results from our robustness analyses are similar to those in the main text – Hypotheses 1, 2 and 4 allowed for identical conclusions, and Hypothesis 3 achieved similar effects despite lack of significance. The lack of significance for H3 can likely be attributed to the fact that fewer messages were considered in the robustness analysis, and hence that less statistical power was given.

Table 2.4 Regression results Hypothesis 4

Multilevel logit regression on making a correct rating decision

	Segregated (Liberal first)		Segregated (Conservative first)	
	Coef.	SE	Coef.	SE
aligned subject				
constant	2.520 ***	0.21	1.430 ***	0.14
position$_i$ (1-25)	0.015	0.01	0.020 *	0.01
misaligned subject				
constant	-0.103	0.29	0.880 *	0.41
position$_i$ (26-50)	0.019 *	0.01	-0.008	0.01
N Observations	1.000		2.000	
N Subjects	500		500	
N Sequences	20		20	

The regression models include an individual-level and a sequence-level variance term in the regression, and standard errors are clustered at the individual level. Only decisions for true messages with $d_{align} < 0.5$; $\bar{d} > 0.5$.

Table 2.5 Regression results Hypothesis 5

Multilevel logit regression on making a correct rating decision

	Segregated (Liberal first)		Segregated (Conservative first)	
	Coef.	SE	Coef.	SE
aligned subject				
constant	-0.988 ***	0.25	-0.914 ***	0.17
position$_i$ (1-25)	-0.007	0.01	-0.001	0.01
misaligned subject				
constant	0.299	0.48	1.563 ***	0.42
position$_i$ (26-50)	-0.002	0.01	0.004	0.01
N Observations	1.000		2.000	
N Subjects	500		500	
N Sequences	20		20	

The regression models include an individual-level and a sequence-level variance term in the regression, and standard errors are clustered at the individual level. Only decisions for false messages with $d_{align} > 0.5$; $\bar{d} > 0.5$.

Chapter 3

How homophily can improve collective decision-making in diverse teams

This chapter is published as: Stein, J., Frey, V., & Flache, A. (2024). Talk less to strangers: How homophily can improve collective decision-making in diverse teams. *Journal of Artificial Societies and Social Simulation*, *27*(1), 14.

Abstract

Identity diversity in teams brings advantages for complex decision-making because it is associated with cognitive diversity among team members. At the same time, homophilic interactions along shared identity dimensions can hinder information exchange among dissimilar individuals and threaten successful exploitation of the team's cognitive diversity. We present an agent-based model to investigate how homophily impacts decision-making quality in diverse teams. Team members communicate information in a 'hidden profile' setting where some pieces of information are known only to single individuals while other pieces of information are known to subgroups with the same identity. While intuition may suggest that homophily impairs collective decision-making, our model reveals how homophilous environments lead to better collective decisions: Homophily fosters temporary disagreements between dissimilar team members, which grant teams additional time to uncover crucial information that would not have been shared otherwise. Longer discussion time comes along with improvements in the quality of the final decision, indicating a trade-off between the time needed to deliberate and decision quality.

3.1 Introduction

A large literature on work teams documents that diverse teams have a greater pool of social, human, and cultural capital translating into a higher potential for team performance (Phillips & O'Reilly, 1998). Yet, when diversity activates social identity processes (Tajfel, 2010), this potential may not be used. Homophily, the tendency to preferentially interact with those similar to oneself is a strong force in humans (McPherson et al., 2001) and prevents team members from communicating with dissimilar others providing them with information needed to reach higher performance. With an agent-based model, we challenge here the intuition that homophily is detrimental to the performance of diverse teams. We demonstrate that homophily can improve team decision making, studying agent teams confronted with a hidden profile task (Stasser & Titus, 1985) which requires team members to share information not known to others in order to collectively find the best solution to a problem. Our finding contrasts with the conjecture that homophily can hamper the performance of diverse teams (Ertug et al., 2022; Estévez-Mujica et al., 2018; Reagans, 2013) and highlights instead the benefits of limiting the flow of information between dissimilar team members.

Structuring collaboration processes and interaction patterns in diverse teams so that they enhance team decision-making has become an increasingly important issue, as a globalized division of labor, rising international migration, and increasingly diverse workforces have led to a ubiquity of heterogenous decision-making groups in organizations (S. E. Jackson et al., 1995). Research investigating how this trend shapes collaborative work processes identifies both challenges and benefits (Carter & Phillips, 2017; Milliken & Martins, 1996; Phillips & O'Reilly, 1998). Successful integration of *cognitive diversity,* referring to the wealth of perspectives, knowledge, and skills present in a team is found to have mostly beneficial outcomes for the quality of a task, especially when tasks are complex because a conjunction of diverse skills and perspectives is expected to enhance team creativity and foster innovative solutions (Page, 2019).

At the same time, *identity diversity* in teams can provide a challenge to the successful integration of cognitive diversity. Identity diversity is sometimes also referred to as 'surface-level diversity' or 'demographic diversity' (Peters, 2021). Including potential demographic or 'surface-level' traits, we focus specifically on identities that are easily observable for others, salient during collaborative work, and plausibly correlated with cognitive traits.[5] Even without assuming that identity diversity leads to intergroup conflict (Lau & Murnighan, 1998) stereotyping and negative outgroup attitudes (Harrison et al., 1998; Northcraft et al., 1995; Phillips, 2003), it has been consistently documented that individuals tend to associate

[5] Identities satisfying these conditions could be, for example, ethnic identities in a team developing a product targeted towards a diverse consumer base or team members' disciplinary identities in a scientific collaboration.

themselves with similar others and that similarities are usually recognized along common identities (Brechwald & Prinstein, 2011; Ertug et al., 2022; McPherson et al., 2001).

Simulation studies on opinion dynamics (Flache et al., 2017; Hegselmann & Krause, 2002) and experimental studies (Bail, 2016; Balietti et al., 2021; Mäs & Flache, 2013) show that such *homophilous preferences* are sufficient to drive groups apart and induce polarization. Empirically, identity traits often correlate with cognitive traits (Phillips, 2003) and theoretical as well as empirical studies have shown that such correlations tend to amplify polarization tendencies (DellaPosta et al., 2015; Mäs et al., 2013; Stark & Flache, 2012). If individuals socially influence each other but tend to interact with similar others, patterns emerge where distinct sets of opinions revolve around similarities in other, seemingly unrelated dimensions.

Models of opinion dynamics highlight how opinion divergence in teams can disable consensus, but they do not clarify how opinion polarization links to the quality of decision making in a team. In this paper we move beyond modeling the dynamics of opinions alone and develop expectations about how homophilous preferences and social influence affects decision-making *quality* in a team. Intuitively, one can expect this theoretical extension to show how homophily in diverse teams can negatively affect decision making quality. For complex tasks where diverse knowledge must be brought together to obtain an optimal decision, a lack of communication between team members with different identities hampers the conjunction of valuable information, leading to failure to realize good solutions. Second, even if some team members find the optimal solution to the task at hand, a lack of consensus endangers the possibility that this solution is adopted by the team. Finally, lacking communication slows down the deliberation process, making decision-making more costly. This expectation aligns with previous experimental and simulation-based research on problem-solving in groups (Estévez-Mujica et al., 2018): especially in those groups where demographic diversity was high, homophilous interactions prevented access to valuable information. This resulted in worse performance on the group level.

Epistemological studies on decision-making in diverse groups, on the other hand, oppose the conjecture that homophily negatively affects group performance in deliberative tasks. Instead, bounded communication between different individuals is beneficial to team decision quality (Zollman, 2010) because restrictions on social influence will prevent individuals from prematurely adopting others' solutions. In other words, boundaries in communication hinder the rapid dissemination of inferior knowledge, ultimately ensuring that individuals explore the full spectrum of possible decisions before exploiting suboptimal knowledge (D. Frey & Šešelja, 2020; J. Wu & O'Connor, 2021; Zollman, 2010). Similar to the opinion dynamics literature, these works also suggest that 'transient diversity' will lengthen the deliberation

process but point out that added discussion time gives room to ensure that knowledge is optimally explored and disseminated.

Studies of transient diversity emphasize possible advantages of boundaries to communication in team decision making, but they leave us in the dark as to how homophily interacts with diversity in teams. Scholars of this canon are primarily concerned with communication processes in science and argue that skepticism and sparse communication can be induced by adapting macro-level incentive structures such as changing funding policies in research. Yet, we would expect that meso-level social processes such as homophily induced by identity diversity can also bound excessive communication and limit exchange between dissimilar members, even if macro-level structural boundaries are absent. This notion is similar to the 'value in diversity' hypothesis (McLeod et al., 1996), arguing that salient markers of diversity can be beneficial to decision-making quality even when they are unrelated to cognitive traits (Levine et al., 2014; Page, 2019; Peters, 2021; Phillips & Loyd, 2006). Easily observable diversity in identities can help groups to apply healthy skepticism, prevent the placement of undue trust, and foster constructive discussion. In sum, it follows from both the transient diversity literature and the value in diversity literature that tendencies to associate with similar over dissimilar individuals improve collective decision-making quality by helping groups to examine information critically instead of converging around early, suboptimal consensus.

Research on opinion dynamics in diverse teams and studies of transient diversity lead to competing intuitions: preferential interactions among similar over dissimilar team members are either beneficial or detrimental to a team's performance. The present paper uses an agent-based model to theorize how homophilous interaction preferences shape decision-making quality in diverse teams. The model combines central aspects that have not been studied in tandem before: first, it evaluates the quality of the decision that is made, which opinion dynamics models have paid little attention to so far. Second, it considers that background traits shape interaction preferences between individuals without having to assume exogenous incentive structures as outlined by the transient diversity literature.

Our model builds on hidden profile tasks, an established paradigm that has been widely used in experimental research to study decision-making in groups (Lu et al., 2012; Schulz-Hardt & Mojzisch, 2012; Sohrab et al., 2015; Stasser & Stewart, 1992; Stasser & Titus, 1985, 2003). In a hidden profile, a team of decision-makers is equipped with a set of information pieces and instructed to deliberate before choosing one of several available decision alternatives, which differ in quality. Individuals are given different pieces of information at the onset of the deliberation task. Pieces of knowledge that point towards inferior options are 'common information', i.e., known to everyone in the group. Common information anchors decision-makers to initially prefer inferior options. Anchoring effects, a tendency to share knowledge supporting one's own views, and social validation from others with

similar decision preferences can prevent members from sharing or accepting dissenting information. This makes hidden profiles difficult to solve (Lu et al., 2012), which resonates with the conjecture that diversity has a much greater impact on the outcome of a task when it is complex and challenging (Wittenbaum et al., 2004). Information supporting the optimal option is 'unique', in that it is known to not more than one individual. However, a conjunction of multiple pieces of unique information will reveal the optimal option, which captures the well-studied phenomenon of cognitive diversity that bringing together knowledge from different individuals will produce better solutions (Hong & Page, 2004). In addition, hidden profiles are a suitable paradigm for the purpose of the present research because they incorporate a number of features that are often hard to observe or hold constant in natural settings: Hidden profiles allow the experimenter to control the distribution of knowledge, as a predefined set of decision options can be perfectly ranked according to their quality, and all communication processes and their outcomes can be observed and subsequently analyzed.

By using the hidden profile paradigm for studying the interplay of identity diversity and homophily in affecting team decision making quality, we also contribute to the hidden profile literature itself. Identity diversity is a feature that is rarely considered in the literature on hidden profiles and has not received much attention in experimental research using this paradigm (see Phillips et al., 2006 for an exception). Since the impact of identity diversity on decision-making quality is a key aspect of our study, our model extends the traditional hidden profile framework by assigning identity traits to individuals, making them either similar or dissimilar to each other. We capture the aspect that identity is associated with the kind of knowledge individuals possess by distributing a separate set of common information to any group of individuals with a common identity. We further condition interaction preferences on identity traits in such a way that higher homophily levels reflect how much identity-similarity increases the chances of communication between individuals. More precisely, the stronger homophily, the higher are the chances of individuals of identical background to communicate relative to the chances of communication between individuals of different background. Our key interest is in assessing the effects of homophily on decision-making quality. Thus, as further outlined in the 'setup of simulation experiments' section, we measure how likely teams were to obtain optimal consensus given different levels of homophily, and how long it took them to reach a decision. In the results section, we show how homophily affects consensus outcomes, uncovering the underlying mechanism and investigating how other discussion features such as deliberation length and belief changes among team members are affected.

3.2 Model description

Our model develops a formal representation of a diverse work team facing a decision problem as implemented in the experimental setup of the hidden-profile paradigm.

Figure 3.1 illustrates key aspects of the model. We implement a setup where a team seeks to identify the best out of a set of possible decision options. Individuals are equipped with different pieces of information that need to be combined to identify the best option. To this end, we assume a team of *N* agents. Each agent belongs to one of *M* groups where each group consists of agents who share a common identity. Identities could represent, for example, different branches in an organization, or different academic disciplines in an interdisciplinary project. For simplicity, we assume throughout that groups are assumed to be of equal size, i.e., we do not consider unbalanced group sizes. We also abstract from power imbalances between agents or the effects of multiple group memberships.

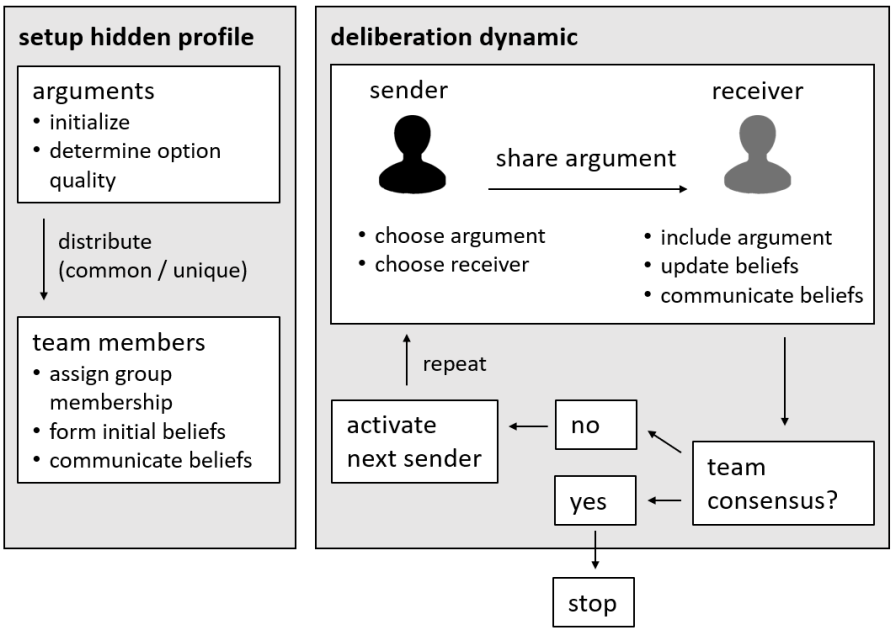


Figure 3.1 Overview of model procedure[6]

The virtual teams in our model face a decision problem, in that the best option o_{max} out of a set of *J* discrete options needs to be identified. Every team member forms her own belief about which decision option is best but is open to influence by other team members. Influence is implemented as a sequence of communication events (right side of Figure 3.1). Agents take turns in sharing an argument with an interaction partner. Every time an argument is emitted, the recipient updates her beliefs and tells her team what option she currently believes to be best. This influence process continues until all agents prefer the same option. This option is the team´s decision. Alternatively, if no consensus is reached after a large number of interaction events (5,000 interactions), the simulation is stopped.

[6] A detailed representation of the model procedure in pseudocode is included in the replication package at https://osf.io/76hfm/.

3.2.1 Decision options and arguments

To create a decision problem as implemented in the hidden profile paradigm, we assume that there is a set of I arguments $A = \{a_1, a_2, \ldots, a_I\}$ pertaining to a predefined set of J decision options $O = \{o_1, o_2, \ldots, o_J\}$. An example of a set of arguments and decision options is presented in Figure 3.2 A. The set of arguments is fixed, implementing a setting in which agents cannot invent new arguments during the deliberation process. Following common practice in hidden profiles, we define the set of arguments and options prior to the start of the decision process. This excludes the possibility that agents come up with 'creative' solutions, which would be hard to design, track, and explain in a simulation. Each argument contains J weights reflecting the degree to which the argument supports the different decision options, i.e. $a_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,J}\}$. Following the standard assumption of the hidden profile literature, we assume that each argument has a positive weight for only one of the options and a weight of zero for the other options. We further assume that there is an equal number of arguments with a positive weight for each decision option. For a given simulated team, weights are randomly drawn from a uniform distribution so that $w_{i,j} \in [0,1]$. The sum of the weights associated with a decision option determines its true quality $Q_j = \sum_1^i w_{i,j}$, and the decision option j with the highest quality is the optimal option, o_{max}.

(A) Arguments

arguments	decision options o_1	o_2	o_3
a_1	.76	0	0
a_2	.17	0	...*
a_3		.63	
a_4		.39	
a_5			.64
a_6			.55
Q_j	.93	1.02	1.19

(B) Argument distribution in the team

Group g_1

agent 1	o_1	o_2	o_3
a_1	.76		
a_4		.39	
$q_{x,j}$	.76	.39	0

agent 2	o_1	o_2	o_3
a_1	.76		
a_5			.64
$q_{x,j}$	.76	0	.64

Group g_2

agent 3	o_1	o_2	o_3
a_3		.63	
a_2	.17		
$q_{x,j}$	.17	.63	0

agent 4	o_1	o_2	o_3
a_3		.63	
a_6			.55
$q_{x,j}$	0	.63	.55

* Empty cells represent weights with a value of zero.

Figure 3.2 Arguments, decision options and option quality. **Panel A:** example set of available arguments and decision options. **Panel B:** argument distribution across groups and agents. Highlighted arguments represent common arguments within groups. Quality scores in red indicate the optimal option in panel A and agents' beliefs about what option is optimal in panel B.

3.2.2 Unique and common arguments

The last ingredient needed to implement a hidden profile task is concerned with the initial distribution of arguments across team members. In hidden profile experiments, participants are provided with so-called 'common' and 'unique'

information. An argument is common when all participants know it already at the outset of the deliberation process. Unique information, in contrast, is provided only to a single participant. This principle is modified for the study of diverse teams: We added that a given piece of common information is provided only to the members of one group (Figure 3.2 B). In this sense, common arguments can reflect certain disciplinary basics that everyone of the same profession was trained with, or specific knowledge shared by everyone working in the same organizational branch. We distribute common arguments in such a way that each agent of group g_m receives the same set of arguments in favor of option o_j. Agents in group g_1 receive C arguments favoring option o_1, agents in group g_2 receive C arguments favoring option o_2, and so on. Like scientists of one discipline thinking that their approach is superior to others or employees being convinced that the 'way things are done' within their organizational branch is best, the distribution of common arguments biases agents of one group to initially favor a specific, but not necessarily optimal option over others. We assume that there are more decision options than groups ($J > M$) so that the optimal option can reside outside of those options supported by common arguments, and thus supported initially by all members of one group.

The procedure by which arguments are distributed initially is designed to ensure that teams face a hidden profile task as described above. Once common arguments are assigned to groups, we consider all arguments that remain. These can pertain to any of the decision options. However, because common arguments are distributed already, those arguments that remain unassigned at this stage most often support options other than those favored by common arguments. In the initialization procedure, agents take turns at randomly drawing from the remaining arguments without replacement until all arguments are distributed. By doing so, these arguments represent 'unique information' that is held by single agents but not by groups. Whether a task is 'hidden' and difficult or 'manifest' and easy is determined by whether the optimal option o_{max} is among those options favored by common arguments. In line with the hidden profile paradigm, we select for our simulation experiments those tasks where only unique arguments favor the optimal option but none of the common arguments in either of the groups do. Hence, to stay within the example of a scientific collaboration, we model a situation in which interdisciplinarity has true advantages because it enables better solutions than those common to either discipline alone. Altogether, our initialization procedure creates the situation of most theoretical interest to us in this study: distinct sets of common arguments ensure cognitive diversity where groups bring different knowledge to the table, with the different groups initially favoring different alternatives and none of the groups favoring the optimal alternative. Unique arguments provide cognitive diversity at the individual level, but they are dispersed over agents across groups and must be brought together to outweigh common stocks of arguments pointing to inferior solutions.

3.2.3 Argument processing and communication

Similar to how objective quality scores Q_j are computed, agents form a perceived quality score for each decision option, $q_{x,j}$, by summing over the weights of the arguments they possess. Agents always believe the decision option to be best that has the highest perceived quality to them and communicate this belief publicly to everyone at the onset of the simulation.

Over the course of the simulation, agents share arguments and update beliefs, thereby deliberating which option is best. Agents are activated sequentially and according to their identifier, i.e., agent 1 in round 1, agent 2 in round 2, and so forth. Additional analyses included in the Section 3.6 show that results do not depend on a sequential activation of agents but are also obtained when agents are activated at random (see Figure 3.9). The first task of the active agent is to select an argument she wants to share. Because the agent has limited capacities to communicate information and can only share one argument per round, she carefully needs to consider which argument is most important to communicate. This consideration is represented by a two-step discrete choice procedure. In the first step of the procedure, the agent chooses an option o_j^* she wants to support. Psychological research suggests that individuals are most inclined to advocate options they deem most preferable themselves (Wittenbaum et al., 2004). For this reason, we assume that the agent is more likely to choose options with higher perceived quality scores relative to the quality score of other options. The probability to choose option *j* at a given moment is formalized by Equation 3.1.

$$p_{o_j} = \frac{e^{\beta * q_{x,j}}}{\sum_{j=1}^{J} e^{\beta * q_{x,j}}} \tag{3.1}$$

The parameter β reflects agents' adherence to choosing an option of higher versus lower perceived quality. When $\beta \to \infty$, the probability of choosing the option with highest perceived quality approximates one while probabilities of choosing other options are zero. When β is one, an option is chosen with a probability proportional to its perceived quality. When β is zero, all options are chosen with equal probability, regardless of their perceived quality.

After an agent has decided which option to support, the second step of the discrete choice procedure determines which argument to communicate. Here, an agent regards all arguments she holds but only considers those weights w_{i,j^*} that contain information on her chosen option o_j^*. She picks one of her arguments with the probability given by Equation 3.2.

$$p_{a_i} = \frac{e^{\beta * w_{i,j^*}}}{\sum_{i=1}^{I} e^{\beta * w_{i,j^*}}} \tag{3.2}$$

Again, the β parameter determines agents' adherence to choosing stronger versus weaker arguments pertaining to her chosen option. As long as β does not approach large numbers, the discrete choice equation assigns all arguments a positive probability of being chosen, including those with a weight of zero. For simplicity, we assume that the value of beta in Equation 3.1 and Equation 3.2 is the same, representing a general tendency to select arguments that most strongly support the alternative an agent believes to be best, given the information she possesses.

3.2.4 Homophilous interactions

After an agent chose which argument to communicate, she decides whom to share it with. Because we are interested in the effects of homophilic interaction patterns, we assume that agents share arguments in dyadic encounters in which they preferentially interact with those of identical group membership. Interactions are regulated through a homophily parameter h which ranges from zero to one. The greater h, the more likely agents are to interact with team members from their own group. Choosing an interaction partner is operationalized as follows: whenever a sending agent x becomes active, we define all remaining team members as potential receiving agents $Y = \{y_1, y_2, \ldots, y_{N-1}\}$. Each agent y_k is assigned a similarity value s_k, which takes on the value of $h / 2 + 0.5$ if sending and receiving agent share the same identity and $1 - (h / 2 + 0.5)$ otherwise. Exactly one of the other members of the team is chosen as recipient, where the probability of choosing agent y_k as the receiving agent is given by Equation 3.3.

$$p_{y_k} = \frac{s_k}{\sum_{k=1}^{N-1} s_k} \tag{3.3}$$

When $h = 1$, homophily is maximal and the sending agent will always choose a member of her own group. When $h = 0$, no homophily is present and all agents are chosen with equal probability. Once the receiving agent has been determined, the sending agent shares the argument she picked before, and the receiving agent updates her set of arguments and beliefs. If the new argument changed the receiving agent's belief about which option is best, she communicates this change immediately with the team. Receiving agents do not forget arguments or value them differently according to recency, frequency of receival, or group membership and beliefs of the sender. The process of activating an agent, sharing an argument, and updating beliefs of the receiver represents one iteration t and is repeated until the team obtains consensus and all agents agree on which option is best. If no consensus is reached, we stop simulations after a large number of iterations ($t = 5000$) and record the belief distribution at the stop moment. We note that if the argument exchange within teams continued even after a consensus is reached, all teams would eventually find optimal consensus. Specifically, because at each iteration, arguments accessible to the sending agent and any potential receiving agent are considered with nonzero probabilities, all members in a simulated team will have access to all arguments and

agree on the best option if a simulation runs long enough. However, as we will show, it almost never occurs that all team members learn all arguments because consensus obtains (on any of the options, optimal or suboptimal) before saturation is reached. In that sense, our results show how homophily affects consensus outcomes without every team member having access to all arguments.

3.3 Setup of simulation experiments

To investigate how homophily affects decision-making quality, we vary homophily while holding all other parameters constant. For each homophily level, we simulate 5,000 teams and observe how often they obtain optimal consensus and how long it takes them to reach consensus. Homophily is varied from $h = 0$ to 0.9 in steps of 0.1. Between $h = 0.9$ and 0.98 we vary homophily in steps of 0.02 because at such high levels of homophily we found particularly strong effects on discussion time. We do not consider teams in which no arguments are exchanged across groups at all ($h = 1$) because they would represent two separate discussions instead of one.

The results section presents the outcomes of simulation experiments designed to assess six main questions. Based on competing expectations derived from the opinion dynamics and the transient diversity literature, our first and foremost question is whether homophily improves or hampers optimal decision-making. To this end, we study how often simulated teams reach consensus on the optimal, second-best or worst decision option given different homophily levels. Teams in which no consensus obtained in the time-horizon of our simulations were very rare (between 0.1% and 1% of all simulation runs with a given homophily level) and are excluded from analysis. To gain better insights into the dynamics of the deliberation process, our second question concerns how homophily affects belief changes that occur among team members. Hence, we compare for different homophily levels how often agents change their beliefs about which option is best, and which option they preferred before and after a change. Our third question is whether our model is consistent with previous claims that bounded interactions increase discussion time (D. Frey & Šešelja, 2020; J. Wu & O'Connor, 2021; Zollman, 2010). Hence, we track the number of iterations needed until a team reaches consensus and compare iteration numbers for different levels of homophily.

Our fourth result concerns the possible reasons behind homophily affecting optimal consensus making. One possibility is that homophily prevents teams from finding the optimal option because they cannot combine crucial arguments across groups. But it is also possible that homophily improves outcomes because it helps groups to find the optimal option on their own, without this process being distorted by influence from another group. To shed more light on the underlying driving effects of homophily, we compare the effect of homophilous interactions on optimal consensus-making for teams in which at least one group can infer the optimal option by themselves versus teams in which arguments from both groups have to be

combined to infer the optimal option. The fifth question asks if our model aligns with existing works arguing that homophily matters more for tasks that are challenging (Harrison et al., 1998; Phillips et al., 2006). For different homophily levels, we thus compare decision outcomes in teams where the quality of different options was hard to disentangle (high difficulty) with those where clear differences between options were easy to recognize (low difficulty). Finally, our last result relates to the question of how homophily affects decision-making quality when discussion time is limited. To elucidate this question, we present the fraction of teams with high versus no homophily that have reached optimal consensus, any consensus or no consensus after different (maximum) numbers of interactions.

To investigate each of the six questions outlined above, we simulate teams with the properties presented in Table 3.1. Our teams have $N = 6$ members, which is a common size among decision-making teams in real-world contexts and hidden profile experiments alike. They are split into $M = 2$ groups, reflecting a setting where relevant identities proxy a binary. In robustness analyses reported in Section 3.6, we also present findings for multigroup settings ($M > 2$). A total of $J = 3$ decision options is chosen and, given the distribution of common arguments, each group initially supports one of the two inferior options. The optimal option, on the other hand, is only supported by unique arguments. Because real-world decision making often involves numerous possible decision alternatives, Section 3.6 includes robustness analyses of the effects of homophily on decision quality for $J > 3$. A total of $I = 18$ arguments is available in a team (i.e., six arguments per decision option given $J = 3$) and each group starts off with $C = 3$ common arguments. There are thus 18 arguments – 2 groups * 3 common arguments = 12 unique arguments, which are distributed across the six agents. Hence, each agent initially holds three common and two unique arguments. This distribution of arguments implies a hidden profile: Common arguments point towards inferior options and agents initially believe such options to be optimal. Throughout the main analyses, we examine teams where agents probabilistically select preferred decision options and arguments according to adherence values of $\beta = 3.5$. This was found to be a reasonable value for agents to select strong arguments supporting options with higher levels of perceived quality while still allowing for small probabilities of stochastic deviation. Apart from the two, additional analyses mentioned above – homophily in multigroup settings and with many options, respectively – Section 3.6 reports on extensive robustness analyses in which we vary the number of I arguments, C common arguments, N team members and adherence values β.

As a further check of our results, we also apply our model to a situation where optimal solutions are not 'hidden' to all team members from the start. Hidden profiles are difficult because arguments pointing towards the optimal option are not commonly known to team members. However, not all real-world decision-making tasks are hidden profiles. Tasks can also be 'manifest' where agents of at least one

group have common information pointing towards the optimal alternative. We expect homophily to be less consequential for decision quality in such situations: homophily will slow down (but is unlikely to fully prevent) the team-wide dissemination of common arguments supporting the optimal alternative. Second, the suggested positive effect that homophily helps to uncover crucial hidden arguments becomes obsolete: arguments pointing to the optimal option are common arguments and do not need to be uncovered.

Table 3.1 Overview of parameters (excluding variations for robustness analyses in Section 3.6)

parameter	description	values used in simulation
N	number of team members	6
M	number of groups	2
I	number of arguments	18
J	number of decision options	3
C	number of common arguments per group	3
β	adherence to selecting stronger arguments and better options (perceived)	3.5
h	homophily	0 – 0.98

3.4 Results

We start off by comparing how often simulated teams reach consensus on each of the three decision options of different quality, given different homophily levels. As Figure 3.3 A shows, more teams form consensus on the optimal option as homophily levels increase and agents tend to interact less with team members with a different identity. The percentage of teams reaching consensus on the second-best option, on the other hand, sinks symmetrically with rising fractions of teams reaching optimal consensus. The share of teams which form consensus on the worst decision option is always below 5 percent, irrespective of the level of homophily. Considering this, and the fact that one group initially favors the second-best and the other group favors the worst option, we conclude that suboptimal consensus is most often made when one group convinces the other of the second-best option. To prevent suboptimal consensus on the second-best option, higher homophily levels could thus be helpful against the diffusion of arguments favoring this alternative from one group to the other.

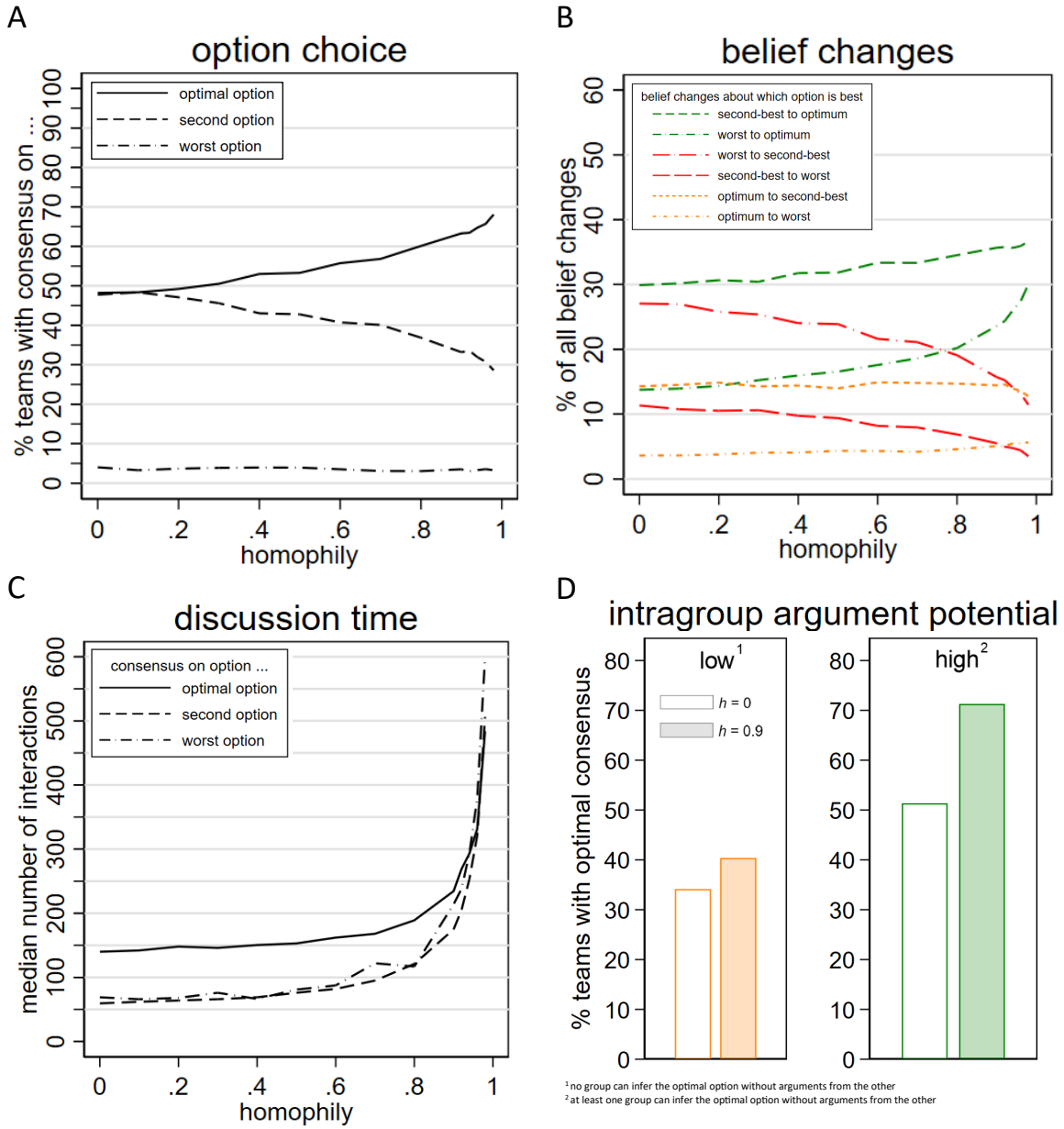


Figure 3.3 Option choice, belief changes and discussion time by level of homophily

The explanation provided here implies that as homophily increases and interactions between members of different groups become less likely, fewer agents should change their belief towards finding the second-best option optimal. Figure 3.3 B supports this conjecture, showing how the overall proportion of belief changes from the worst option to the second-best option decreases from 28 % to 12 % between $h = 0$ and $h = 0.98$. Similarly, since homophily limits communication between groups both ways, fewer arguments supporting the worst option are being passed over to the group favoring the second-best option, and fewer belief changes towards the worst option occur. Conversely, belief changes from both the second-best and the worst option towards the optimal option become increasingly frequent

in higher homophily levels. While changes towards this option are obviously necessary to obtain optimal consensus, they are not easily explained. If homophily hinders the exchange of arguments across groups, including those that point towards the optimal option, why do more team members change their belief towards the optimal option?

3.4.1 Longer deliberation uncovers optimal arguments

An explanation to this is that because homophily limits argument exchange between groups, disagreement in beliefs across groups is preserved, and neither group can convince the other of their initially preferred option. Hence, discussion continues. Figure 3.3 C illustrates this, showing how higher homophily levels result in higher median discussion time. Prolonged discussions, in turn, enable arguments favoring the optimal option (called 'optimal arguments' hereafter) to be revealed and spread within a respective group. This is so because optimal arguments are unique arguments, which need more time than common arguments to be selected. In comparison to common arguments, unique arguments face a sampling disadvantage and are initially disfavored by agents' argument selection procedure. But because this procedure is stochastic, small probabilities of choosing optimal arguments remain. When exchange between groups is limited and premature consensus kept at bay, optimal arguments are selected and spread within a group, and agents' perceived quality of the optimal option rises. Once all members in one of the groups realized what the optimal option is, they are unlikely to be swayed: optimal arguments tend to have the highest weights and are difficult to surpass by other arguments. Hence, as soon as one group has discovered the optimal option, the danger of a suboptimal consensus is minute, which gives this group ample time to still convince the other group.

This explanation implies that homophily has a much more powerful effect when at least one of the groups has sufficiently strong optimal arguments to identify the best option by themselves and without the help of the other group. Figure 3.3 D supports this conjecture, showing how homophily increases the share of teams with optimal consensus to a great extent (i.e., from 51 % with $h = 0$ to 71 % with $h = 0.9$) when the arguments initially provided to one group are sufficient to infer the optimal option. However, when neither group can infer the optimal option without the other, homophilous interactions still increase the chances of making optimal consensus, but only slightly (34 % to 40 %). The smaller effect is explained by the fact that homophily still prolongs discussion time, making it more likely that optimal unique arguments spread within groups by chance. Altogether, it follows that homophily improves consensus quality mostly because it grants one group with the time to uncover unique arguments and arrive at the optimal option. At the same time, it hinders another group from quickly convincing the team to prefer a suboptimal option through the dissemination of inferior arguments.

3.4.2 Homophily is crucial when tasks are especially difficult

In Figure 3.3, we have shown that homophily improves decision-making quality because it prevents teams from prematurely adopting a second-best decision option. To further test this proposition, we investigate if optimal consensus is less likely when the second-best and the optimal option are close to each other in quality, and therefore hard to distinguish, and if the worst option has much lower quality than the second-best option and is therefore likely to be neglected quickly. For each simulation run, we compute a distance score reflecting the difference in quality between the optimal and the second-best option, and the second-best and the worst option. If the mechanism works as we have suggested, a *smaller* quality difference between the second-best and the optimal option should make it easier for the group initially supporting the worst option to get persuaded into the second-best option, thus reducing the chances of reaching optimal consensus. Similarly, a *large* difference in quality between the second-best and the worst option should make it easier for the second-best group to convince the worst group of the second-best option. In both cases, homophily should have a bigger effect because it is needed more to prevent suboptimal consensus on the second-best option.

In Figure 3.4 A, we present the percentage of homophilous versus non-homophilous teams with optimal consensus by their differences in quality between the optimal and the second-best option, and the differences between the worst and the second-best option. As suspected, a smaller difference between the optimal and the second-best option increases the chances of finding optimal consensus. Greater differences between the second-best and the worst option, on the other hand, lead to lower proportions of optimal consensus. Comparing the fraction of populations with optimal consensus under high homophily versus no homophily, the previous finding persists that higher homophily levels render more populations with optimal consensus. This supports our explanation and shows that the positive effects of homophily we observe generalize to a wide set of different combinations of option quality.

However, Figure 3.4 B also indicates that the positive effect of homophily varies among problems with different quality combinations across options. In line with our proposed mechanism, homophily appears to matter especially for those problems where chances to obtain optimal consensus are low to begin with. Here, homophily provides the crucial barrier to the team-wide adoption of second-best arguments that are dangerous precisely because they are either almost as strong as arguments supporting the optimal option, or because arguments pertaining to the worst option are weak in comparison. As becomes evident from the figure, the increase in teams obtaining optimal consensus under high homophily levels is largest when differences between the second-best and the optimal option are small, and differences between the worst and the second-best option are large.

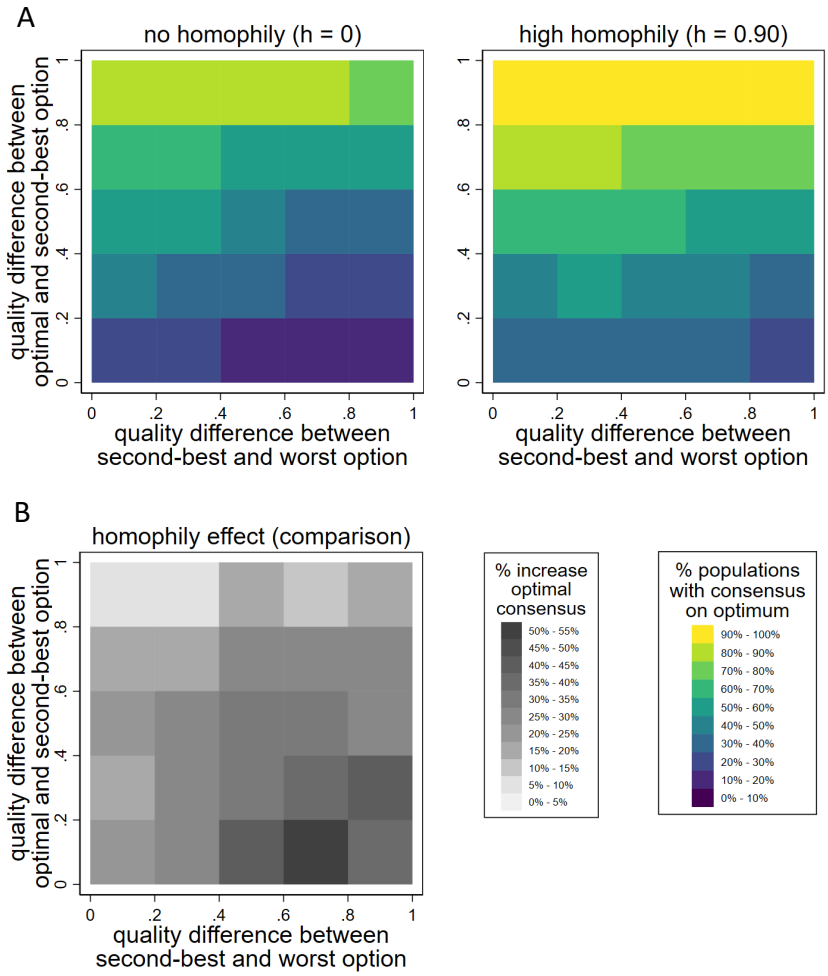


Figure 3.4 Optimal consensus by quality difference in options[7] and homophily level

3.4.3 Homophily facilitates optimal intragroup consensus

The results from Figure 3.3 suggest that homophily fosters optimal consensus by granting one of the groups the crucial time to exchange their optimal arguments and, in a subsequent step, convince the rest of the team. To further elucidate this mechanism, Figure 3.5 shows the fraction of teams with high versus no homophily that have reached optimal consensus across the whole team after a given number of interactions, and the first occurrence of optimal consensus within any of the two groups. If the mechanism works as described, we should find that optimal consensus within either group is more frequent, and occurs sooner under high homophily than in teams with no homophily.

A comparison of teams with and without homophily reveals that at any discussion length, more teams will have reached consensus on the optimal option in either

[7] Data grouped by quintiles of the distribution of distances scores to ensure that each cell represents at least 200 observations.

group when homophily is high (dashed lines in Figure 3.5). This finding aligns with previous results that homophily facilitates optimal consensus within a group without influence from the other group (compare Figure 3.3 D). However, teams with high homophily also need more time from the first occurrence of optimal consensus within a group until team-wide optimal consensus is established. The reason behind this is that homophily also slows down the sharing of arguments from the group with consensus on the optimal option to the other group. This contributes to increased discussion time under homophily.

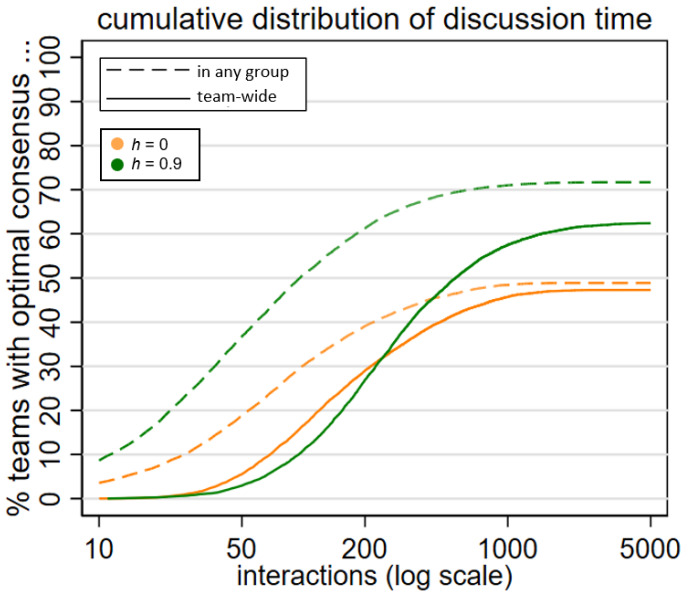


Figure 3.5 Discussion time until the first occurrence of optimal consensus in any group (dashed lines) and until optimal consensus in the whole team (solid lines), by homophily level

For both teams with high homophily and no homophily, most teams that reach optimal consensus in a group also establish optimal consensus on a team level. Some teams, however, fail to achieve optimal consensus despite having reached optimal consensus in a group before. This becomes apparent from the fact that also at $t = 5000$ interactions – a cutoff point at which more than 99% of teams have established any consensus – there remains a vertical difference between the dashed and the solid lines. Suboptimal team consensus despite previous optimal intragroup consensus can materialize because unless all optimal arguments have been uncovered already, belief changes within team members from the optimal to a suboptimal option can still occur. Surprisingly, this happens more frequently in teams with high homophily. The explanation behind this is that while one group may have found the optimal option, homophily tends to tighten intragroup consensus in the other group, enabling them to establish a strong belief that another (suboptimal) option is best. As discussions go on, the group with optimal consensus fails to receive

additional optimal arguments from the other group that would be needed to strengthen their belief. Receiving other arguments instead, group members revoke their belief and suboptimal consensus is made. In sum, however, this tendency is insufficient to offset the mechanisms that optimal consensus in a group predates and fosters optimal consensus in the team, both of which occurring more often under homophily.

3.4.4 Homophily and decision quality in manifest profiles

In the above analysis, we have shown that homophily is especially important when a hidden profile task is difficult to solve. However, in real-world decision tasks it is rarely clear if a hidden profile is at hand in the first place. Common arguments do not necessarily need to support inferior options, they can also point towards the optimal option. In this case, multiple team members start off with optimal beliefs and a 'manifest profile' is present. Given that homophily only has strong positive effects when the task is difficult, can we still assume homophily to foster optimal consensus-making when a task is an easy manifest profile?

Intuitively, because optimal arguments are common arguments in one group and therefore likely to be shared early in the discussion, homophily hinders the communication of optimal arguments across groups. Hence, we expect that optimal consensus obtains less frequently. Figure 3.6 A shows that this is indeed the case, albeit only to a small degree. The fraction of teams making optimal consensus decreases only slightly from 95 percent to 88 percent over the full range of the homophily parameter. This can be attributed to ceiling effects – when problems are very easy to solve, homophily will only delay the deliberation process but cannot thwart optimal decision–making. Just like in hidden profiles, homophily leads to longer discussion time. However, simulation runs ending in optimal consensus need slightly less time than those with suboptimal consensus (Figure 3.6 B). This is because arguments pointing towards suboptimal options may be present only in the form of unique arguments, which comes along with more time necessary to uncover and share them.

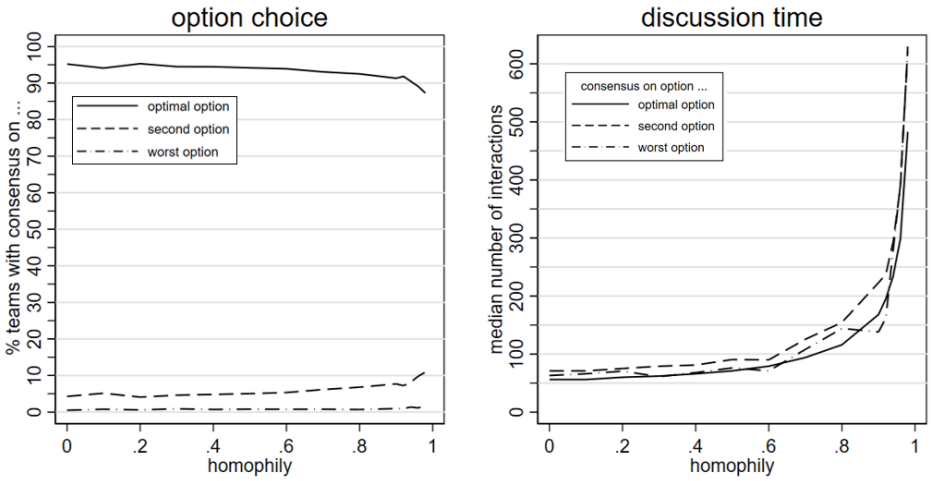


Figure 3.6 Option choice and discussion time in manifest profiles

Altogether, our analyses reveal that in hidden profiles, homophilous interactions along common group identities substantially improve collective decision-making quality. Additional robustness analyses reported in Section 3.6 reveal this result to be remarkably robust. When profiles are manifest, disadvantages in decision quality are limited. In both profiles, higher levels of homophily prolong discussion time, especially at the upper end of the homophily parameter range.

3.5 Discussion

Our model identifies homophilic interactions in diverse teams as a key factor to the quality of a team consensus in difficult decision-making tasks. Preferential interaction with similar others prevented team members from being convinced by inferior solutions proposed by those with identities different than their own. As discussions continued, this gave room for communicating crucial arguments that would not have been shared otherwise. This finding resonates with the notion of 'transient diversity' (Zollman, 2010). However, our investigation extends this notion to the new realm of hidden profile problems and revealed a mechanism that had not been considered by the transient diversity literature before. While transient diversity models find that unbounded communication leads to the insufficient *generation* of diverse information, our hidden profile task showed that even when all information had been created prior to the task, its *distribution* could lead to unfavorable situations in which non-homophilous interactions resulted in suboptimal decisions. Extensive robustness analyses reported in the following section reveal that this finding extends to different team sizes, argument distributions, and multigroup settings.

Our results thus run contrary to expectations suggested by the opinion dynamics literature – namely, that homophily will undermine team functioning. Regardless,

our simulations feature two central commonalities with prominent opinion dynamics models (Flache et al., 2017; Hegselmann & Krause, 2002; Mäs et al., 2013): homophilic interactions prolong discussion time, and they produce disagreement. Different from opinion dynamics research, however, disagreements had a transitionary character that eventually enabled better decisions. Many opinion dynamics models produce irresolvable and lasting intergroup dissent because they have different behavioral assumptions: agents do not search for optimal solutions but strive to maintain consensus with ingroup members and differentiate themselves from outgroup members. In consequence, emerging factions increasingly distance themselves from each other to such an extent that the odds of interaction between them become zero. Integrating assumptions of differentiation from those with different opinions, a possible opportunity for 'model docking' with opinion dynamics models (Axtell et al., 1996) and extension to this model is to condition homophilic encounters on endogenously changing interaction preferences based on beliefs. While this may result in unresolvable disagreements, the necessity to study alternative means for making decisions arises. A common approach in real-world teams is to rely on voting procedures and other aggregation rules when failure to obtain consensus is immanent (Levy, 2007). Hence, studying decision quality while assuming voting procedures within the context of this model, or an adapted version thereof, provides a promising avenue for future research. Our results also run contrary to those of Estévez-Mujica et al. (2018) who found homophily to limit problem-solving potential in diverse groups. We attribute this divergence to the fact that their task resembled a manifest profile more than a hidden profile: because optimal pieces of information were relatively easy to recognize by agents but needed to be disseminated across groups, homophily hampered performance.

Considering that in our model, individuals truthfully communicate the information they consider most valuable and process information in an unbiased manner, the question arises how homophily also improves team decision-making when individuals are less rational. Previous works on hidden profiles identify anchoring heuristics, recency biases, and needs for social validation that can lead individuals to report and process information less accurately (Stasser & Stewart, 1992; Stasser & Titus, 2003). While such biases are not considered by our model, it is unlikely that fully accounting for them would fundamentally change the mechanism by which homophily improves decision-making. Due to the setup of the hidden profile and in line with empirical research, behavioral heuristics make it less likely that unique information is shared and accepted (Lu et al., 2012; Wittenbaum et al., 2004). Making the task at hand more difficult, accounting for them would likely amplify the effect that homophily has. In a similar vein, future research may consider status distinctions or different group sizes that would introduce inequality in the influence that one group has over another. Here, homophilous interactions could again provide a crucial mechanism to improve deliberation tasks that would

otherwise have been dominated by the group with the greatest influence. This notion is supported by robustness analysis included in Section 3.6, showing that homophily is especially important when the group initially supporting the second-best option is larger in size.

A feature that is inherent to hidden profiles is that individuals ultimately share the same goal and are likely to agree on one option to be best when faced with complete evidence. While this is applicable to many real-world situations, it abstracts from the possibility that team members of different identities may have group-based interests that make them attach different values to decision options, or even attach value to maintain disagreement with other groups. Such a case would make it necessary to redefine what an 'optimal' solution is and poses interesting distributive and ethical questions. While outside the scope of this paper, an extension of this model could be used as a starting point to investigate whether homophily is helpful in reaching decisions that maximize welfare for the team as a whole versus solutions that optimize payoffs for some groups of team members at the expense of reduced team performance.

For the purpose of this paper, probabilistic encounters between team members represented homophilic interaction *preferences*. However, the same encounters can also be seen as a manifestation of underlying *social foci* that structure team deliberation (Feld, 1981). Translating the insights of this paper to such a perspective implies that better decisions will be made in settings where team members are structurally guided to interact with similar over dissimilar others more frequently. This resonates with suggestions made by the transient diversity literature (Zollman, 2010). Simultaneously, studying the effects of interactions in structurally embedded environments calls for a possible extension of our model in which encounters are not probabilistic but occur along a network that specifies who exchanges information with whom. If the mechanism proposed here holds, networks in which members of different groups are increasingly kept apart should also feature better decisions.

In a more general sense, the model presented here can also be seen as an example of a larger class of phenomena exhibiting puzzling and often unexpected social change. Similar to how information transmission in networks is highly sensitive to 'percolation thresholds' (Newman & Watts, 1999), residential segregation can emerge from minor preference shifts (Schelling, 1971), and spread of new attitudes critically depends on the way supporters of new beliefs are spatially located (A. Nowak & Vallacher, 2019), our model exhibits phase transitions where unlikely sharing of optimal information can determine failure or success to obtain optimal consensus.

While the simulation results reported here convincingly show how homophily fosters the quality of the team decision, it is important to note that homophilous interactions may have other, unintended consequences. Limiting interactions

between members of different identities may amplify social identity processes that can lead to negative outgroup attitudes, lower levels of trust, and less cooperative behavior in general (Carter & Phillips, 2017; Lau & Murnighan, 1998). This raises the question whether improved decision-making can be reached through alternative means. Within the context of our model, such means could involve increased skepticism towards information coming from dissimilar members. However, this would involve that spillover effects from increased skepticism resulting in negative outgroup attitudes had to be tempered all the same. Similar to the extension suggested in the paragraph above and in line with simulation research suggesting that the timing of outgroup contacts matters crucially for multigroup discussions (Flache & Mäs, 2008), an alternative to achieve improved consensus while minimizing negative affective consequences is to structurally embed conversations. For example, deliberation could be broken up into phases where groups are first kept apart and given enough time to uncover crucial information without influence from other groups, and only then brought together to find a consensual solution. In addition to future computational experiments, the effectiveness of such an intervention could easily be tested in an empirical setting where teams with a structurally embedded deliberation procedure likely made better decisions than those without.

Lastly, the finding that keeping groups apart has positive effects on decision quality relies to some degree on at least one of the groups having sufficient information to infer an optimal solution by themselves. This provides a scope condition for the mechanism found by this paper, but also raises the important question whether homophily can be helpful in cases where only a specific conjunction of arguments from different groups can reveal the best solution. While exceeding the scope of this paper, this calls for a promising model extension in which this is addressed more explicitly – namely, a model where a complex underlying function enables certain argument combinations to have nonlinear impacts on team members' beliefs.

In conclusion, the work presented here provides a novel insight on how to better shape interactions in diverse teams with regards to their decision-making abilities. When tasks are difficult, unbounded communication among team members can cause cognitive diversity to pose a liability. In such cases, homophilous interactions improve decision quality because they keep individuals from convincing dissimilar others with their suboptimal responses too quickly. Pointing to a trade-off between decision quality and efficiency, homophily also resulted in increased discussion time. This is an important finding to consider when finite resources have to be weighed against convex returns to optimal over inferior solutions.

3.6 Appendix

In our main results, analyses pertained to small teams with no more than two groups, three decision options, and a limited number of arguments. Here we investigate if the positive effect of homophilous interactions on decision quality also persists under different parameter settings. In each of the panels of Figure 3.7, we measure discussion time and the fraction of teams obtaining optimal consensus under high (h = 0.9), medium (h = 0.5) and no homphily (h = 0), and vary one of the default parameters described in the section 'Setup of simulation experiments'. Finally, we vary one component not originally covered by the variable parameters in the model: the proportion of the group initially supporting the second-best option relative to the size of the whole team. All analyses are conducted to investigate if our findings hold in more generalized ecological environments. The finding that homophily improves discussion quality turns out to be remarkably robust across all analyses.

Figure 3.7, Panel A shows that regardless of the number of total available arguments I, homophily improves decision quality. This is because the mechanism by which homophily operates – enabling one group to find the optimal decision while limiting influence from the other – stays the same, regardless of the number of arguments. However, as visible from the decreasing fractions of teams with optimal consensus, a greater number of available arguments also makes problems more difficult to solve correctly. As more arguments pertain to each option and weights pertaining to arguments are drawn randomly, the law of great numbers makes the true quality scores of the different options more similar and therefore harder to differentiate. More arguments also increase discussion time because there are more possible argument sets team members can hold to form beliefs: in turn, longer deliberation is needed to align everyone's beliefs.

From Figure 3.7, Panel B it becomes evident that a greater number of common arguments per group leads to decreasing fractions of teams with optimal consensus but amplifies the effect of homophily. This is because an increasing number of suboptimal common arguments will lead to their increased circulation and hence, chances rise that one group will be able to convince the other of a suboptimal option. This makes homophilous interactions even more important, providing the necessary barrier to the diffusion of inferior arguments across groups.

In Figure 3.7, Panel C we show that the positive effect of homophilous interactions on decision quality persists under a rising number of available decision options J. To ensure that each decision option had a constant number of available arguments pertaining to it, we increased the number of total available arguments by six in each additional option (hence, $I = J * 6$). As J rose, the effect of homophily on optimal consensus-making decreased slightly but remained positive. We attribute this to the fact a greater number of options diversifies team members' sets of unique arguments and weakens the correlation between initial beliefs and group membership. This weakens the effect of keeping groups separate. Additionally,

greater *J* increased discussion time, which is intuitive because an increasing number of options makes it increasingly difficult to align all team members' beliefs about which option is best.

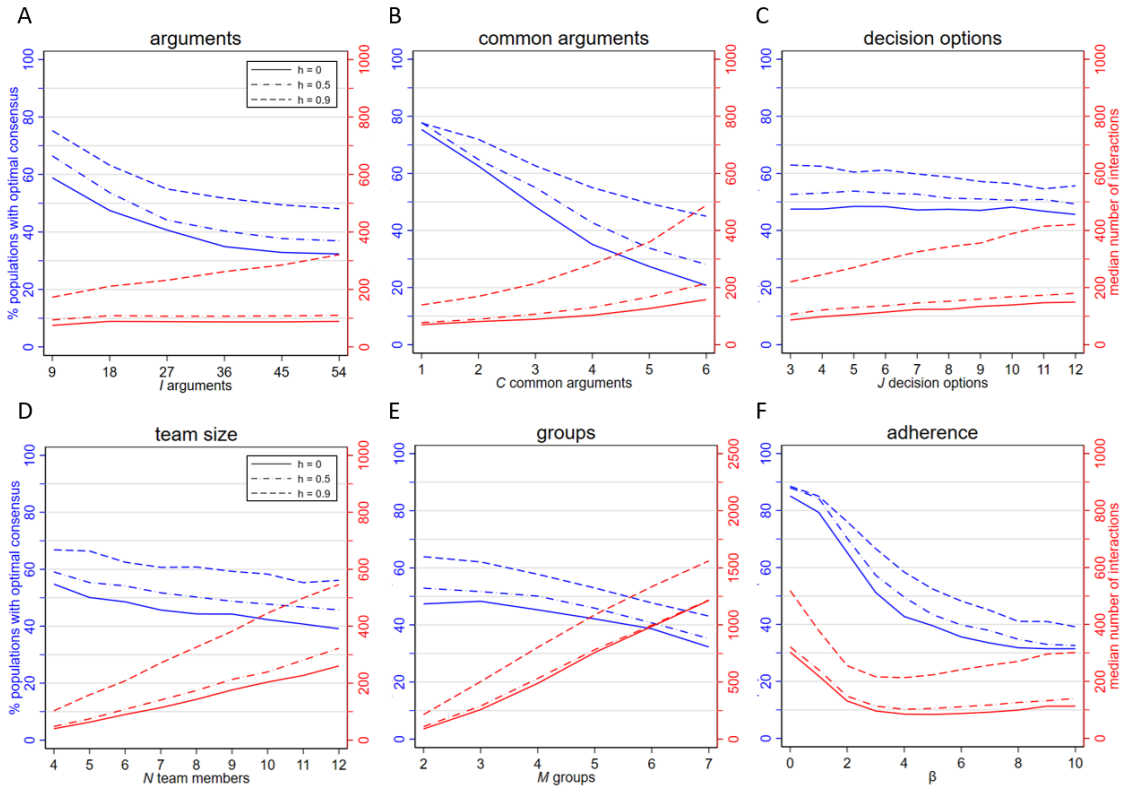


Figure 3.7 Optimal consensus-making and discussion time for variables *I*, *C*, *J*, *N*, *M* and β under high, intermediate, and no homophily

Figure 3.7, Panel D confirms that homophily improves decision quality irrespective of the number of team members. As team size increases, the fraction of teams with optimal consensus sinks while the positive effect of homophilous interactions becomes larger. An explanation to this is that as *N* rises while the number of available arguments *J* and common arguments per group *C* are kept constant, each agent holds fewer unique arguments relative to their common arguments. Similar to the effects found in Figure 7B, this leads to an increase in common arguments that are circulating, which in turn makes homophily even more important to prevent suboptimal consensus. In addition to the fact that it takes longer to align everyone's beliefs when the team is larger, the increase in (redundant) common information that circulates prolongs discussion time.

Figure 3.7, Panel E shows that homophilous interactions increase the fraction of teams with optimal consensus also when more than $M = 2$ groups are present. To ensure that each group featured sufficiently many members, we increased team size in steps of three with each group added. Because each group needs to hold common information different from those of other groups, we added one additional decision

option with each group, and six arguments with each option (hence, $N = M * 3$; $J = M + 1$ and $I = J * 6$). As the number of groups and decision options rose, chances decreased that one group alone had the arguments necessary to infer the optimal option without information from others and hence, the effect of homophily was smaller when the number of groups was large. Given that additional groups, arguments, and options also increased the overall complexity of the deliberation process, discussion time increased greatly.

Figure 3.7, Panel F reveals robust homophily effects under varying levels of the adherence parameter β by which agents select a decision option to support and an argument favoring this option. The homophily effect is smaller at low levels of β, which is due to a ceiling effect. Most teams reach optimal consensus also in the absence of homophily. The high fraction of teams with optimal consensus results from the fact that lower β values result in an increasingly random selection of arguments. This helps optimal unique arguments to be uncovered. High β values, on the other hand, cause quasi-deterministic sharing of arguments that support agents initial yet inferior beliefs. Discussion time is maximal when β is lowest because randomly selected arguments allow for greater diversity in beliefs, which prevents consensus and delays the deliberation process. Discussion time slightly increases again at high β because increasingly deterministic argument sharing leads to greater redundancy in argument sharing, which is insufficient to convince one group of another group's preferred option. This leads to lack of consensus, especially when interactions are homophilous.

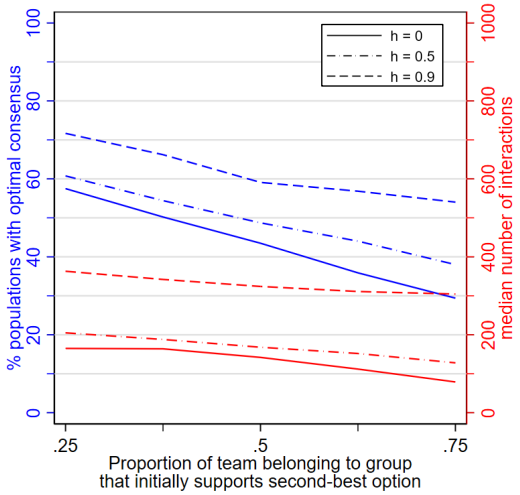


Figure 3.8 Optimal consensus-making and discussion for teams of unbalanced group sizes under high, intermediate, and no homophily

Figure 3.8 presents the effect of homophily for teams of unbalanced group sizes. All simulated teams have eight members and the size of the group initially supporting the second-best option ranges between two and six members. All other

parameters have the default values reported in the main paper. As becomes evident from the figure, homophily makes optimal consensus-making more likely throughout the imbalance levels tested. Optimal consensus is least frequent and homophily effects strongest when the fraction group of members initially supporting the second-best option is large. This is so because creating more team members that prefer the second-best option gives this option additional support, which increases the risk of a second-best consensus. To prevent this, homophily is especially important.

Lastly, Figure 3.9 presents the main results from Figure 3, but this time using simulated teams where the sending agent is not activated sequentially but randomly. A comparison with Figure 3.3 reveals that results are virtually identical, leading us to conclude that the model is robust to potential statistical artefacts resulting from a fixed order activation of agents.

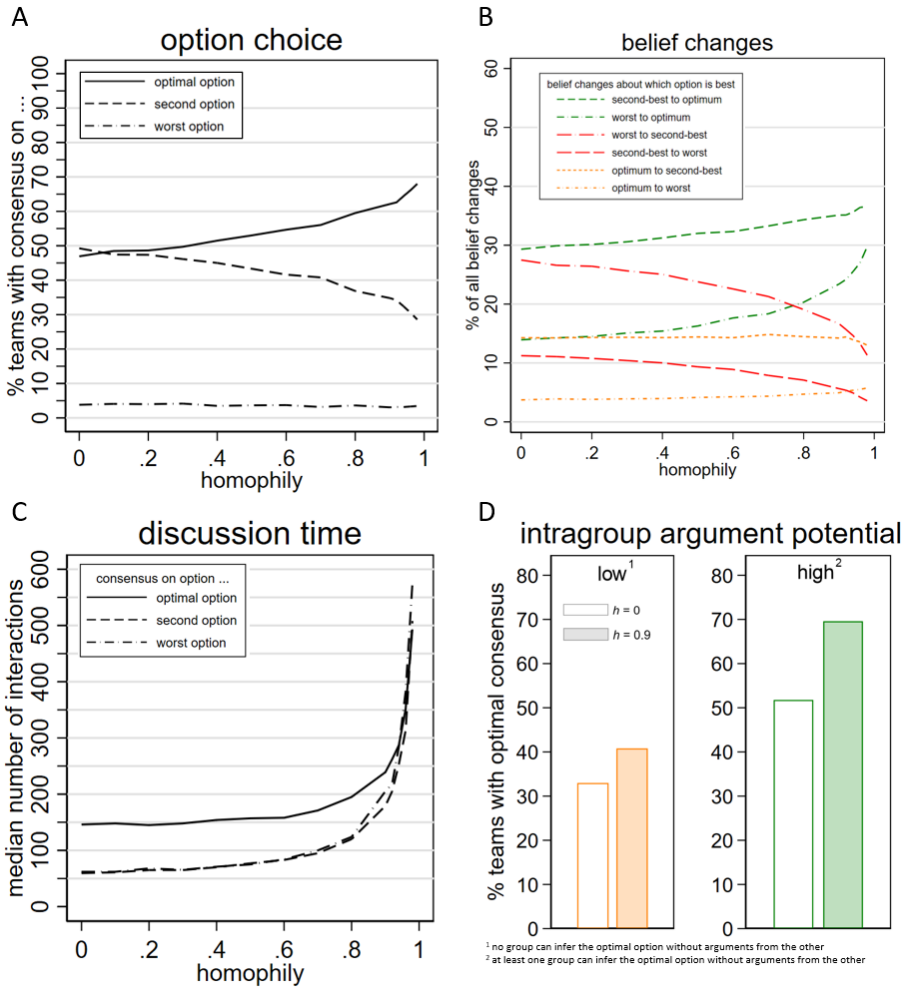


Figure 3.9 Option choice, belief changes and discussion time by level of homophily using random sender activation.

Chapter 4

Perceived cognitive differences facilitate complex social learning

This chapter represents joined work with Vincenz Frey and Maxime Derex. A revised version of it is under consideration at an international journal.

Abstract

Humans learn from each other, enabling them to innovate tasks by integrating their own approaches with those of others. Ample literature shows that individuals are more likely to observe and integrate unfamiliar practices if they come from people who are similar to themselves. Here we conjecture that this is indeed true when strong identities inflate the perceived competence of ingroup peers and lead to the rejection of behavior of dissimilar others. However, if comparisons are made along cognitive characteristics instead, dissimilarity can create expectations of novelty and individuals preferentially learn from those who are different from themselves. We test our expectations in a preregistered experiment in which 859 individuals develop suboptimal approaches to a problem-solving task and are subsequently exposed to the behavior of a fictional demonstrator. The demonstrator pursues a different but likewise suboptimal approach; and participants can learn to optimize their approach by integrating it with that of the demonstrator. A 2x2 experimental design varies whether observed behavior comes from a demonstrator with shared or dissimilar characteristics, and whether (dis-)similarity relates to ideological leaning or the outcome of a bogus 'cognitive style test.' Participants optimized their approach most often after exposure to someone with a different 'cognitive style,' even though participants expected to learn most from an ideologically similar peer. Participants also indicated a preference to observe ideologically similar and cognitively dissimilar sources. Our findings challenge the common assumption that people learn best from those who are similar to them, suggesting that perceived cognitive differences can create expectations of novelty and foster observational learning.

4.1 Introduction

Humans are avid social learners and solve problems by integrating their own approaches with those of others (Derex & Boyd, 2016). And yet, successfully bringing together different practices can be difficult to achieve. Not only does any recombination involve greater complexity than simple copying. It also involves potentially many trial-and-error attempts, and it forces individuals to closely observe others' behavior. Humans have limited attention and cognitive resources, making observation costly. This makes them often reluctant to study some else's behavior in the first place. Anchoring heuristics pull individuals to remain with approaches they deem 'good enough' (Furnham & Boo, 2011), and risk aversion keeps people from exploring alternatives, especially when their performance is difficult to observe (Kahneman & Tversky, 2012). Hence, individuals frequently rely on heuristics and proxies in their choice to pay close attention to someone else's behavior or not. A prominent proxy in this choice are other, easily observable characteristics of the person demonstrating new behavior, and how they relate to the observer.

A wide range of literature shows that similarity between demonstrator and observer fosters attention, even if this similarity is unrelated to the task at hand. Humans seem to preferentially learn from and adjust their behavior towards others with shared attributes (Smaldino & Velilla, 2025). From a very young age, children are more likely to mimic same-gender peers (Chartrand & Lakin, 2013), and preferentially learn languages from people with the same native dialects (Kinzler, 2021). A bias towards similar others continues among adults, who learn cultural knowledge from same-age peers (Reyes-García et al., 2016), place more trust in the buying offers of sellers with the same ethnicity (Levine et al., 2014), and are more easily swayed by same-partisan peers to correct their false interpretations of climate trends (Guilbeault et al., 2018).

Reasons for the widespread prevalence of similarity-biased social learning can be summarized as follows: The first reason relates to the expected usefulness of new behavior. Individuals often navigate complex and challenging environments, and these environments can differ from person to person (Aoki & Feldman, 2014; Boyd & Richerson, 1988; McElreath et al., 2013). Likeness on some trait can be an indication that someone else's behavior will be applicable to oneself. Say that, for example, an individual needs to decide what game to hunt. By observing the hunting strategies of others with similar traits, such as their equipment or physical condition, the person tries to maximize chances that the observed strategies will also be successful for themselves.

Second, aligning behavior often comes with coordination benefits (Ellison, 1993). For example, speaking a certain language is only useful if one's interaction partner also understands and speaks it. Higher probabilities of interaction often correlate with shared social categories, such as gender or ethnicity (McPherson et al., 2001).

Provided that shared categories imply increased chances of interaction, this can be a reason to learn behavior from similar others and adopt it accordingly (Cooper, 1999). In the same line of reasoning, increased interaction chances can also facilitate adaptation towards mutually cooperative behavior (Efferson et al., 2008; Gross et al., 2023; M. A. Nowak, 2006), implying that interactions with members of the same category are less likely to be exploited.

In addition, social-psychological factors can influence the decision to dedicate attention to similar others. We like to think of ourselves positively, and, by extension, of those we have much in common with (Tajfel, 2010). This can make others with a shared identity appear more competent, and hence individuals will have greater expectations that their solutions are successful and viable. This is demonstrated by a large experimental literature on social identity (Ellemers et al., 2002) and conformity (Cialdini & Goldstein, 2004), showing that even arbitrary and ad-hoc categories can lead to favorable evaluations of (similar) ingroup members and the devaluation of dissimilar others.

However, similarity-biased social learning can also have drawbacks. Because many of the tasks we face are complex and multi-faceted, they involve a wide range of possible solutions and oftentimes require costly trial-and-error to improve (March, 1991). Limiting one's observation to the behavior of similar others makes adjustments less risky, but it will also mean that not all approaches are considered. In the long run, and as research on polarization shows (Bail et al., 2018; Hegselmann & Krause, 2002; Mäs & Flache, 2013), preferential interactions with similar others can lead to clusters with correlated opinions, preferences, and behaviors (DellaPosta et al., 2015) and hinder the transfer of potentially valuable knowledge and skills (Ertug et al., 2022). Hence, clusters often represent stable local optima because an existing array of solutions is 'good enough', while finding a better one would involve searching and adjustment without a guarantee of success.

Yet, the very fact that behavior often tends to cluster in individuals with shared characteristics could also be a reason to strategically seek out different individuals instead. If people associate the categories they belong to with certain behaviors - which is a frequent assumption among humans (Schultner et al., 2024) - they may deduce that observing someone similar to them renders redundant interactions. Observing dissimilar individuals, on the other hand, could be a way to gain access to valuable behaviors that are not available in the local environment. In other words, someone's different characteristics could be used as a proxy for learning potential and novelty (Homan et al., 2007); and positively influence willingness to observe their solutions. Such a strategy should also be favorable over individual trial-and-error, which offers the most opportunities for novel exploration, but also renders a high risk of failure (Lazer & Friedman, 2007; McElreath et al., 2013).

Nevertheless, few studies investigate whether perceiving someone as different can enhance openness to learn from them. Studies on dyadic influence show that

status discrepancies (V. Frey et al., 2024), and reputation effects (Baldini, 2013) enhance influence, but these imply a hierarchical difference between sender and receiver. Individuals listen to others because they have information making them appear more successful or competent, and not simply 'different' from themselves. However, some experiments on group decision-making suggest that visible differences could enhance learning: Phillips (2003) and Phillips & Loyd (2006) show that demographic differences make it more likely that novel information is taken seriously and not ignored or forgotten, and work on mock jury trials (Sommers, 2006) demonstrates that racially diverse groups exchange a wider range of information and make fewer errors. Similarly, Antonio et al. (2004) find that ethnic minority members have a positive influence on how much information groups can remember and consider simultaneously. It hence appears that demographic differences help to surface a diverse array of information, which in turn has a positive influence on the decisions groups make. However, these works focus on the role of information exchange, demographic differences and discussion quality in groups, leaving unaddressed if people seek out and preferentially learn from dissimilar others to adjust their behavior in response to an individual problem as well.

In this study, we theorize that dissimilarity can indeed foster behavioral adaptation, but only if observed differences lead to expectations of learning potential. The mechanism for this is as follows: in complex environments that are difficult to navigate, people anticipate that others pursue strategies that may be complementary to their own. In the absence of immediate indicators of performance, individuals rely on easily observable characteristics to build beliefs about the viability of their behavior. If people connect someone's characteristics with the strategies they pursue, dissimilarity will lead to expectations of novelty, and they will observe this person more closely. Similarity, on the other hand, will lead to expectations of redundancy and diminish the attention paid to that person. The argument we make here connects to earlier research on small group research, suggesting that visible diversity stimulates people to think in unfamiliar ways (Homan et al., 2007; Phillips, 2003; Phillips & Loyd, 2006).

Our conjecture rests on two important scope conditions. Dissimilarity-biased social learning should mostly occurs in situations where payoffs do not depend on strategic interaction with others. Otherwise, similarity may be associated with frequent interaction and more cooperative behavior, and the expected advantages arising from aligning behavior with ingroup peers can outweigh potential benefits of learning from dissimilar others. Second, it must be clear that observable characteristics do not influence the payoffs a behavior generates. If a certain approach is only viable for someone with specific characteristics (say, the physical condition to engage in hunting of a certain animal), then the behavior of dissimilar others will be ignored.

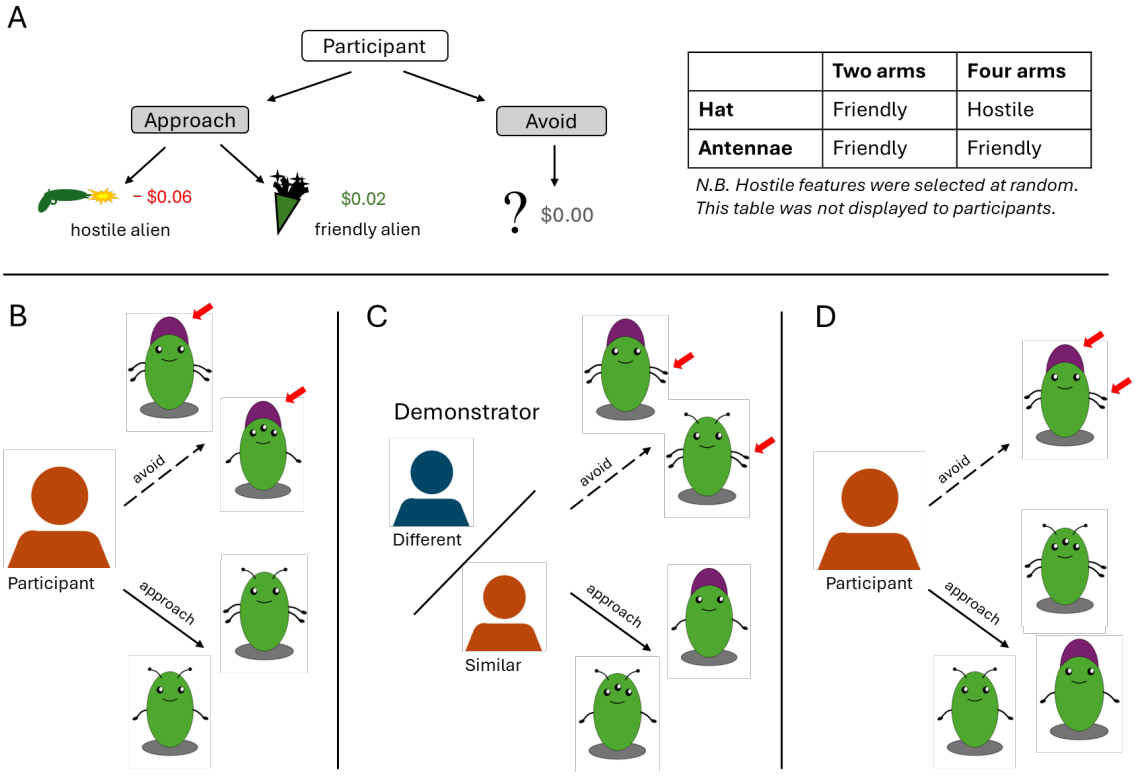
We also argue that even under the above scope conditions, dissimilarity-biased social learning will not always be present. Namely, if observable characteristics are tied to a strong notion of social identity, outgroup devaluation makes dissimilar others appear less competent and overpowers associations of behavioral novelty. If someone upholds positive stereotypes about their ingroup – for example, that they are decent and sensible people – and negative stereotypes about an outgroup – that they are irrational people whose actions are driven by false beliefs – then they will brush off unfamiliar behavior from dissimilar others as nonsensical and ignore it. This conjecture integrates research on identity-based social learning (Guilbeault et al., 2018; Zou & Xu, 2023).

Our aim is to empirically compare the conditions under which behavioral adaptation becomes similarity-biased versus dissimilarity-biased. In doing so, we integrate two competing strands of literature and shed new light on the occurrence of social learning. We test our expectations in a preregistered task where participants must learn to navigate an unknown two-dimensional environment. They are first led to develop suboptimal approaches and are subsequently exposed to the behavior of a fictional demonstrator who pursues a different but likewise suboptimal approach. Through close observation and trial-and-error, participants can realize the flaws in their own and the demonstrator's behavior and optimize their approach. As outlined in the next section, a 2x2 experimental design varies whether the behavior is demonstrated by someone with shared or dissimilar characteristics, and whether (dis-)similarity relates to an identity-based category or 'cognitive traits' that could be linked to their behavior. Identity is represented by political ideology, which is known for its strong connection to affective devaluation of opposite-partisan others (Iyengar et al., 2019; L. Mason, 2018; West & Iyengar, 2022; Westfall et al., 2015). Cognitive traits are based on a fictional 'cognitive style' test that uses image recognition, and was designed specifically for the context of this study. The outcome of the cognitive style test aims to maximize associations that it is related to someone's behavior, while simultaneously giving no indication of overall performance. The actual behavior of the demonstrator was kept constant across conditions, giving all participants equal opportunities to learn from the demonstrator. In the Results section, we compare across experimental treatment how often participants integrated their own behavior with that of (dis-)similar others, thereby optimizing their approach. The Discussion section compares our findings in light of existing research.

4.2 Experimental design

Testing our expectations requires a controlled comparison of two aspects: first, a social learning task in which demonstrator characteristics imply a strong, preconceived notion of group identity on the one hand versus novel characteristics that could be linked to a behavioral strategy. The second aspect involves a comparison of learning from dissimilar versus similar others. We implement both in

a 2x2 preregistered experimental design with 859 US participants recruited via Prolific.[8] In the first phase of the experiment, participants were exposed to a stream of experiences that led them to develop suboptimal and stable solutions to a problem-solving task which we adapted from Rich & Gureckis (2018). In the second phase, participants continued with the same task but could observe the behavior of a fictional demonstrator pursuing a complementary approach to their own. The demonstrator possessed characteristics that made them similar or dissimilar to the participant, and these characteristics either related to the outcome of a bogus 'cognitive style test' or self-reported political ideology (liberal / conservative). A final phase measured whether participants optimized their approach or failed to integrate the information provided to them.



	Two arms	Four arms
Hat	Friendly	Hostile
Antennae	Friendly	Friendly

N.B. Hostile features were selected at random. This table was not displayed to participants.

Figure 4.1 Task and procedure. We instructed participants to 'approach friendly aliens and avoid hostile ones'. Approaching friendly (hostile) aliens resulted in financial reward (punishment), avoiding aliens did not affect payment. Aliens were only hostile if two features were present (**A**). Participants developed behaviors where they avoided all aliens where one hostile feature was present (**B**). We introduced a demonstrator with dissimilar (Cond. A) or similar characteristics (Cond. B). The demonstrator avoided all aliens with the second hostile feature and approached aliens if this feature was not present (**C**). Through close observation, participants could integrate behaviors and learn to only avoid aliens with both features (**D**).

[8] Preregistration, survey and replication package available at: https://osf.io/fjtac/?view_only=d1e9be0558b9443cac7490cfc9b9e985

4.2.1 Participant recruitment

We recruited 1600 US participants by posting a targeted advertisement on the online crowd working platform Prolific. To maximize reliability of ideological identity, we only invited participants with the following criteria: they either had to have voted for Donald Trump in the 2020 US national election, identified as conservative and considered themselves republican (25% of the sample, or 400 participants), or they had to have voted for Joe Biden, identified as liberal, and considered themselves democrats (75%, 1200 participants). Ideological leaning was checked and verified during the experiment (see 'Experimental conditions'). The study was advertised as a 'Task about pattern recognition and learning'. Upon arrival at our experimental platform, participants received instructions and gave informed consent.

4.2.2 Approach-avoid task

The study consisted of an approach-avoid task adapted from Rich & Gureckis (2018). We showed participants pictures of fictive creatures we called 'aliens', one picture at a time, and let them decide to avoid or approach the alien by clicking on a corresponding button. As illustrated by the left side of Figure 4.1 A, participants were instructed to approach friendly aliens, in which case they were rewarded 2ct, and avoid hostile aliens, in which case their bonus payment stayed unaffected. If they approached a hostile alien, we subtracted 6ct from their bonus payment. Whenever a participant approached an alien, they were immediately informed about whether the alien was friendly or hostile. We did not reveal if an alien was friendly or hostile if participants avoided an alien. Aliens differed along three binary dimensions (antennae or hat, two or four arms, three or two eyes). One dimension was selected as irrelevant. The other two dimensions made an alien hostile if a combination of two specific features was present, for example, four arms *and* hats; otherwise, the alien was friendly (Figure 4.1 A, right side). Hostile features and relevant dimensions were selected at random. Through trial-and-error, participants could learn to recognize the features that made an alien friendly or hostile and thereby improve their bonus payment.

Participants were compensated with $2.00 upon successful completion, plus a bonus payment according to performance between $0.00 and $1.20. The study was approved by the ethics committees of the University of Groningen and the review board of the Institute of Advanced Studies, Toulouse. Data were collected between December 2, 2024 and January 10, 2025.

4.2.3 Learning traps

Participants made approach/avoid decisions on maximally eleven 'blocks' of eight aliens each. During the 'private learning phase' consisting of the first five blocks, participants developed 'learning traps' (Rich & Gureckis, 2018): that is, they identified one feature that had to be present for an alien to be hostile but failed to

identify the second feature (Figure 4.1 B). Learning traps serve as a persistent and common representation of behavior that is not optimized but stable (Liquin & Gopnik, 2022; Spreng & Turner, 2021): Individuals avoid all hostile aliens (where both features were present) but also avoid some friendly aliens where the first feature is present but the second is not. Because avoiding aliens means that their nature remains unknown, participants do not learn that some are in fact friendly.

The first two blocks were designed to lead participants into a specific learning trap: In these blocks, we excluded aliens where the first hostile feature was present, but the second one was not. Because of this, participants were exposed to a stream of experiences where aliens were always hostile when the first feature was present, and therefore quickly arrived at the impression that this was the only feature that mattered. In all subsequent blocks three and following, we presented aliens in a random sequence without replacement such that each block featured all 8 different aliens. We classified behavior as consistent with a learning trap when a participant avoided all aliens with one relevant feature and otherwise approached aliens for 8 out of 8 decisions within one block. If a subject's behavior was consistent with a learning trap in block three or four, they immediately continued with block six.

4.2.4 Experimental conditions

Prior to block six, each participant was told that they would be 'matched with an earlier participant' [i.e., a fictional demonstrator], shown 'what decisions the earlier participant made when they encountered the same aliens as you' and that 'in some cases, observing their behavior could help you earn a higher bonus'. Before being exposed to the demonstrator's behavior, participants indicated their political ideology and took part in a 'cognitive style test'. Representing experimental conditions, we later used ideology and the outcome of the test as characteristics that would make participants similar or dissimilar to a demonstrator. Cognitive style was measured by presenting participants with an ambiguous image of the rabbit-duck illusion[9] and asking them what they saw first in the picture (a rabbit or a duck). We instructed participants that cognitive style was "related to how people perceive patterns in the alien task" but left them otherwise uninformed about demonstrator performance. We assessed political ideology using a standard 7-point self-identification question ranging from 'extremely conservative' to 'extremely liberal' (ANES, 2020). Participants whose self-reported ideology did not match their pre-screened ideology were excluded from the study.

We then assigned participants to experimental conditions at random and as illustrated in Figure 4.2. In condition 1A, we told participants that the demonstrator had a 'different cognitive style' as themselves, and in condition 1B that the

[9] The rabbit-duck illusion is a popular ambiguous image dating back to the late 19th century. For our experiment, we used a modern and less well-known adaptation available at https://www.ocf.berkeley.edu/~jfkihlstrom/images/Jastrow/DuckRabbitStudios.png.

demonstrator had the 'same cognitive style'. In doing so, we exposed participants to cognitive characteristics that could either be associated with behavioral novelty or redundancy. In condition 2, the demonstrator either had a different political ideology than the participant (Cond. 2A), or the same ideology (Cond. 2B).

We subsequently presented participants an information screen with the characteristics of the demonstrator. The screen was followed by an 'exposure phase' with blocks six to nine, in which we exposed participants to the behavior of the demonstrator, along with a graphical representation of their own and the demonstrators' characteristics. Demonstrator behavior was chosen such that it was consistent with a different learning trap, i.e. focused on the second feature the subject had not yet identified. For example, and as illustrated in Figure 4.1, if a participant avoided all aliens with hats, the demonstrator avoided all aliens with four arms. Following close observation and further trial-and-error, participants could recognize their own learning trap and that of the demonstrator. Integrating their own approach and the demonstrator's approach enabled participants to optimize: i.e., avoid all hostile aliens but approach all friendly aliens. Experimental procedures, including the demonstrator's behavior, were identical across conditions. Only the demonstrator's characteristics were varied.

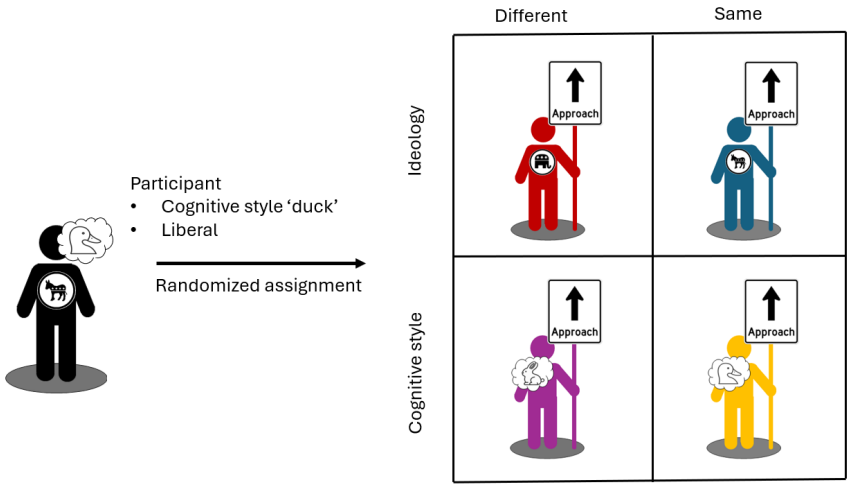


Figure 4.2 Experimental conditions. Participants completed a fictional 'cognitive style test' and self-reported their ideology (left side). They were then matched with one of the four different demonstrator types (right side), representing experimental treatments.

4.2.5 Optimization

The final two blocks 10 and 11 represented an 'assessment phase' in which we measured whether a participant successfully integrated their own approach with the approach of the demonstrator. These blocks included no information on the behavior of the demonstrator and assessed if the subject correctly identified the

combination of two features that would negatively affect their payment (Figure 4.1 D). We classified behavior as optimized if participants identified 8 out of 8 aliens in one block correctly. Participants could finish the experiment early if they optimized their approach from block 8 onward, in which case they were directly routed to a study completion page informing them about their payment. After the study, a detailed debriefing informed all participants that the demonstrator was not a real person and that 'cognitive style' was a fictional concept.

4.2.6 Sample of analysis

We aim to assess when individuals integrate their own suboptimal behavior after observing the behavior of someone else. This requires that individuals developed a suboptimal solution on their own in the first place, which was not the case for all participants. As pre-registered, we therefore restrict the final sample of analysis to those participants who had identified a suboptimal solution before they were exposed to information about the demonstrator's behavior. That is, for participants to be included in our sample, they had to have developed the specific learning trap we designed for them prior to block six: they avoided all aliens with the first of the two relevant features that made an alien hostile and otherwise approached all aliens for 8 out of 8 decisions within one block. This was the case for 859 out of 1600 participants recruited for the task. Of the remaining 741 participants, 180 optimized early, meaning that they identified both relevant features before being exposed to the demonstrator. 77 participants developed an 'unintended' learning trap, i.e., they identified the second hostile feature instead of the first, which meant that participant behavior was identical with demonstrator behavior, rendering exposure to the latter redundant. 471 participants followed no discernible approach, and 13 were excluded because they failed to verify their pre-screened ideology correctly, tried to participate twice, or had implausible completion times.

In the final sample of analysis, average completion time was 10 minutes and 55 seconds. Participants made decisions on 68 aliens on average and were correct in 79 percent of their choices. At an average bonus payment of $0.66 and $2.00 compensation for successful completion, mean hourly compensation was $14.63. The median age of the sample was 38 years and 57 percent identified as female. Participant demographics in the final sample were statistically indiscernible from the 741 individuals who were not included in it.

4.2.7 Data quality

We undertook the following steps to ensure high data quality. First, we only recruited online participants with an approval rating above 99 percent, meaning that they had to have completed nearly all of their previous tasks on Prolific successfully. Second, we included timers on all instruction screens to avoid that participants did not read them carefully. Third, we verified participants' political ideology on our experimental platform, excluding individuals whose pre-screened

ideology did not match their ideology when we asked them again. Fourth, we included an attention check during the experiment, asking participants in block seven if they could recall the decision of the demonstrator to approach or avoid the alien they were previously exposed to.

In line with our preregistration, the final sample of analysis also included participants who failed this attention check, which was the case for 14.4 percent of participants. This was done to exclude the possibility that we falsely excluded participants who executed their task carefully but had decided to disregard demonstrator behavior either because of their characteristics, or because they were completely immersed in their learning trap. For sensitivity reasons, Section 4.5.2 presents the main analyses without participants who failed the check, showing that results stay similar and retain their significance at conventional levels.

4.3 Results

Figure 4.3 presents an overview of participant behavior in our final sample of analysis, pooling across experimental conditions. 859 participants developed a learning trap before the end of the private learning phase, i.e. they identified one feature that indicates hostility before block six, avoiding all aliens with that feature and approaching aliens otherwise. 495 of these participants retained their learning trap at the end of the experiment (57.6%) and were not influenced by the demonstrator. During the exposure phase in blocks six to nine, between 6.9 percent and 8.9 percent of participants left their learning trap and instead adopted the demonstrator's approach. However, this behavior shrunk back to 2.4 percent in the assessment phase, suggesting that participants either copied the demonstrator without learning the behavioral rule behind it, or returned to their own learning trap because they failed to integrate approaches.

Optimized behavior started to appear among 12.1 percent of all participants at the start of the exposure phase and rose steadily to 30.7 percent at the end of the experiment. It persisted after demonstrator behavior was made unavailable during the assessment phase (block 10 onwards), suggesting that some participants understood the two-feature rule that made aliens hostile and did not simply use an intuitive combination of demonstrator behavior and their own. Within the context of our experiment, exposure to demonstrator behavior was sufficient to interrupt stable learning traps and led to optimization among almost one third of participants, even though the demonstrated behavior was not optimized itself.

As preregistered, we compare optimization after exposure to similar versus dissimilar demonstrators, and separately for the cognitive style conditions and the ideology conditions. If participants associated cognitively different and ideologically similar demonstrators with learning potential, and observed them more closely, then this is where optimization would occur most frequently. Figure 4.4 A provides

an overview of this comparison. Unless indicated otherwise, test results are derived from two-sided randomization tests.

In line with our expectations, more participants optimized after exposure to a cognitively different demonstrator (36.7%), as compared to only 25.1 percent of participants who optimized after observing a cognitively similar demonstrator (p = 0.006, N = 426). Keeping in mind that demonstrator behavior was the same across conditions, this shows that an indication of cognitive dissimilarity tipped participant behavior towards close observation and subsequent optimization. Ideological (dis−)similarity, on the other hand, did not affect how often participants integrated observed behavior: More participants optimized after being exposed to a demonstrator with a similar as opposed to a different ideology (32.5% versus 28.3%), but this difference was not statistically significant (p = .41, N = 433).

To gain more insight into the mechanism behind observational learning from (dis-)similar others, we implemented two additional measures. First, we asked participants to indicate their agreement to the statement 'I think that I can learn something from the earlier participant' [i.e., the demonstrator] after being exposed to the demonstrator's approach/avoid decisions for one block (i.e., 8 aliens). If cognitively dissimilar and ideologically similar demonstrators are perceived as more competent, more participants would agree they can learn from them. Similar to our analysis of optimization, we compare perceived learning potential separately for the cognitive style conditions and the ideology conditions.

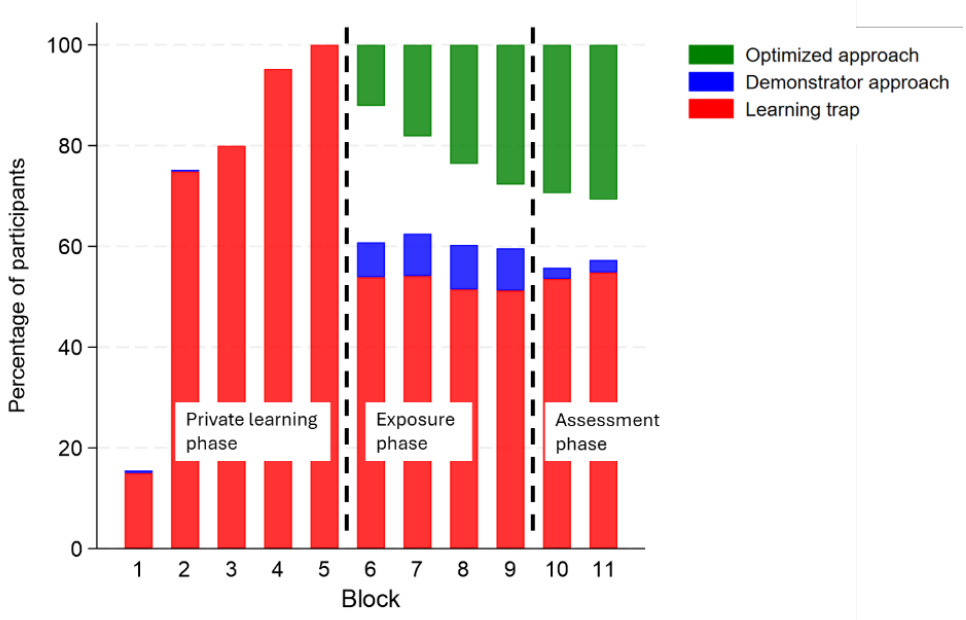


Figure 4.3 Overview of participant behavior. Participants developed learning traps during the private learning phase. Optimization started shortly after the introduction of a demonstrator in the exposure phase. In the assessment phase, demonstrator copying diminished when participants made decisions without access to the demonstrator's decisions, while optimized approaches persisted.

Figure 4.4 B shows that, as expected, participants more often agreed or fully agreed if they were exposed to an ideologically similar demonstrator as opposed to a demonstrator from the opposite ideological camp (48.7% versus 37.6%, $p = 0.025$, $N = 433$). This was the case even though demonstrator behavior was identical across conditions, underpinning that ideological differences alone were sufficient to impact impressions of competence. A greater proportion of participants (fully) agreed that they could learn from a demonstrator with a different rather than a similar cognitive style (53.0 versus 46.0 percent), though the result did not reach statistical significance ($p = .168$, $N = 426$).

Second, before exposing participants to the demonstrator, we asked them if they preferred 'to be matched with someone with the same cognitive style or a different cognitive style', and someone 'with the same ideology or a different ideology'. According to our proposed mechanism, participants would prefer a cognitively dissimilar and an ideologically similar demonstrator if they expected greater learning potential from them. Because demonstrator preferences were indicated by all participants before the exposure phase, and independent from randomized condition assignment, we simply test over the whole sample if preferences were statistically different from chance.

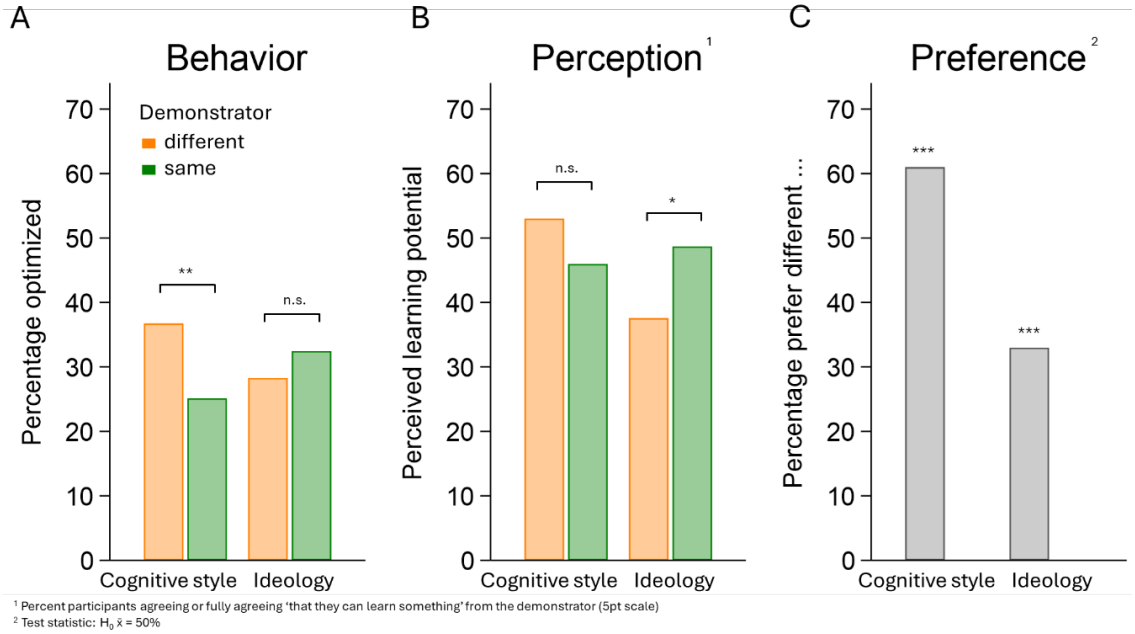


Figure 4.4 Dissimilarity-biased social learning arises from perceived cognitive differences but not from ideological differences. Participants optimize their solution more often after being introduced to a cognitively different demonstrator (A). They also attribute greater learning potential to ideologically similar peers (B) and prefer learning from cognitively different and ideologically similar demonstrators (C). n.s. $p \geq 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

As illustrated in Figure 4.4 C, the majority of participants (60.7%) preferred a demonstrator with a different cognitive style over one with the same style (H_0 $\bar{x}$ = 50%; $p < 0.001$, $N = 859$). For political ideology, the pattern reversed: Only 33.4 percent preferred a demonstrator from the opposite ideological camp ($p < 0.001$, $N = 859$). Considering that participants were incentivized to make correct decisions and told that the demonstrator could be helpful in earning a higher bonus, this finding suggests that they projected greater financial rewards onto learning from cognitively different others and ideologically similar peers.

We then tested if higher rates of optimization in the cognitive dissimilarity condition were independent from whether participants' preferences were met, which is important for the following reason: Figure 4.4 C showed that most participants preferred ideologically similar others. Hence, due to random assignment, more participants in the cognitive dissimilarity condition (1A) had their preferences fulfilled than in the cognitive similarity condition (1B). This could have led to a larger number of disappointed participants in the cognitive similarity condition, spuriously causing lower rates of optimization. We tested this possibility with a logistic regression on the sample of participants in the cognitive style conditions ($N = 426$). Optimization was the dependent variable, and met preferences and similarity treatment were the independent variables. The analysis revealed no significant effect for met preferences ($\beta = -.07$, $p = .743$) while the effect of cognitive similarity persisted ($\beta = -.57$, $p = .011$). This shows that cognitive dissimilarity fostered optimization, irrespective of whether participants sought it out or not.

Similarly, higher agreement to perceived learning potential in the ideological similarity condition was unaffected by whether participants' preferences were met. We selected the sample of participants in the ideology conditions ($N = 433$) and conducted a logistic regression with agreement as the dependent variable (binary, "agree" or "fully agree" = 1). Met preferences and similarity were included as independent variables. The effect of similarity remained positive ($\beta = .55$, $p = .010$) while the effect of met preferences was not ($\beta = -.25$, $p = .245$), giving no indication of confounding influence.

In sum, our results reveal the mechanism behind the occurrence of dissimilarity-biased social learning. Exposure to behavior from a cognitively dissimilar demonstrator resulted in more optimization, was tied to expectations of greater learning potential, and was preferred by participants. Ideological leaning was accompanied by a preference for similar others and an enhanced perception of competence of same-partisan peers, but ultimately did not affect optimization.

Although not in the primary focus of our study, the question arises if similar effects can be found among participants who did not develop a learning trap on their own in the first place. To this end, Section 4.5.2 includes a replication of Figure 4.4 on the sample of participants with no discernible approach before block six

(N = 471). The results reported there show that participants preferred ideologically similar demonstrators but otherwise no significant treatment effect of cognitive or ideological (dis-)similarity. Potential reasons for a lack of treatment effects are discussed ibidem.

4.4 Discussion

In this paper, we examined the conditions under which individuals improve their suboptimal solutions after observing someone else take a likewise imperfect but useful approach. We showed that, as hypothesized, learning and optimization increased when demonstrator characteristics pointed towards cognitive differences, whereas ideological leaning did not affect whether participants optimized. Our findings identify dissimilarity to foster attention and integration of unfamiliar behaviors, but only if this dissimilarity leads to expectations of novelty and learning potential. In contrast to empirical literature showing that people tend to adopt behaviors from similar others (Chartrand & Lakin, 2013; Guilbeault et al., 2018; Reyes-García et al., 2016), our findings present a more nuanced and optimistic view of human learning. Rather than merely following ingroup members, humans are capable of valuing dissimilar others more highly when they anticipate greater learning opportunities from them.

In emulating a task where a combination of solutions from different individuals produced a better outcome, our research sheds new light on an evident puzzle in the literature. On the one hand, our findings resonate with those of Phillips (2003) and Levine et al. (2014), suggesting that visible differences can foster openness to thinking in unfamiliar ways. But they also connect to work that highlights diversity's role as a 'double-edged sword' (Milliken & Martins, 1996; Phillips & O'Reilly, 1998): Diverse populations perform better because of a greater wealth of perspectives and skills (Hong & Page, 2004; Page, 2019) but identity diversity can make it challenging to fully use their potential (Flache & Mäs, 2008; Lau & Murnighan, 1998). In line with this, our study showed that individuals disfavored ideological difference and attributed lower learning potential to it.

To overcome the challenges associated with different identities (Guilbeault et al., 2018; van der Does et al., 2022), previous research has explored ways to build trust and understanding among dissimilar peers. One possibility involves emphasizing 'deep-level' similarities, such as shared interests and values (Peters, 2021; Phillips & Loyd, 2006). However, the fact that in our experiment, participants optimized less often when introduced to a 'cognitively similar' demonstrator suggests that highlighting deep-level similarity can sometimes be counter-productive. In a globalizing world where populations become increasingly diverse, our findings imply that practitioners should not try to emphasize similarity at all costs but make those differences salient that enable individuals to expect novelty and learning potential from each other.

Research on surface- versus deep-level diversity (Harrison et al., 1998, 2002) inspires us to theorize how results would have been affected in a design where participants were able to observe several characteristics simultaneously. It seems natural that ideological affiliations and cognitive characteristics would interact in such a way that people preferentially integrate behavior from others who are ideologically similar but also cognitively different. However, other alternatives are thinkable, calling for further empirical investigation. For example, depending on the setting and the salience of different stimuli, cognitive dissimilarity could also cancel out effects of ideological differences, or vice versa.

An important consideration is whether our results are not only valid within dyadic learning tasks but also translate to group settings. Groups naturally constitute much more complex interactive environments but also carry potential to amplify the effects found here: Groups could exacerbate identity-driven similarity bias because individuals want to signal alignment with ingroup members (van der Does et al., 2022). At the same time, groups could also make ingroup homogeneity appear more obvious and therefore lead individuals to seek out different others (Phillips, 2003). The complexity of group dynamics suggests that before we can meaningfully theorize about the effects of diversity at the group level, we must first understand how individuals interact in pairs. This study contributes to that understanding.

Group settings also bring forth the question of how learning was affected if the viability of a solution depended on who else is using it. Many human tools and cultural practices involve coordination benefits (Henrich & McElreath, 2003). This can make it more advantageous to strategically align actions with ingroup members if those are the ones we are primarily interacting with (Gross et al., 2023; McPherson et al., 2001). In line with the scope conditions outlined in the introduction, and to test our proposed mechanism of dissimilarity-biased social learning in a clean and parsimonious manner, we did not consider such possibilities in the present study. However, advantages through coordination with ingroup members versus expected gains from interactions with outgroup members could also be tested within the same experimental setting, for example through research that features strategic interdependencies in larger populations (Ehret et al., 2022) or networked environments (Centola, 2010).

In addition, the focus of our study laid on observational learning, which included that behavior was always transmitted truthfully, and with little room for misunderstanding, misrepresentation, and noise. Real-world interactions may involve strategic reasons to obfuscate certain behaviors, both to ingroup members and for reasons of 'brokerage' (Stovel & Shaw, 2012), but also to outgroup members to protect themselves from exploitation (M. A. Nowak, 2006).

Another question that remains is why in our study, perceived cognitive differences enhanced social learning while ideological similarity did not. We

attribute this to the fact that our experimental design excluded factors that amplify similarity bias, such as coordination advantages, peer sanctioning, or a task where the viability of different approaches may be related to the social categories someone belongs to (Aoki & Feldman, 2014; McElreath et al., 2013). We do not exclude the possibility that social psychological causes of identity are sufficient to trigger behavioral adaptations. However, these causes may instead be more applicable to settings where incentives to learn unfamiliar behavior are less strong (Guilbeault et al., 2024), and in more natural settings of human interaction, such as the exchange of opinions in online social networks (van der Does et al., 2022)

Our study is circumscribed by the fact that we were primarily concerned with an optimization task where individuals had already identified a suboptimal solution. Section 4.5.2 of the Appendix presents an analysis of participants without a discernible approach and compares the conditions under which these individuals learned to adopt observed behavior. We did not find that perceived ideological similarity or cognitive differences enhanced learning in this sample, likely because the experimental paradigm we used was not designed for it, and because this sample includes many participants who were not paying much attention at all. Future research should investigate if dissimilarity can also foster attention and influence if individuals had not identified a viable approach in the first place.

Lastly, while research on the evolution of culture has identified reasons for why technological boundaries tend to coincide with social boundaries (Flache, 2018; Roux et al., 2017), our study provides a novel mechanism that can help understand how innovative practices sometimes do emerge from interactions across groups. We hope that the mechanism we identify here will stimulate new avenues in research, such as integrating dissimilarity-biased learning into agent-based models of cultural evolution and experimental studies.

4.5 Appendix

4.5.1 Sensitivity analyses

This section provides sensitivity analyses of the main results, selecting only those participants who passed the attention check (see section 'data quality'). 735 participants, or 85.6% of the sample correctly recalled the behavior of the demonstrator in a previous decision in block seven and therefore passed the check. Figure 4.5 gives an overview of results on this sample. As becomes evident from the figure, results stay largely unaffected by the subsample selection. Similar to the main results, more participants optimized their approach after exposure to the behavior of a demonstrator with a different cognitive style (41.4%) as compared to the same style (26.5%, $p = .004$, $N = 362$). Participants also agreed more often that they could learn from an ideologically similar demonstrator (51.3%) as compared to an ideologically different one (39.3%), and this difference was statically significant ($p = 0.026$, $N = 373$). Lastly, 62.3 percent prefer a cognitively dissimilar demonstrator (H_0 $\bar{x} = 50\%$; $p < .001$, $N = 735$), while only 32.8 percent prefer an ideologically different one (H_0 $\bar{x} = 50\%$; $p < .001$, $N = 735$). Repeating regression analyses from the main analyses on this subsample revealed that differences in perception and behavior were unaffected by whether participants' preferences for a (dis-)similar demonstrator were met (preference met, perception: $\beta = -.23$, $p = .298$; behavior: $\beta = -.002$, $p = .994$).

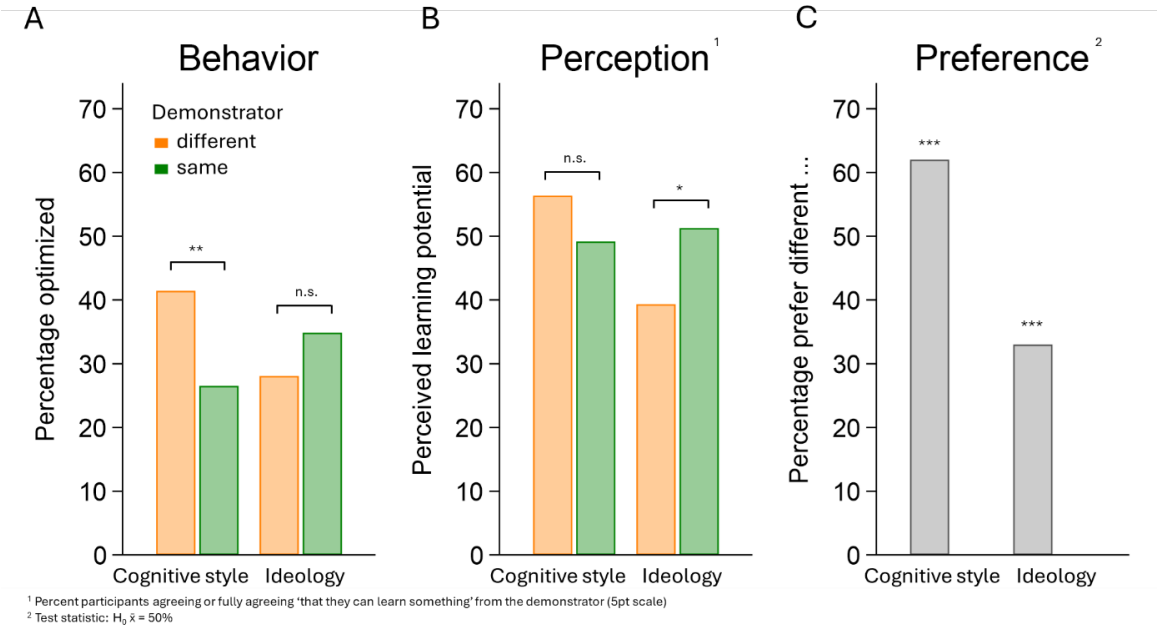


Figure 4.5 Main results on subsample of participants who passed the attention check ($N = 735$).

4.5.2 Analyses on sample without learning trap

This section investigates if participants who did not develop a learning trap before the exposure phase in block six were nevertheless influenced by cognitively or ideologically (dis-)similar demonstrators. We conducted the same analyses as in Figure 4.4, but this time on the sample of participants with no discernible approach (N = 471). Figure 4.6 A shows that exposure to demonstrators with different characteristics did not lead to significant differences in the percentage of participants who optimized their approach, neither in the cognitive style conditions (dissimilar 13.7%, similar 14.4%; p = .873, N = 235), nor in the ideology conditions (dissimilar 18.8%, similar 11.1%; p = .105, N = 236).

Overall, 75.6 percent of participants agreed or fully agreed that they can learn from the demonstrator, but Panel B shows that perceptions of learning potential did not differ between participants exposed to a cognitively dissimilar versus similar demonstrator (76.9% vs. 72.0%; p = .392, N = 235), or ideologically different versus similar demonstrator (78.1% vs. 75.0%; p = .574, N = 236). Panel C illustrated that participants preferred ideologically similar demonstrators (66.9%, H_0 $\bar{x}$ = 50%; p < .001). However, unlike our main sample of analysis, participants had no preference for cognitively dissimilar demonstrators (53.9%, H_0 $\bar{x}$ = 50%; p = .088).

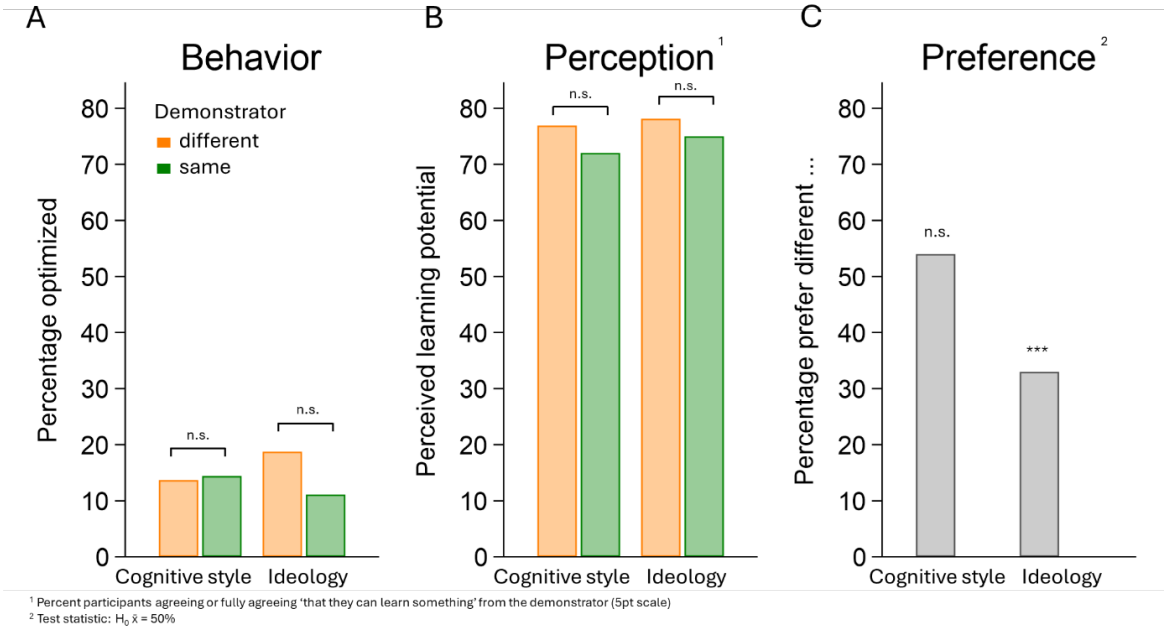


Figure 4.6 Main results for participants who did not develop a learning trap before exposure to the demonstrator (N = 471).

Despite the lack of differences in optimization rates, the possibility remains that participants without a discernible approach at least learned to adopt the demonstrator's approach later on, and that this was more often the case for participants who were exposed to ideologically similar and cognitively different demonstrators. To test this possibility, we compared across experimental conditions how often participants adopted the demonstrator's learning trap in the assessment phase in block 10 and 11. The analysis revealed that overall, only 14.4 percent of participants were able to do so. We did not detect any statistically discernible differences in adoption rates across experimental conditions: 16.2 percent of participants took over the observed learning trap when the demonstrator had a different cognitive style, as opposed to 12.7 percent with the same cognitive style (p = .543, N = 235). 11.7 percent adopted the learning trap from an ideologically different demonstrator compared to 15.7% from an ideologically similar one (p = .478, N = 236).

We attribute a lack of treatment effects across experimental conditions to the following reasons: First, the sample is much smaller, meaning that analyses could be underpowered. Second, optimization is much harder without viable approach in the first place – overall, only 14.4 percent of participants in the sample optimized their approach by the end of the experiment. This meant that a lack of different optimization rates across treatments could have been caused by a floor effect. Lastly, overall data quality was likely to be lower in this sample, either because the group of participants without a discernible approach was inflated by participants who did not pay enough attention to the task from the beginning, or because they became frustrated with failing to identify a solution on their own.

Chapter 5

How argumentation styles and preference perceptions affect deliberation outcomes in groups with conflicting stakes

This chapter is published as: Stein, J., Romeijn, J. W., & Mäs, M. (2025). Ill-informed Consensus or Truthful Disagreement? How Argumentation Styles and Preference Perceptions Affect Deliberation Outcomes in Groups with Conflicting Stakes. *Erkenntnis*, 1-26.

Abstract

In groups where members deliberate with limited information, consensus can emerge where, under complete information, fundamental disagreement would prevail. Using an agent-based model, we explore the factors contributing to group consensus by comparing argumentation styles in two types of groups: agents in groups of *advocates* communicate arguments for options perceived as personally beneficial. Agents in groups of *diplomats* do the same but avoid disagreement in that they bring up arguments supporting a second-best option whenever their interaction partner perceives to benefit the least from what the sender finds best. Results show that consensus depends on argumentation style, but also on what members initially perceive as preferred. Diplomats are more likely to form consensus when initial perceptions accurately align with full information preferences, which diverge within the group. Conversely, and perhaps counterintuitively, in the presence of inaccurate initial perceptions, groups of advocates converge while diplomats part in disagreement. Our results imply that the ideal argumentation style must be considered carefully in light of both the desired outcome and the initial information distribution: when conflicting stakes produce a trade-off between consensus and truthful perceptions, polite versus selfish ways of deliberation may produce one or the other outcome, depending on the initial information members are equipped with.

5.1 Introduction

Since James March's seminal work on 'Exploration and exploitation in organizational learning' (1991), much research has investigated the notion that optimal group performance requires a delicate balance between individual's efforts to seek new solutions and their ability to adopt existing approaches (Bernstein et al., 2018; Lazer & Friedman, 2007; Levinthal, 1997; W. Mason & Watts, 2012). Across the contexts studied, a common denominator is that individuals do not exhibit conflicting stakes, in that the benefit of a solution is the same to everyone. Disagreement among group members arises from heterogenous access to information, but not from group members evaluating the same information differently.

Yet, in many groups, different solutions have inherently different qualities to individual group members. Here, disagreement about a 'best' solution may prevail even when everyone is faced with the same information: For example, members of a hiring committee may prefer different qualities in candidates, and therefore disagree on the choice of the right applicant. In the same vein, organizational board members can have strategic motives to weigh executive decisions differently, or the assessment of a political decision may depend heavily on what stakeholder is involved in it.

In such situations, decisions are often obtained through voting. However, unless the decision is unanimous, voting requires that everyone accepts the decision irrespective of their own preferences. Alternatively, some or all actors must compromise or bargain, knowingly disregarding what they prefer the most in exchange for an alternative choice. Such compromises can be difficult to obtain as they depend on actors' willingness to sacrifice and may come along with prolonged negotiations, tension, and conflict (Priem et al., 1995).

Here, we point to another possibility: ill-informed consensus. In the process of deliberating arguments pro and con alternative decision options, there may be a phase when all group members agree on the same option, simply because they are not aware of all existing information. If the group makes the decision at this moment, it would be based on a consensus despite diverging preferences. This approach is different from compromising because group members do not settle for an option they deem suboptimal. Instead, every group member individually comes to the – warranted yet faulty – conclusion that a given decision option is best for themselves, given the information they currently possess.

We propose an agent-based model to investigate such an emergence of consensus despite diverging preferences. In the model, agents communicate arguments to form and adapt perceptions of their preferences over three decision options. These agents are categorized into two subgroups, each with distinct stakes that, under full information, lead them to prefer different options. Preferences are modeled as a zero-sum situation: the more one subgroup gains from an option, the less does the other and hence, the greater the divergence of preferences. A third option represents

middle ground in the sense that it is a second-best choice with equal value to everyone. A group reaches consensus when deliberation has led everyone to perceive that they prefer the same option. Notably, since truthful (full information) preferences diverge, consensus can only be present when at least some group members have inaccurate perceptions.

Our paper is structured as follows. In the next section we motivate our modeling approach and spell out our specific goals with it. The model itself is presented in the next section, and in section 5.4 we specify the simulations that we carry out on this model. Section 5.5 contains the results of the simulations, and these are further discussed and evaluated in section 5.6. Finally, section 5.7 summarizes the main conclusions.

5.2 Motivating the model

The main concern of this study is to investigate how argumentation styles affect decision-making in settings with conflicting stakes and incomplete information. In addressing this question, we do not aim to determine how, in actual fact, argumentation styles impact on decision-making, by making the models maximally descriptively adequate. Instead, we offer possible scenarios and deliberative mechanisms for how such argumentation styles might impact on decision making under the given circumstances, with the aim of drawing tentative normative conclusions about them. In line with simulation studies elsewhere in philosophy, we intend to explore the conditions under which social deliberation is beneficial or detrimental to collective opinion formation, focusing on distinctions between conditions that are both empirically and theoretically salient.

To be sure, our model is certainly not free from empirical constraints. To the contrary, and much in line with the sociological literature that our research is embedded in, our modeling choices are underpinned and motivated by empirical studies. In fact, we believe our model fares relatively well in approximating the empirical facts of social deliberation on a number of relevant aspects. As evidenced by the references below, the argumentation styles that we distinguish are ideal-typical simplifications of how people have been observed to deliberate in sociological experiments and field studies. Moreover, as argued in the next section, the agents in our model advance arguments in a deliberation according to procedures that resemble behaviors observed among actual deliberators. As said, we do not strive for the full descriptive adequacy of our models. But since we want to use the models for exploring possible deliberative mechanisms and, ultimately, for drawing tentative normative conclusions about forms of social deliberation, we need to ensure the approximate descriptive adequacy of our models in particular respects.

In what follows, we will review a number of key modeling choices and provide further motivations for them, mixing theoretical and empirical considerations.

Further empirical motivations can be found in the next section, in which the model specifications are reviewed more elaborately. Towards the end, the current section also briefly discusses our models in view of a broader philosophical literature.

When asking how argumentation styles affect decision-making in settings with conflicting stakes and incomplete information, what styles are we taking into consideration? Considering an exhaustive list of argumentation styles hardly seems possible and at any rate exceeds the scope of a single study. Instead, the model is inspired by empirical and theoretical research on human behavior in deliberative settings (Cialdini & Goldstein, 2004; Deffuant et al., 2000; Mercier & Sperber, 2011; Wittenbaum et al., 2004) and sets out to compare populations of agents using either of two ideal-typical argumentation styles. *Advocates* represent individuals who communicate arguments supporting what they currently prefer, given the information they possess. In this sense, advocates represent individuals who raise information that is consistent with their own beliefs, without taking into account any characteristics of the agent they are talking to. This specific type of agent is inspired by empirical research on discussion settings (Mercier & Sperber, 2011; Stasser & Titus, 2003; Wittenbaum et al., 2004), suggesting that individuals are usually inclined to raise arguments in favor of their own opinion. Theoretical models on social influence (Flache et al., 2017; Hegselmann & Krause, 2002) and collective deliberation (Madsen et al., 2018; Olsson, 2013) usually assume similar behavior according to which individuals freely bring forth arguments supporting what they find best.

Intuitively, one would expect that groups of advocates rarely end up with consensus when conflicting stakes are present. Pushing for what one finds selfishly beneficial when a conversation partner has little to gain seems an unlikely way to convince them. Communicating arguments with others who share the same interests, on the other hand, will amplify latent preferences: new arguments will fall on fertile ground and strengthen their belief in that option. As both types of interactions are repeated many times within the group, patterns should emerge where members with identical stakes become more similar in their convictions but fail to agree on a decision with those opposed to it.

Despite the theoretical (Flache et al., 2017) and empirical (Mercier & Sperber, 2011) basis motivating our choice of advocate-type agents, many situations can be thought of in which individuals deviate from arguing for what they find best. Social conformity, for example, is a strong force in human behavior (Asch, 1956; Cialdini & Goldstein, 2004) and may prompt individuals to steer clear of disagreements with their conversation partners. Discussions in 'Hidden Profile' settings (Stasser & Titus, 1985) show that group members often fail to raise dissenting information because they underestimate its significance (Lu et al., 2012; Wittenbaum et al., 2004). Drawing on Social Judgement Theory (Sherif & Hovland, 1961), literature on 'bounded confidence' (Deffuant et al., 2000; Hegselmann & Krause, 2002) argues that individuals will reject or ignore information that is too different from their own

convictions. Prominent works on repulsive influence (Baldassarri & Bearman, 2007; Flache & Macy, 2011; Mark, 2003) assume that trying to influence someone with information perceived as too dissonant may provoke even greater opposition in them. In all of these cases, strategic considerations would lead individuals to express convictions more similar to their conversation partner than they actually are, either out of fear of being sanctioned, or because a more honest expression would get rejected right away.

For these reasons, we introduce a second argumentation style: *Diplomats* aim to convince others of the option they currently prefer as well but are cautious to not offend their conversation partner. Such agents arguably follow an ideal of reasonable discussion that has deep roots in pragmatist philosophy: they are guided by a communicative rather than a strictly instrumental rationality in their social deliberations (Habermas, 1985a, 1985b) and occupy a shared space of reasons (Brandom, 1994). In our simulations, as further discussed below, we attempt to capture these ideas on reasonable debate in a specific deliberative format.

To be clear, philosophers like Habermas, Rawls and Brandom have argued for an ideal of reasonable discussion primarily because of its potential to overcome diverging preferences. But the same reasonableness may also prove its worth when preferences are irreconcilable. In particular, we might expect diplomats to reach a state of ill-informed consensus more often than advocates. This is so because diplomats will raise arguments supporting a second-best option instead of trying to convince their conversation partner of an option they prefer the least. On a group level, this should enhance the circulation of arguments in favor of an option both subgroups find neither worst nor best. In consequence, states may emerge where so many arguments in favor of what is in fact the second-best option have been shared that everyone ends up convinced that this is the most beneficial option.

Further developments in the philosophical understanding of social deliberation, both in social epistemology and in political theory, lead us to question this intuitive connection between an ideal of reasonable debate on the one hand and the feasibility of consensus on the other. In the social epistemology of science, researchers have discovered that a certain degree of fragmentation and dissensus, fueled by constraints on information sharing, may be beneficial to the results of social deliberation (Zollman, 2010). In political theory, the broadly Habermasian ideal of reasonable debate has been challenged by so-called agonistic pluralism (Mouffe, 1999), i.e., the view that a focus on reasonable debate cannot properly accommodate the depth of disagreement between deliberators and in fact hampers political representation. In short, both formal social epistemology and activist political theory have offered arguments towards the overall idea that social deliberation benefits from vocal participants that sustain and act out dissensus.

While these arguments against reasonable diplomats and in favor of vocal advocates pertain in first instance to settings in which there is, at the level of the

collective, a shared goal, like finding a correct scientific theory or achieving adequate political representation in a pluralist society. They are not principally geared towards the phenomenon of ill-informed consensus. Nevertheless, it is conceivable that the beneficial effects of dissensus and advocacy once again carry over to settings in which preferences are irreconcilable and in which consensus can only be reached through faulty preference perceptions. It seems clear that if the positions of the deliberators are fundamentally at loggerheads and entirely transparent to themselves, consensus formation is impossible. However, if deliberators have incomplete information about what serves their interest best, then following a communication strategy that embraces rather than eschews conflict may offer advantages.

All in all, we conclude that the debate over social deliberation offers arguments pulling in opposite directions. Speaking generally and acknowledging that there will be other ways to partition the debate into camps, we can place a Habermas-inspired ideal of reasonable debate, represented by diplomatic interlocutors, against a Mouffe-inspired ideal of agonistic debate, represented by vocal advocates. Both camps have arguments going for them and this leads us to ask: which of these strategies is more conducive to achieving an outcome of truthful disagreement or ill-informed consensus?

In what follows we study such questions by means of simulation studies. We describe how argumentation styles influence groups to either arrive at ill-informed consensus or part ways in truthful disagreement, and what factors play a role in the deliberation dynamics leading up to it. Because plausible settings exist where one or the other outcome is more advantageous, we refrain from interpreting simulation results normatively. For example, issues such as impending health crises or natural catastrophes can be pressing enough that taking any kind of decision - even if partially based on inaccurate knowledge - is preferable over further disagreement. Conversely, 'agreeing to disagree' may be preferable in simple transactional situations when disagreeing group members can find more fitting decision partners elsewhere, and when failure to obtain consensus bears little consequences in the first place. For this reason, lessons from our simulations must be assessed in the light of the specific contexts in which decisions are made.

To this day, to our knowledge at least, the model we built is the first to study collective deliberation in a context where group members' individual preferences diverge. There are of course game-theoretic models and simulations in which agents are self-interested. But our focus is not on agents coordinating their actions to each other, but rather on agents deliberating, i.e., exchanging arguments and investigating the possibility of consensus. Most prominent models of normative deliberation (Hegselmann & Krause, 2006; W. Mason & Watts, 2012; Olsson, 2013; Zollman, 2010) assume that the best option is the same to everyone, and that full information will lead to natural convergence around this option. In our model, consensus only

emerges when at least some group members have faulty perceptions about their preference. Full information, on the other hand, prompts agents to discern their individual stakes and preferences, and then disagreement is the natural outcome. In light of the many plausible contexts where insurmountable disagreement is the most truthful conclusion, we deem the approach taken here an important yet understudied setting.

In addition, our model is among the few to study how argumentation styles affect consensus-making in groups (see van Veen et al., 2020 for a related study). Simulation research has investigated how network structure (Bernstein et al., 2018; Lazer & Friedman, 2007; W. Mason & Watts, 2012; Shore et al., 2015), homophilous interaction preferences (Stein et al., 2024) and cognitive characteristics of the recipient of an information (Madsen et al., 2018; Zollman, 2010) shape collective deliberation. However, less is known about the possible group-level consequences on how people choose information they disclose - although it appears that the latter should greatly affect the former. By comparing groups of agents who advocate for what they think is best with agents who avoid harsh disagreement, our study takes a first step in this direction. The choice of advocates and diplomats can also be seen as a comparison of direct versus indirect communication styles across organizational or cultural contexts (Hall, 1976). Of course, the two argumentation styles we outline do not remotely capture all ways according to which humans reason with each other. Instead, the focus of this study is to compare the status quo in how most deliberation models assume actors to argue (Hegselmann & Krause, 2006; Madsen et al., 2018; Olsson, 2013; Zollman, 2010) with a more nuanced, and perhaps more realistic argumentation style (Hahn & Harris, 2014). As mentioned, competing theoretical expectations about whether diplomats or advocates form consensus more often exist, underlining the importance of simulating discussion outcomes in groups with different argumentation styles.

5.3 Model description

To represent a discussion setting where group members face diverging preferences, we assume a group of N agents consisting of two subgroups $\{\alpha, \beta\}$. Every agent i affiliates with one of the subgroups g. Agents are endowed with arguments that support either option o out of three decision options $\{A, B, C\}$. Taken together, all arguments pertaining to a decision option reveal the *true preferences* of a given subgroup of agents for that particular option. Arguments contain subgroup-specific weights so that the true preferences for a given option of one subgroup are different from the true preferences of the other. The creation of options, arguments and argument weights is outlined in the description below. Subgroups are assumed to be of equal size.

During group deliberation, agents operate under limited information and form *preference perceptions* based on the arguments they currently possess. At each

round t of the simulation, one agent attempts to influence another agent by sharing an argument. The receiving agent then integrates the argument into their argument set and updates their preference perceptions.

The discussion ends when the group forms consensus, i.e., all agents have identical perceptions in terms of which option they prefer the most. Note that the model is set up such that consensus is only possible before all agents possess all arguments. Under full information, agents' perceptions equal their true preferences, which are different for the two subgroups. Unless consensus is obtained, we stop the simulation when enough arguments have spread so that agents' perceptions approximate their true preferences, and no argument combination they receive would be capable of changing their conviction.

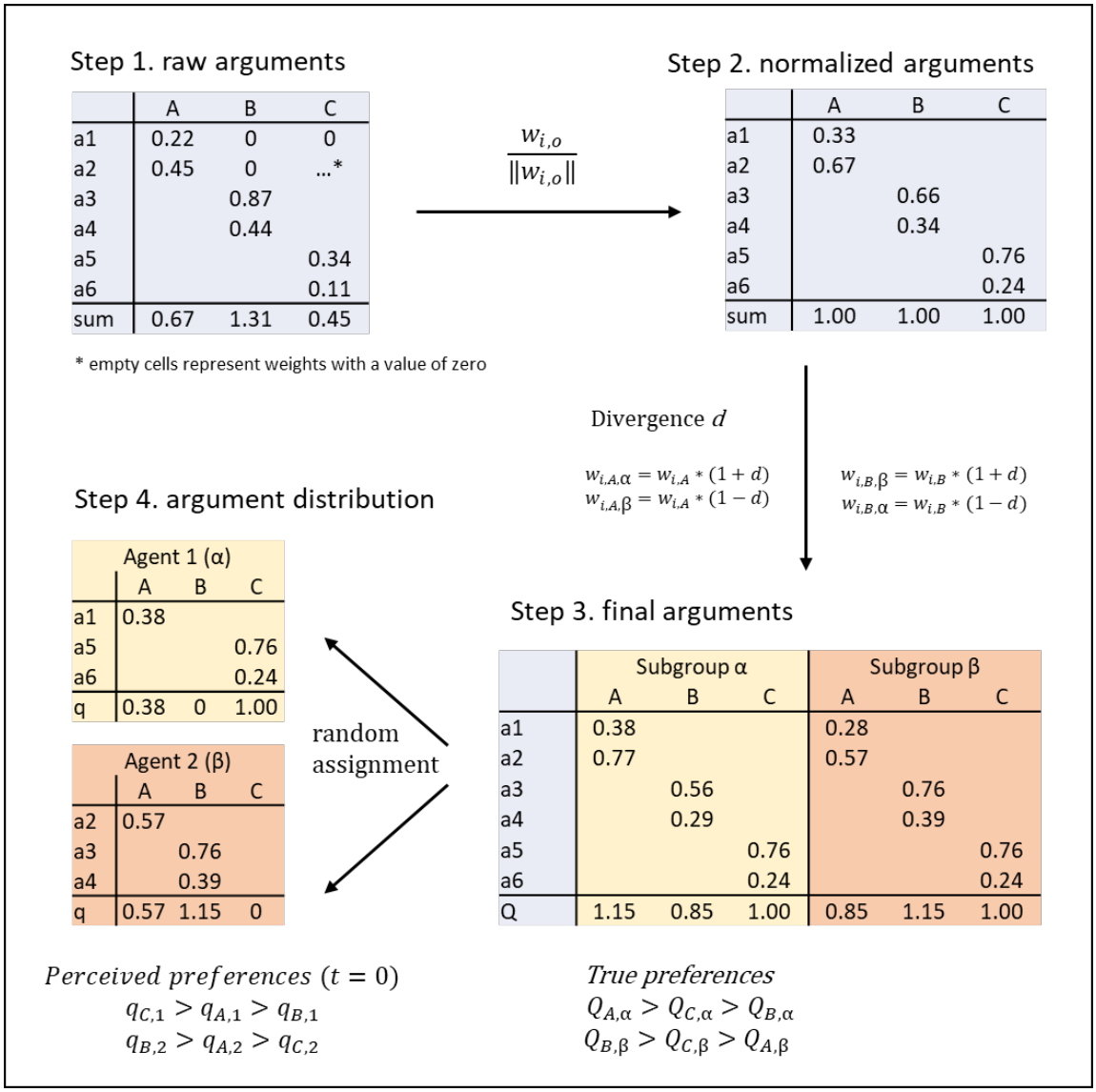
5.3.1 Initialization

Prior to the deliberation process, we assume a fixed set of three decision options $O = \{A, B, C\}$ and I available arguments $A = \{a_1, a_2, \ldots, a_I\}$. Fixed sets exclude the possibility that agents redefine options or create new ones, which would be hard to design, track and explain in a simulation. Each argument a_i contains weights for each decision option, i.e. $a_i = \{w_{i,A}, w_{i,B}, w_{i,C}\}$ representing information about the benefits of the different options. Following related models on consensus-making and argumentation in groups (Stein et al., 2024; van Veen et al., 2020), we assume that each argument cannot support more than one option simultaneously. That is, we assign each argument a positive weight to only one of the options and a weight of zero for the other options. We further assume that each decision option has an equal number of arguments with a positive weight pointing towards them. For a given simulated group, weights are first randomly drawn from a uniform distribution so that $w_{i,o} \sim U\{0,1\}$. Subsequently, weights are normalized ($w_{i,o} / \| w_{i,o} \|$), such that each option's sum of weights equals that of another (Figure 5.1, Step 1 & 2).

In Step 3 of the argument creation process, we manipulate all arguments pertaining to option A and B such that their weights are different for the two subgroups. In doing so, we capture the aspect that options have divergent benefits for members of different subgroups, and that this is reflected in the weights of the arguments pertaining to them. We introduce a divergence parameter $d \in [0,1]$. All weights pertaining to option A are multiplied by $1 + d$ for α agents and by $1 - d$ for β agents, and all weights pertaining to option B are multiplied by $1 + d$ for β agents and by $1 - d$ for α agents. Weights pertaining to option C remain unchanged, i.e. they have the same value for members of either subgroup.

For an agent of a given subgroup, the sum of weights associated with a decision option represent their *true preference*: $Q_{o,g} = \sum_{i=1}^{I} w_{i,o,g}$. Summation reveals that for any value of divergence $0 < d < 1$, preferences of α members will correspond to $A > C > B$, while preferences of β members correspond to $B > C > A$. True preferences thus crucially capture the theoretical scope of the study, namely, a divergence of

preferences among group members under full information, with a higher value of *d* implying higher divergence. Note that although every individual agent has a strict preference ranking, the incomparability of interpersonal utility (Hausman, 1995; Robbins, 1938) makes it problematic to rank options on an aggregate (group) level. We therefore do not make assumptions about the collective benefits (or 'optimality') of either of the decision options.



	A	B	C
a1	0.22	0	0
a2	0.45	0	...*
a3		0.87	
a4		0.44	
a5			0.34
a6			0.11
sum	0.67	1.31	0.45

	A	B	C
a1	0.33		
a2	0.67		
a3		0.66	
a4		0.34	
a5			0.76
a6			0.24
sum	1.00	1.00	1.00

Agent 1 (α)	A	B	C
a1	0.38		
a5			0.76
a6			0.24
q	0.38	0	1.00

Agent 2 (β)	A	B	C
a2	0.57		
a3		0.76	
a4		0.39	
q	0.57	1.15	0

	Subgroup α			Subgroup β		
	A	B	C	A	B	C
a1	0.38			0.28		
a2	0.77			0.57		
a3		0.56			0.76	
a4		0.29			0.39	
a5			0.76			0.76
a6			0.24			0.24
Q	1.15	0.85	1.00	0.85	1.15	1.00

Figure 5.1 Creation and distribution of arguments, argument weights and preferences

The last step of the initialization procedure is to assign arguments to group members (Step 4 in Figure 5.1). Here, assuming systematic biases in agents' information acquisition prior to discussion would make it possible to assign arguments such that they mainly correspond to agents' true preferences. However, since we do not have a particular theoretical or empirical motivation that would lead us to assume such assignment, we let agents take turns at randomly drawing from

the total set of arguments, one at a time without replacement, until all arguments have been assigned. The appendix section of this chapter includes additional analyses assuming lopsided initial argument distributions. Note that because argument weights for agents of different subgroups diverge, assigning arguments at random still generates initial preference perceptions that correlate with agents' true preferences. How agents form perceptions is outlined below.

5.3.2 Argument processing and communication

Similar to how a true preference $Q_{o,g}$ is computed, an agent forms a *preference perception* $q_{o,x,t}$ for each option *o* by summing over the weights of the arguments they possess at round *t*. This allows agents to rank options from being perceived as most to least preferred. Preference perceptions are based on incomplete information and do not necessarily overlap with the true preferences of a group member. Because preference perceptions of an agent can shift over time, they are denoted with a subscript *t*. Subscript *x* denotes the individual agent.

Over the course of the simulation, agents communicate arguments, influencing other group members with the arguments they share. Agents who receive arguments integrate them and update their perceptions, using the weights that correspond to their subgroup.[10] Each round t of the simulation consists of the following steps:

1. A *sending agent* and a *receiving agent* are activated.
2. The sending agent selects an argument to share with the receiving agent.
3. The receiving agent integrates the argument and updates their perceptions.

We compare groups in which the sending agent selects an argument according to either of two argumentation styles: Agents who are *advocates* select arguments that tend to strongly support the option they perceive to prefer the most, regardless of the perceptions of the receiving agent. *Diplomats* do the same but avoid selecting an argument that supports the option their receiver perceives to prefer the least. Diplomats thus only differ from advocates when preference perceptions between agents are opposed, and otherwise behave in an identical fashion.

Similar to many empirical settings, both diplomats and advocates are unaware of the exact set of arguments of other agents, making it possible that agents communicate an argument that others are already aware of. However, we do assume that group members are aware of each others' perceived preferences. This reflects the notion that real-world decision-makers often do have an intuition of each other's positions but that underlying arguments remain private information. Awareness of

[10] Take the following example as a short illustration of this process: if Agent 1 from Figure 5.1 shared argument *a1* with Agent 2, Agent 2 would add a value of 0.28 to their perceived preference for option *A*, even though the same argument has a stronger weight (0.38) to Agent 1. The perceived preference for option *A* of Agent 2 would thus be 0.57 + 0.28 = 0.85.

perceptions enables diplomatic agents to know what options to avoid during interaction, and lets agents realize when consensus is present.

Following standard procedure of canonical social influence models (Deffuant et al., 2000; Flache & Macy, 2011; Keijzer et al., 2018), we randomly select a sending and a receiving agent at each round of the simulation. The sending agent selects an argument according to a two-step softmax function (Daw et al., 2006). Softmax functions are commonly used for modeling human decision-making across a range of fields (D. Guo & Yu, 2019; Harlé et al., 2015; Sutton & Barto, 2018)[11], are capable of predicting observed decision-making in experimental tasks (Daw et al., 2006; Witt et al., 2024; C. M. Wu et al., 2024), and have properties that make them plausible approximations of human choices (Reverdy & Leonard, 2015): a softmax function assigns the highest choice probability to the option with the highest reward, choice probabilities are sensitive to distances between options, and choices are influenced by an adjustable degree of random deviation and noise. Similar to related collective deliberation models (Stein et al., 2024; van Veen et al., 2020), we implement our softmax procedure as follows. In the first step of the procedure, the agent chooses an option o^* they want to support. If the agent is an *advocate*, they consider their preference perceptions and choose an option o according to the probability

$$p(o) = exp(\tau * q_{o,x,t}) \,/\, \sum_{o \in \{A,B,C\}} exp(\tau * q_{o,x,t}) \tag{5.1}$$

where the slope parameter τ controls the degree of adherence (as opposed to randomness) in agents' decisions. The higher τ, the more their choice is determined by selecting the option they perceive to prefer the most.

Diplomats select an option in a very similar manner, but with one important difference: they exclude what their receiver perceives to prefer the least from the set of options considered by Equation 5.1. Thus, when a diplomat considers which option to argue for, they act as if the receiver's least preferred option did not exist and choose from the remaining options instead. In consequence, whenever the option a diplomatic sender perceives to prefer the most is simultaneously the option that the receiver perceives to prefer the least, she is most likely to choose the option corresponding to her perceived second preference instead.

The second step of the discrete choice procedure concerns the selection of the argument to be shared. This step is the same for advocates and diplomats alike. Here, an agent regards the set of arguments $A_{x,t}$ they currently hold and considers those argument weights $w_{i,o^*,g}$ that correspond to their chosen option o^*. They pick one of their arguments with the probability given by

$$p(a_i) = exp(\tau * w_{i,o^*,g}) / \sum_{i \in A_{x,t}} exp(\tau * w_{i,o^*,g}) \tag{5.2}$$

[11] Although the name varies across fields. 'Softmax' is used within biology, cognitive and neuroscience studies, economics, sociology and consumer studies use mathematically equivalent 'discrete choice' functions (Blume et al., 2011; Greene, 2009).

Again, the parameter τ determines agents' adherence to choosing stronger versus weaker arguments pertaining to her chosen option. We assume that the value of τ in Equations 5.1 and 5.2 is the same across simulations. By default, we set the adherence parameter to $\tau = 2$ such that agents make choices neither deterministically nor randomly, but according to probabilities that lie somewhere in between these two extremes. This behavioral assumption is consistent with empirically validated models employing similar choice functions (Daw et al., 2006) and coherent with our understanding of human deliberation (Wittenbaum et al., 2004). As shown in the appendix section of this chapter, main results do not depend on the exact value of the adherence parameter but hold for any $\tau > 0$ (cf. Figure 5.6 A).

After the sending agent has chosen an argument to be shared, the receiving agent integrates the argument and concludes the round by updating her perceived benefits. Since arguments represent information about the benefits of different decision options, which diverge among subgroups, a receiving agent always integrates an argument using the weights corresponding to their own subgroup. These weights can be different from the sender's weights.

The process of randomly activating an agent, sharing an argument, and updating benefit perceptions of the receiver is repeated until either of two states are reached: (1) all agents align in their perception of what they prefer the most, i.e. they form consensus. (2) Two agents of opposing interest groups perceive to prefer what they actually prefer, and no argument combination they receive can possibly change their perception. In this case, consensus becomes impossible, and the simulation would continue until all agents had received all arguments. Note that because this state becomes more likely the more arguments are exchanged, consensus must emerge early enough for all group members to be able to align on one of the options.

5.4 Setup of simulation experiments

We conducted simulation experiments to investigate deliberation outcomes in groups with diverging preferences and different argumentation styles, tracking whether the simulated groups reached one of the following states: (i) everyone in the group perceives their second-best option *C* to be best. (ii) Everyone perceives *A* or *B* as best, which is the best option for half of the group but the worst option for the other. (iii) The group disagrees about which option is best and the discussion has reached a point where perceptions cannot be altered. Argumentation styles were represented by groups that either consist of advocates or diplomats. We operationalized preference divergences by the parameter *d*, with higher values of *d* implying higher divergence. Theoretical intuition led us to expect that diplomats more often form consensus than advocates but gives no indication of the strength of preference divergence to assume. For this reason, we started the analyses with a simple comparison of the probability of ill-informed consensus in groups of

diplomats and advocates under high (d = 0.6) and low (d = 0.2) divergence.[12] Because this simple comparison revealed that discussion outcomes crucially depended on the level of divergence, subsequent analyses vary d from very low (0.05) to very high (0.95) levels in steps of 0.05, and compare discussion outcomes in groups of advocates versus diplomats at each level.

For every parameter combination, we simulated 1,000 independent discussion processes.[13] We assume groups of N = 6 agents, which is not an unrealistic size in decision-making settings. A set of I = 90 arguments was used (30 arguments per decision option) to create a set of arguments that is sufficiently large for deliberation outcomes to not be determined by the coincidental spread of single arguments. The adherence parameter τ is set to 2, meaning that agents choose options and send arguments that correspond well to their argumentation style while still allowing for a small degree of randomness in argument communication. Figure 5.6 in Section 5.8.2 demonstrates robustness of the main results at different values for τ. Main results include additional analyses in which we vary group size between 4 and 12 in steps of two, and the number of arguments between 30 and 300 in steps of 30.

The findings of this paper rest on a setting where arguments are initially distributed at random. Yet, contexts can be thought of where cognitive heuristics (Mercier & Sperber, 2011) or homophilous information networks (McPherson et al., 2001) would lead individuals to acquire information selectively prior to discussion. For this reason, additional analyses reported in Section 5.8.1 elucidate that discussion outcomes may differ when initial argument distributions correlate with subgroup membership, but that the underlying mechanisms stay the same. Concluding sensitivity analyses investigate if mixed groups of advocates and diplomats result in more or less consensus compared to groups using either argumentation style exclusively.

5.5 Results

We start off by comparing how often agent groups of advocates versus diplomats find ill-informed consensus under two levels of preference divergence (Table 5.1). Under high divergence, groups of diplomats converge much more often on either option (41%) than advocates (8%), which is consistent with the intuition that Diplomats' avoidance of harsh disagreement fosters consensus formation. Counter this intuition, however, only 16% of diplomat groups converged on a consensus under low divergence while more than 90% of advocate groups do. Intuitively, lower divergence should foster consensus because weights are more similar for members

[12] For reference, high divergence means that the true preference score of group members' most preferred option is 400% higher than their least preferred (1.6 and 0.4, respectively). Low divergence, on the other hand, implies an increase of 50% (1.2 versus 0.8).
[13] Simulation code, output and syntax to replicate this study available at https://osf.io/76hfm/

of opposing subgroups, making it easier to converge. While this is clearly the case for advocates, why can the same not be said of diplomats?

Table 5.1 Percentage of groups reaching consensus on either option, by group's argumentation style and level of preference divergence *d*

	preference divergence	
	low (d = 0.2)	high (d = 0.6)
Advocates	95%	8%
Diplomats	16%	41%

Figure 5.2 hints at a possible explanation for this puzzling finding, presenting a more fine-grained examination of discussion outcomes across the divergence parameter range. The dashed lines in Figures 3A suggest that both advocates and diplomats are less likely to experience consensus on *A* or *B* as divergence levels rise. This is explained by the fact that under higher divergence, weights of arguments in favor of *A* or *B* are increasingly different for members of opposing subgroups, making it harder for the group to converge around either of these options. Advocates are especially likely to find consensus on *A* or *B* under low divergence because their argumentation style has self-reinforcing characteristics: random initial majorities in favor of an option will convince other group members, who will then advocate for this option as well, leading to swift convergence. Diplomats, on the other hand, are limited by what their interaction partner perceives to prefer least, making it hard to find consensus when single individuals find worst what a majority finds best.

Among advocates, consensus on option *C* becomes increasingly rare under higher divergence as well (Figure 5.2 A, solid red line), despite arguments weights in favor of *C* being unaffected by *d*. An explanation for this finding becomes apparent from the solid black line in Figure 5.2 B: as divergence gets larger, agents' initial perceptions increasingly align with their true preferences, meaning that members of different subgroups start the discussion with diametrically opposed perceptions already. As discussions evolve, advocates amplify this initial disagreement when raising arguments according to what they perceive to benefit from the most, up to a point where discussions are deadlocked and consensual agreement becomes impossible. Figure 5.2 C supports this explanation, showing that the number of *C* arguments sent in groups of advocates decreases in higher divergence (solid red line).

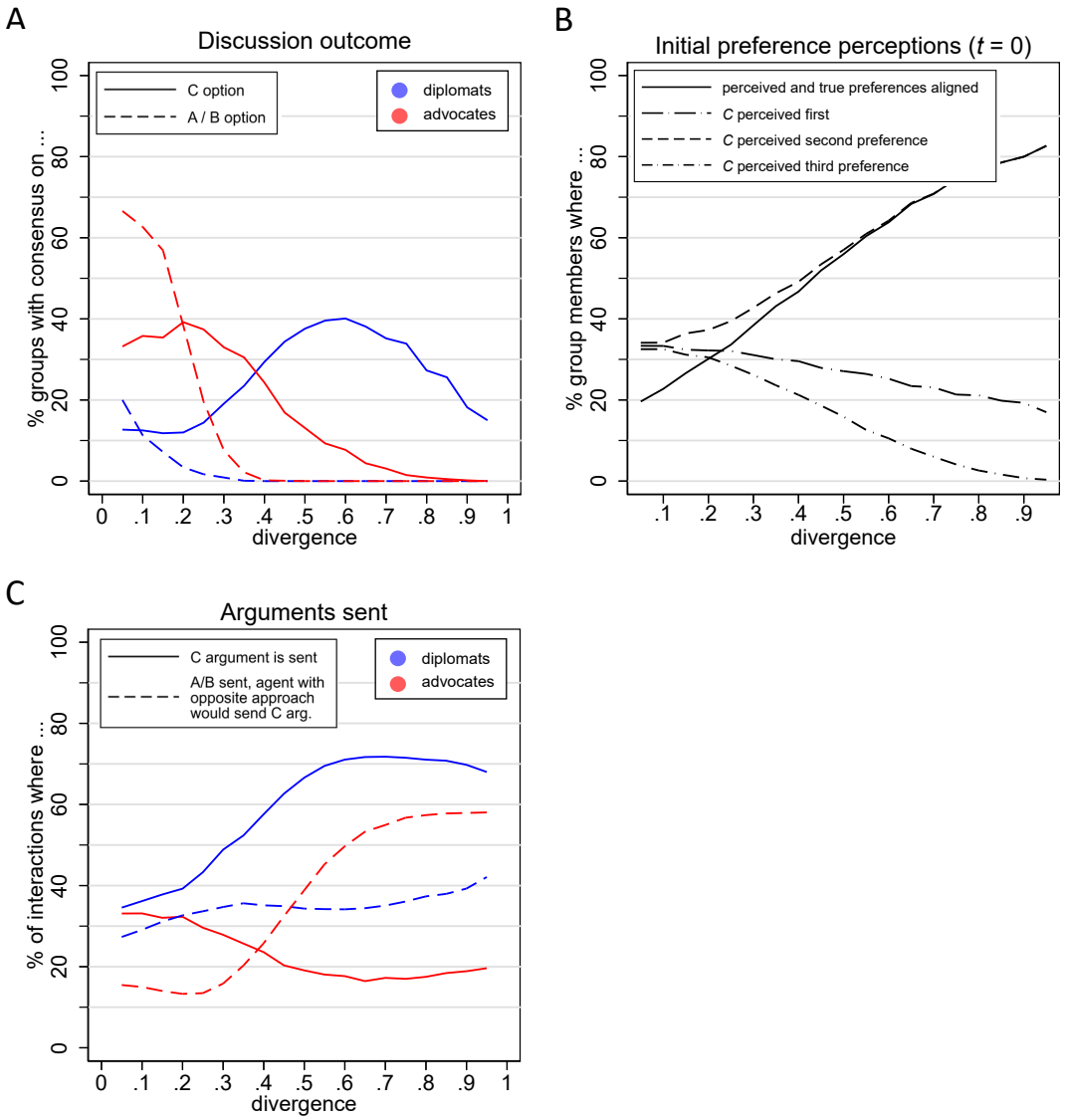


Figure 5.2 Discussion outcome, initial preference perceptions and agent sending behavior, by divergence

Ill-informed consensus on option *C* in groups of diplomats, on the other hand, follows a complex pattern (Figure 5.2 A, solid blue line): higher divergence implies greater chances of consensus until $d = 0.6$. But divergence levels beyond 0.6 negatively impact the proportion of groups with consensus on *C* again. Contrary to intuition, consensus in groups of diplomats occurs less often than among advocates until $d < 0.4$. A closer look at Figure 5.2 B offers an intuition why diplomats rarely form consensus: at low divergence levels, the proportion of group members initially preferring *C* the least is relatively high (short dash-dotted line). Because of their argumentation style, diplomats will avoid sending arguments favoring *C* to such agents, making consensus on *C* unlikely. Higher divergence levels, on the other hand, make it easier for diplomats to form consensus: more agents start the discussion

with perceptions that correspond to their true preferences, meaning that agents of different groups will have opposing perceptions about options *A* and *B*, and more agents perceive *C* as second best. As divergence rises, diplomats raise more arguments in favor of *C* (Figure 5.2 C, solid blue line), explaining a greater fraction of groups with consensus on this option until $d = 0.6$.

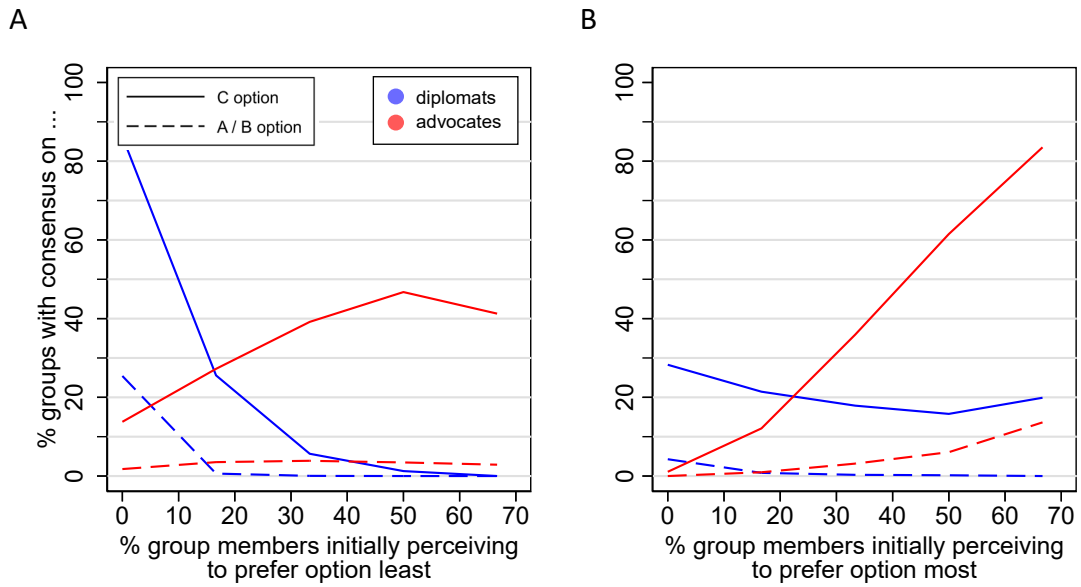


Figure 5.3 Discussion outcome at $d = 0.30$, by initial perceptions of group members

Divergence does not affect the initial distribution of arguments but changes the way arguments are perceived by an agent. Higher divergence implies that an *A* argument will have a much greater impact on an α-agent's perception that *A* is best, but a smaller impact on the perception of a β-agent. From this follows that at low divergence, assigning arguments at random creates perceptions that are more random, while random argument assignment at high divergence implies perceptions that correspond closely to the ranking of true preferences among group members.

In sum, the results obtained here reveal a striking finding: at low divergence levels (< 0.2), advocate groups are more than three times as likely to find an ill-informed consensus on either option, but the opposite is the case for higher divergence levels (> 0.6). The explanation we propose is that divergence impacts the initial distribution of perceptions, which in turn interacts with agents' argumentation style and their chances of finding consensus: advocates are more likely to find consensus on an option when more agents initially perceive to prefer this option the most. Diplomats, on the other hand, are more likely to establish consensus on an option the fewer agents perceive to prefer this option the least.

We test the proposed explanation by zooming in on groups of diplomats and advocates at a moderate level of $d = 0.3$ and analyze how often they find consensus, depending on the proportion of agents in the group initially perceiving an option to

be least (Figure 5.3 A) or most beneficial (Figure 5.3 B). The results overwhelmingly support the proposed explanation: the proportion of diplomat groups finding consensus sharply declines the more group members initially perceive an option as worst, while advocates become more likely to build consensus the higher the proportion of agents initially perceiving an option as best. Note that the two variables are positively correlated: due to the random assignment of arguments during initialization, allocating disproportionately few arguments in favor of an option to some agents by chance implies that others will receive disproportionately many. Because of this, almost 50% of advocate groups find consensus on C even when two thirds of the group initially perceive this option as least preferred.

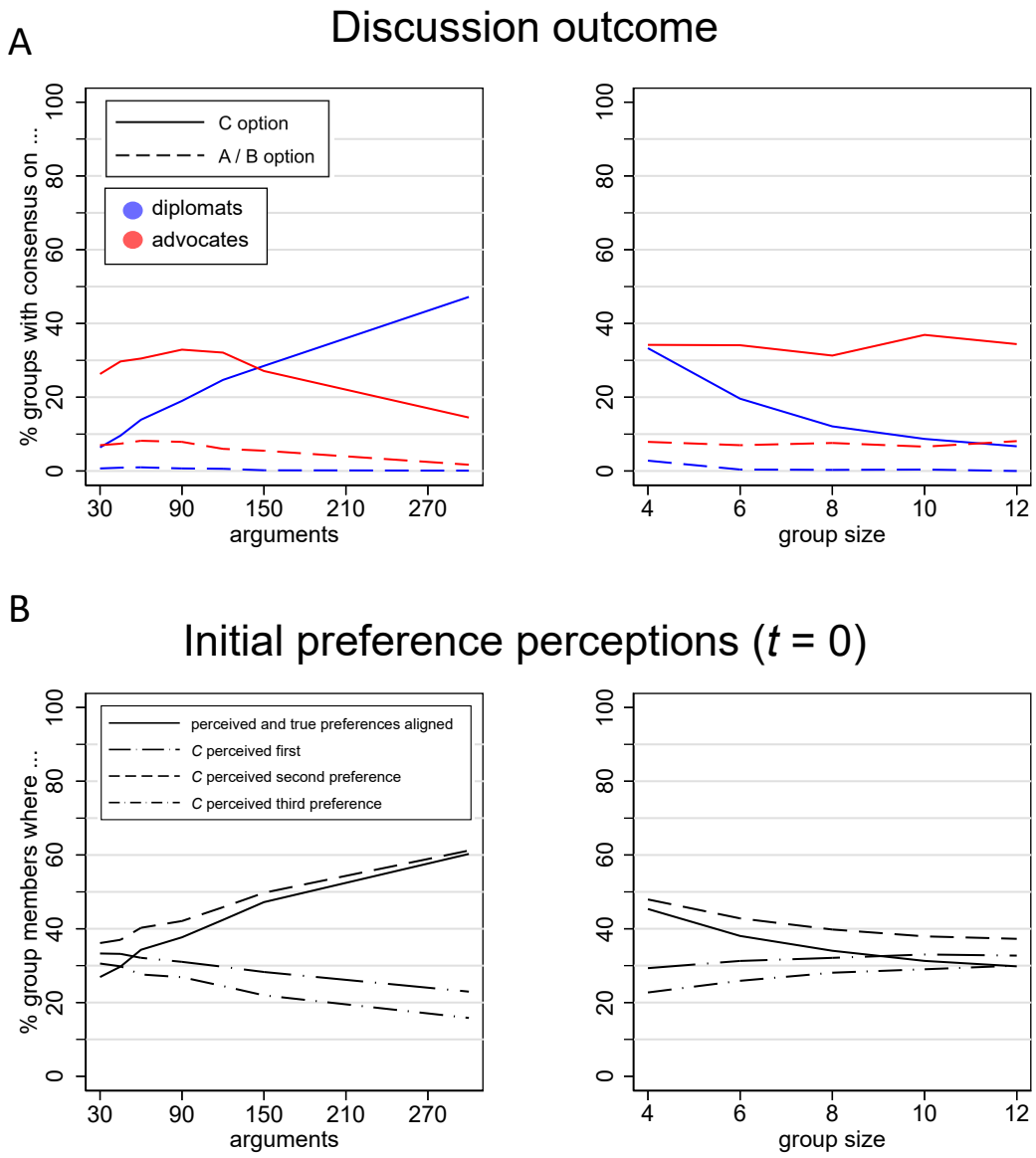


Figure 5.4. Discussion outcome and initial preference perceptions of group members, by argument pool and group size (d = 0.3)

Figure 5.2 showed that divergence affects the consensus-making capacities of advocates and diplomats through its influence on the initial distribution of perceptions. We now investigate whether the same is the case for other substantial features of the discussion setting, namely the number of available arguments and group size. If the general explanation holds, any factor leading to a closer alignment between initial perceptions and true preferences should positively impact consensus-making for diplomats but negatively for advocates. Figure 5.4 shows that this is indeed the case. As the number of arguments grows, agents hold more arguments at the start of the discussion. Here, the law of great numbers implies that initial perceptions that are based on more arguments will more closely resemble their expected value (i.e. $E(q_{x,t0,o}) = Q_{o,g} / N$) and are hence more often aligned with agents' true preferences (Figure 5.4B, left side). In consequence, more groups of diplomats and fewer groups of advocates find consensus as the number of arguments grows (Figure 5.4A, left side). Group size, on the other hand, has an opposite effect (Figure 5.4A, right side): as groups get larger, the same number of arguments is distributed over more agents, leading to less alignment of initial perceptions with the ranking of true benefits, and, in turn, less frequent consensus among diplomats.

In sum, the results presented here reveal striking insights into the deliberation outcomes of simulated groups with diverging preferences: when many group members initially misperceive an option as preferred, advocating for this option regardless of others' perceptions will be a very effective way to emerge with a – likewise ill-informed – consensus. But when group members have initial perceptions that closely align with their true preferences, consensus requires an argumentation style that makes members bring forth arguments in favor of a second-best alternative. Additional analyses in Section 5.8.1 show that this finding persists when arguments are initially distributed in a lopsided fashion. Results are robust to the level of strategy adherence τ, do not depend on random as opposed to homophilous interaction preferences, and remain similar when a decision is made at only four or five out of six group members perceiving to prefer the same option. Further analyses reported in Section 5.8.2 reveal that results are not artifacts of perfectly homogenous groups either; but persist in mixed groups where minorities of agents using the opposite argumentation style are introduced to a population.

5.6 Discussion

Considering the complex relationship between discussion outcomes, argumentation styles and preferences perceptions, our results offer a new perspective on existing theoretical narratives. The simulations support the overall conclusion that the two argumentation styles under consideration have their own merits and defects relative to context. As indicated before, we do not believe that our models fully match real-life deliberations but we believe them to be sufficiently descriptively adequate to lend

normative force to the conclusions, in the sense that they can back up tentative claims about the desirability, relative to a context, of one or other argumentation style.

Importantly, in this new perspective we cannot and do not take a stand on the desirability of the outcomes of social deliberations: the merits and defects of ill-informed consensus and truthful disagreement are context-dependent to the extent that choosing an overall favorite would be nonsensical. Our main result is that relative to what is deemed beneficial in a certain context, and on the assumption that group members harbor a latent disagreement but start off with incomplete information about their actual preferences, there are further specifics of the deliberation setting that will make a diplomatic or an advocating argumentation style the more effective one.

This relatively modest and qualified conclusion challenges overly simplistic applications of ideal-type deliberative formats. In particular, assuming that the context makes a consensus desirable, a simplistic reading of the agonistic approach of Mouffe (1999) might suggest that groups in which individuals advocate according to their best evidence will arrive at superior collective conclusions. However, our results support this idea only when initial evidence is fragmented and noisy, leading agents away from their true preferences from the onset, and then only if the differences in preferences are not too pronounced. Conversely, echoing the perspective of Habermas (1985a, 1985b), it may have seemed that groups where members prioritize avoiding offensive engagement and maintain a common ground often exhibit superior deliberative performance. Yet again, our findings only partially align with this expectation: diplomatic argumentation enables consensus only when agents have sufficiently accurate initial convictions, or when preference divergence is high enough for disagreement to be obvious under more abrasive argument sharing.

We hope that our results, apart from revealing the context-sensitivity of styles of deliberation, stimulate a broader engagement of philosophical debate with insights from computational sociology and formal social epistemology. It bears repeating that we do not view our results, or indeed other results from these formal and computational disciplines, as providing stand-alone support for some or other communication style or format for public debate. Rather, we believe that studying the consequences of such styles and formats from up close will help us in our attempts to make public debate and social deliberations more fruitful, transparent, and fair, through an understanding of the dynamics that drive them.

With our emphasis on the fact that our results depend on context, we do not mean to suggest that contextual considerations cannot be accommodated, at least partially, in further model developments. Argued from a utilitarian point of view, the optimal outcome depends on the post factum consequences of the group decision: disagreement can incur costs for both individuals and create unforeseen externalities, but so can an ill-informed consensus. In our model, we assume that the

consequences of the group decision - beyond what can be learned from the arguments raised in the deliberation process - are unknown to group members, and accordingly that they do not feature in the deliberation. However, relaxing this assumption is certainly possible, and this would open a new array of strategic considerations to be done by agents accompanied by further theorizing that exceeds the scope of the paper.

Like in any simulation model, simplifying assumptions were made for reasons of parsimony and to the benefit of understandability. This involves, for example, that the relatively basic, additive argument structure used here enabled us to intuitively trace and understand agents' preference formation process. Bayesian preference updating (Assaad et al., 2023; Madsen et al., 2018) is regarded as more 'rational' from the agents' point of view, suggesting that exploring this updating approach is a valuable avenue for future modeling work. Similarly, recent advances in the development of large-language models make it possible to formally represent argument communication in terms of realistic human language rather than the abstract representation of arguments applied here (Betz, 2022; Du et al., 2023). However, while LLMs are powerful tools to generate meaningful text, it is unclear whether they also reliably represent human behavior in complex settings like a debate in which individuals deliberate despite having competing preferences. Another simplifying assumption concerns the focus on a dyadic interaction setting. Restricting argument communication to only one receiver per interaction facilitates an easy, straightforward formalization of an argumentation style that takes receiver preferences into account. Multilateral communication, on the other hand, would involve weighing a multiplicity of receiver preferences against specific persuasion goals, introducing additional complexity and exceeding the scope of this study.

The two argumentation styles elucidated here served as prototypes of self-serving versus cautious communication. Obviously, alternative ways to implement these styles exist, and many alternative argumentation styles can be thought of that should be considered by future research. Having studied groups that only consisted of either advocates or diplomats, the question arises if new patterns of discussion outcomes emerge if groups consist of a mix of individuals with different argumentation styles: for example, whether groups with just one or two diplomats are enough to steer the group towards consensus under high preference divergence. Additional analyses reported in the appendix section of this chapter explore this possibility and suggest that the deliberation outcomes do not depend on the assumption that groups are homogenous in argumentation style: instead, mixed groups simply produce discussion outcomes that resemble a linear combination of outcomes in homogenous groups.

Next to alternative argumentation styles and mixed groups, another factor that can potentially affect discussion outcomes is whether groups will only make decisions after having reached consensus or whether they rely on voting procedures

instead (Priem et al., 1995). Of course, investigating the full spectrum of possible voting rules exceeds the scope of any sensitivity analysis, but we show in the appendix that results are at least robust under a majority-based voting rule: low preference divergence still facilitates decision-making among advocates when only four or five out of six members must align in their convictions, while high divergence fosters decision-making among diplomats.

In our model, subgroups impose a preference structure that is symmetric and straightforward. But extensions of our model could weaken this symmetry and allow for a multiplicity of individual stakes next to subgroup membership. Likewise, additional biases and heuristics in the interaction between group members can be implemented. A first step in this direction is investigated in Section 5.8.3, where we show that discussion outcomes remain similar if agents interact in homophilous (McPherson et al., 2001) instead of random encounters. Although group-wide consensus becomes less likely the higher the homophily level, diplomats will continue to find consensus more often than advocates under high preference divergence, while the opposite is the case at low divergence.

Next to homophily, literature on affective polarization (Iyengar et al., 2019) and bounded confidence (Hegselmann & Krause, 2002) suggest that individuals may reject information that comes from disliked or dissimilar sources. While not part of our model in a theoretical sense, our robustness analyses on homophily formally include such behavior: whether individuals encounter outgroup members at a lower probability or accept their arguments with lower likelihood is mathematically interchangeable.

Besides advancing theoretical development to refine the model and validate the robustness of our findings, empirical investigation holds promise in advancing the discourse. Our model's key innovation, the incorporation of agents' conflicting preferences, presents fertile ground for empirical exploration. Preferences can be rigorously quantified and experimentally manipulated in experiments along the paradigm of behavioral game-theory (Camerer, 2011; Fehr & Gächter, 2000), making it possible to create laboratory settings where human participants operate with preferences akin to those in our model. An important empirical question to answer is, for instance, what argumentation styles individuals are using and whether there are conditions under which humans adopt different styles. Experimental work in social psychology, for instance, suggests that individuals may strategically misrepresent their positions when discussing with members of outgroups holding different positions (Hogg et al., 1990).

Likewise, theoretical work would benefit from empirical work on consensus formation in groups. We focused our analyses on the first time a group experienced consensus in that all members perceive to prefer the same option. This consensus, however, is ill-informed since a continuation of the discussion to the point of full information would reintroduce disagreement. An important empirical question is

whether individuals notice that they have reached consensus and stop the debate or whether and under what conditions, they continue the exchange of arguments.

5.7 Conclusion

Our simulation analyses suggest that in groups with diverging preferences, deliberation is shaped by the way members raise arguments as well as their initial preference perceptions. Advocating for what one finds personally beneficial only led to truthful disagreement when group members started the discussion with accurate perceptions about their preferences already. Conversely, when initial perceptions were noisy and inaccurate, random initial majorities often convinced the rest of the group of an option they disfavored had full information been present. We compared the behavior of such 'advocates' with that of 'diplomatic' agents who avoid disagreement at the cost of speaking one's actual mind. In these groups, initially accurate (divergent) preference perceptions led to an ill-informed consensus. Inaccurate initial perceptions, on the other hand, eventually resulted in truthful disagreement. Here, the avoidance of disagreement made consensus hard because majorities failed to convince other group members of an option they found least preferable.

5.8 Appendix

5.8.1 Lopsided argument assignment

The main results pertained to groups where arguments were distributed at random prior to deliberation. Here we investigate discussion outcomes of simulated groups where arguments are initially assigned in a selective fashion. We introduce an additional parameter σ, regulating the probability by which agents draw arguments in support of their most preferred option. They take turns at choosing from the set of available arguments *A'*, one at a time without replacement according to

$$p(a_i) = exp(\sigma * w_{i,o_{max},g}) / \sum_{i \in \{A\}} exp(\sigma * w_{i,o_{max},g}) \quad (5.3)$$

where o_{max} represents the option members of a subgroup g prefer the most. At $\sigma = 0$, arguments are drawn at random, mirroring the setup of discussion groups in the main results. At $\sigma = 2$, argument distribution is highly lopsided, with high chances of A arguments being drawn by α members, B arguments drawn by β members, and C arguments being drawn by members of both subgroups with equal probability. To avoid that initial preference perceptions strongly correlate with group members' true preferences independent of the level of σ, the simulation experiments presented here use a low divergence value of $d = 0.1$.

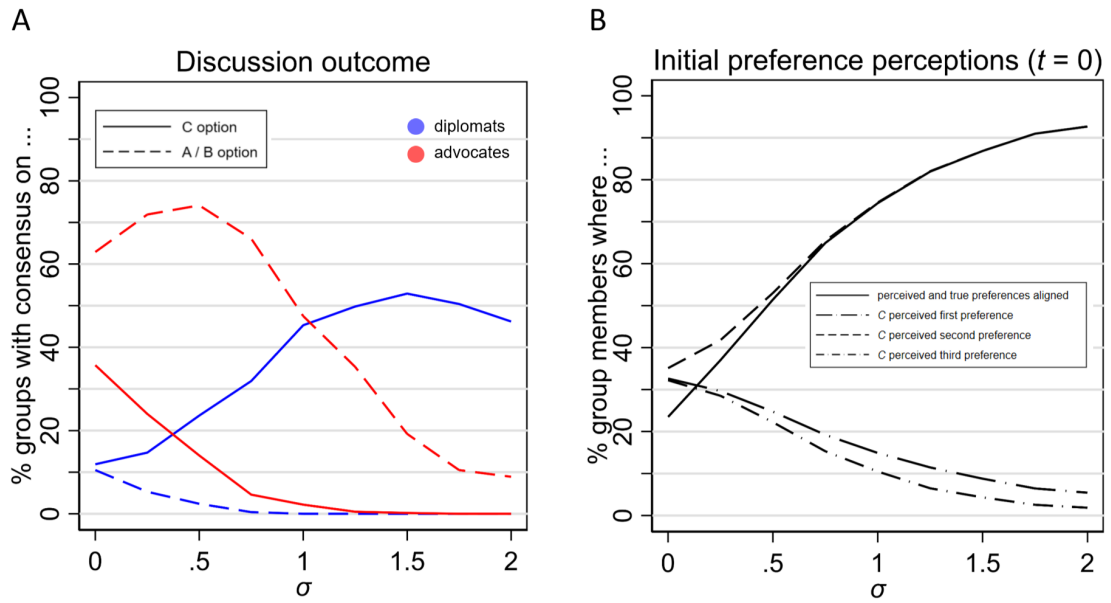


Figure 5.5 Discussion outcomes and initial preference perceptions by sigma at $d = 0.1$

Figure 5.5 shows that higher σ leads to discussion outcomes similar to higher preference divergence (compare Figure 5.2). This is explained by the fact that both parameters tighten the correlation between initial preference perceptions and true preferences (Panel B), either directly through lopsided allocation (σ) or indirectly through differences in argument weights (d, see Figure 5.2 B). Variations of the σ

parameter always produce more consensus on option *A* or *B* than on *C* among advocates, with the reason being the low value of preference divergence: at low *d*, options *A* and *B* are more similar to members of both subgroups, making it easier to align on either of these options.

5.8.2 Strategy adherence and group composition

Figure 5.6, Panel A reveals that across the range of the adherence parameter τ and consistent with the main results, diplomats are more likely to form consensus than advocates at high divergence. At low divergence, consensus is more likely among advocates than diplomats. Differences in the effects of the two argumentation styles become bigger in τ. This is explained by the fact higher adherence implies less randomness and a more deterministic selection of arguments according to agents' argumentation styles. At very low adherence ($\tau < 0.5$), argument selection for both styles approximates randomness, resulting in similar probabilities to find consensus.

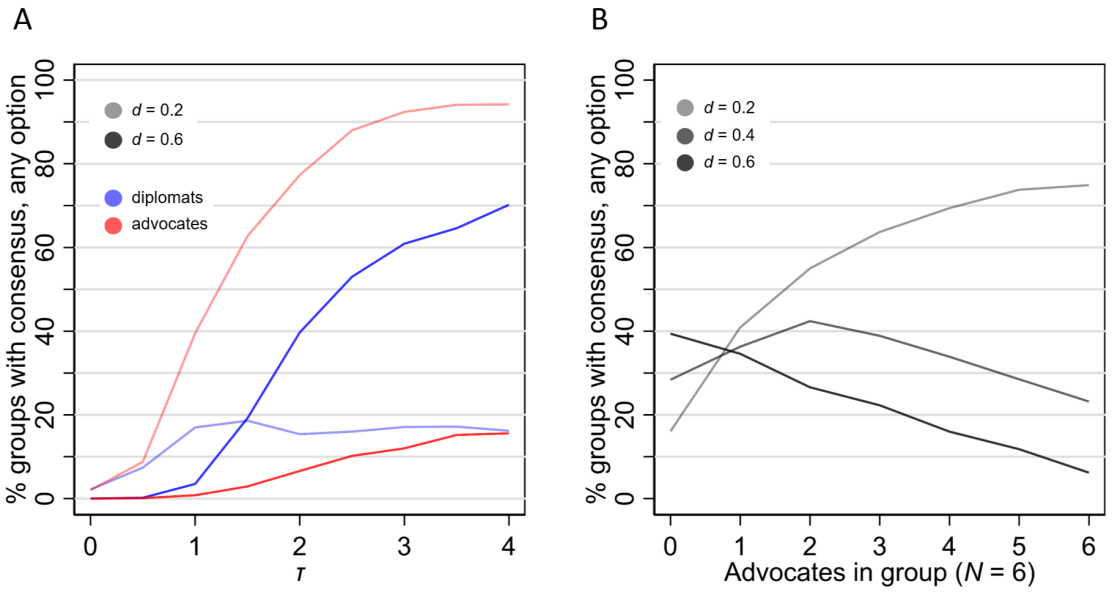


Figure 5.6 Discussion outcomes by adherence τ (A) and number of advocates in group (B)

Figure 5.6, Panel B elucidates discussion outcomes at different divergence levels for mixed groups of advocates and diplomats. At high divergence, the probability of consensus sinks almost linearly in a greater fraction of advocates and rises monotonously at low divergence. Both results are consistent with the main results, showing that consensus occurs more often among diplomats than among advocates at high divergence, while the opposite occurs at low divergence. Only at moderate divergence, a peak in consensus propensity appears at two diplomats and four advocates, exhibiting a curvilinear relationship between consensus propensity and the fraction of advocates in the group. Here, frequent consensus results from diplomats' tendency to raise *C* arguments, combined with advocates' ability to raise arguments even if they go against their opponents' preference perceptions.

5.8.3 Preferential interaction

The simulations that underlie the main results of the paper assume that agents select any interaction partner with equal probability. Here, we test if results are robust to two types of preferential interaction, namely, similarity in perceptions and similarity based on subgroup membership. We regulate interactions between agents by introducing a homophily parameter *h*, ranging from -1 to 1. The greater *h*, the more likely a sending agent is to choose a receiving agent with the same trait. A sending agent chooses a receiving agent according to

$$p_y = s_y / \sum_{y \in \{Y\}} s_y \tag{5.4}$$

where *y* denotes the individual receiver and *Y* the set of potential receiving agents. s_y represents the trait similarity between sender and receiver, taking on the value of $h / 2 + 0.5$ if sending and receiving agent share the same trait and $1 - (h / 2 + 0.5)$ otherwise.

Consistent with the main results, Figure 5.7 shows that advocates are more likely than diplomats to form consensus at low levels of preference divergence ($d = 0.2$), while diplomats are more likely to form consensus at high preference divergence ($d = 0.6$). This is true for the entire parameter range of perception-based homophily (Panel A) and group-based homophily (Panel B). Generally, higher homophily levels indicate lower probabilities of finding consensus. This is because preferential interactions between similar individuals, either in perceptions or subgroup membership, solidify convictions in line with true preferences and result in disagreement as the final outcome of the discussion. Interesting to note is the increase in consensus at low divergence as subgroup homophily reaches higher levels, pointing to potential 'transient diversity' effects (c.f. Chapter 3).

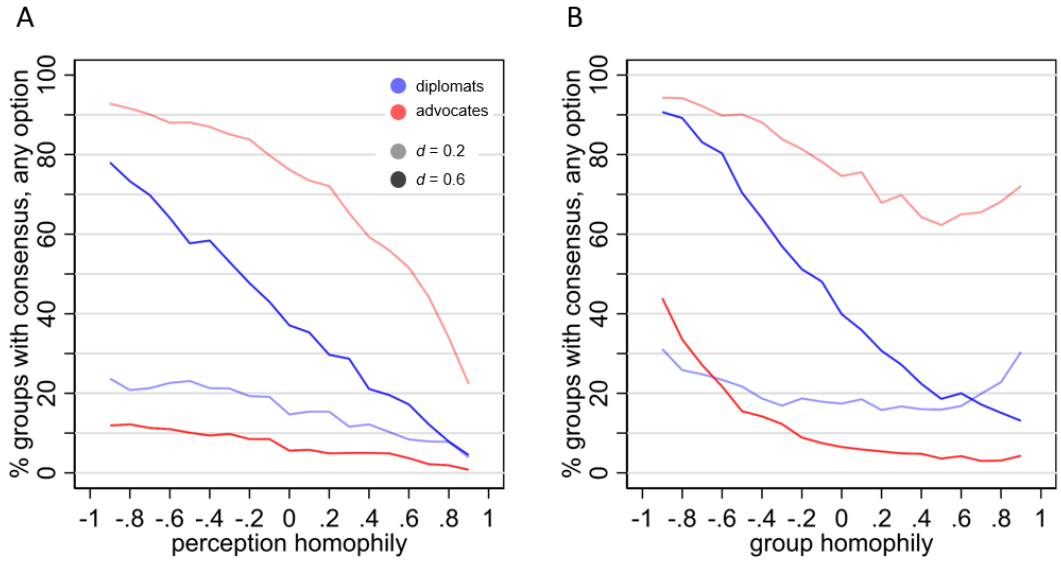


Figure 5.7 Discussion outcomes by perception-based (A) and subgroup-based homophily (B)

5.8.4 Decision-making without consensus

Figure 5.8 reveals that groups would reach similar discussion outcomes if decisions were not made according to full consensus, but according to a 5 / 6 or 4 / 6 majority rule instead. Advocates are less likely to converge around any option as divergence levels rise, regardless of whether 4 or 5 group members perceive to prefer the same option. Diplomats tend to disagree more often as divergence levels rise as well, but this tendency is offset by a local peak at around $d = 0.6$, regardless of the underlying decision-making rule. The latter is, again, explained by diplomats' ability to raise arguments in favor of option *C* as divergence levels rise (Figure 5.8 B).

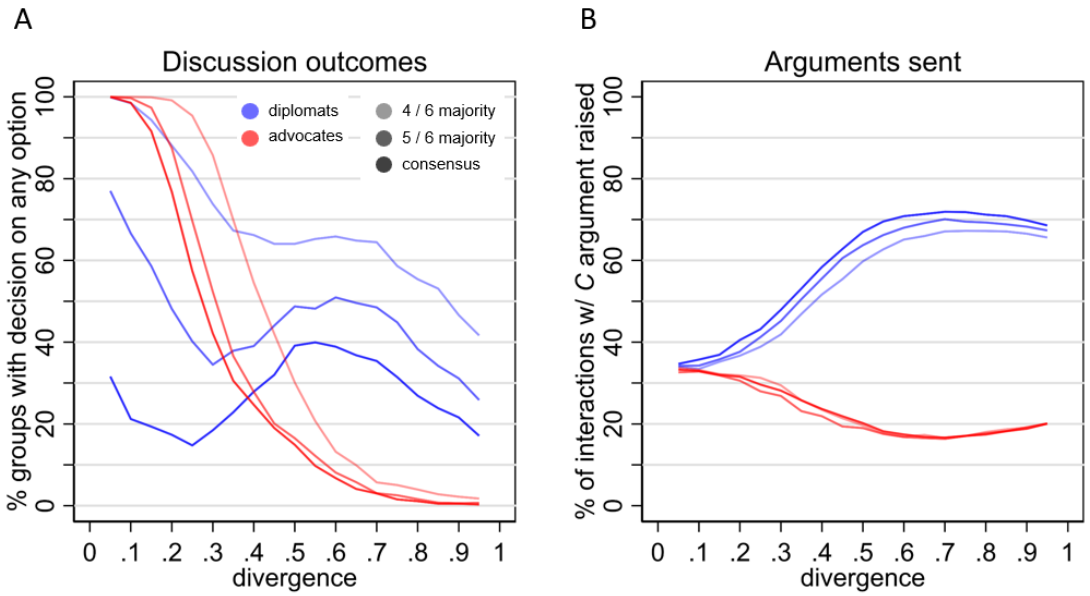


Figure 5.8 Majoritarian decision-making, by divergence

Propositions

1. Identities are not only linked to what people know, they also influence how individuals share knowledge with one another (this dissertation).

2. Ideological segregation can undermine crowd-based misinformation detection (Chapter 2).

3. In diverse teams, homophily can prevent early, suboptimal consensus and improve team decision quality (Chapter 3).

4. Homophily can be beneficial or detrimental for collective decision-making, depending on the setting (Chapters 2 and 3).

5. Ideological differences can reduce the perceived usefulness of information from others, regardless of its actual value (Chapter 4).

6. When facing difficult problems, individuals may prefer to learn from someone who is cognitively different to themselves, which can increase the likelihood of finding a better solution (Chapter 4).

7. Identity diversity can facilitate but also hinder the integration of epistemic diversity in groups (Chapters 2–4).

8. When group members have strongly conflicting interests, avoiding disagreement can help with finding consensus (Chapter 5).

9. Persuasive argumentation can foster consensus when group members have initially inaccurate perceptions of what they think is best for themselves (Chapter 5).

10. *All of you are partly right and partly wrong – each of you is telling it differently because you touched a different part of the elephant.*

English summary

Diversity poses a 'double-edged sword' to collective decision-making in groups. Epistemic diversity – i.e., differences in perspectives, skills, and knowledge – can enhance the accuracy, creativity, and effectiveness of collective decisions. However, epistemic differences are often tied to identity differences among group members. Different backgrounds – cultural, ideological, or professional – can make it challenging to build the mutual trust and open communication necessary to exploit diversity in the process of making better group decisions. Therefore, group diversity can be a potential asset, but not one that is automatically utilized.

Different scholarly fields have tried to identify ways to reap the benefits of epistemic diversity while avoiding pitfalls of identity diversity. Organizational research has gathered much observational data on the demographic composition of work teams and their performance. Sociological simulation models have illuminated how small-scale interaction patterns between individuals can breed large-scale phenomena such as polarization. And formal philosophical work in social epistemology has investigated the conditions under which groups can arrive at epistemically justified, truthful decisions.

While specialization in these fields allows for a detailed examination of specific aspects of the problem, it simultaneously risks that insights generated in one domain may not hold when factors studied in other domains are taken into account. A reliance on large-scale surveys in organizational research can make it challenging to pin down precisely which micro-level interaction patterns give rise to group-level performance. Sociological models mostly study opinion distributions, making it difficult to draw conclusions about the quality of the opinion distributions at which these populations arrive. And philosophical accounts may overlook the influence of non-epistemic factors that are central to human behavior.

This dissertation aims to bridge these gaps by integrating philosophical and sociological modeling with empirical insights from organizational and social psychology. In doing so, its aim is to outline pathways *how identity diversity and epistemic diversity can be reconciled for better decision-making in groups*.

Chapter overview

The first dissertation chapter provides an overview of the research problem. It brings together different theoretical perspectives and proposes an integrative framework: Decision-making in diverse groups is a complex problem. Researchers need to take into account settings, group-level features of identity diversity and epistemic diversity, deliberation dynamics in groups, and the epistemic consequences of these deliberations. This framework is then used to connect Chapters 2 – 5, which zoom in on specific aspects of group diversity and its consequences for information exchange and decision-making in groups. Chapter 1 synthesizes the findings of these chapters and provides an outlook for future research.

Chapter 2 approaches group diversity and decision-making from a network perspective: It investigates how veracity ratings of informational messages evolve in online bipartisan communities. Simulation-based analysis and a large-scale empirical experiment reveal that ideologically segregated networks undermine users' ability to make accurate rating decisions: Network segregation initiates a dynamic where early biased ratings influence subsequent users' judgments, resulting in a cascade of inaccurate evaluations. Ideologically integrated networks, on the other hand, cause a positive feedback loop that improves overall rating accuracy. Findings show how ideological diversity does not pose a challenge to misinformation detection per se, but rather the network segregation that often comes along with it.

Chapter 3 presents simulated discussions in small teams with subgroups of different professional identities. It examines how homophily, a tendency of individuals to preferentially interact with similar others, affects collective decision-making quality. It shows how homophily initially creates disagreement between subgroups, which extends deliberation time and allows critical arguments to surface and spread. As a result, homophilous teams reach better decisions, while teams without homophily are prone to consensus around inferior options. Findings resonate with accounts of 'transient diversity', suggesting that group problem-solving can benefit from boundaries in how much individuals should influence one another.

Chapter 4 explores how identities affect peer-to-peer learning in problem-solving tasks. A behavioral experiment shows that differing political identities (as opposed to shared ones) are associated with perceptions of reduced competence. However, when identities point to different cognitive styles, perceived dissimilarity enhances attention and influence. This resulted in more frequent problem solving, challenging the common notion that people learn best from those who are similar to themselves.

Chapter 5 investigates how groups can find consensus when subgroup members have stakes in differing decision options. A simulation model shows how consensus depends on both the way members raise arguments and the extent to which

preferences diverge: Conciliatory argumentation styles foster consensus under high divergence, whereas persuasive deliberation increases chances of consensus when subgroup interests are similar. Findings add nuanced insights into the complex dynamics of deliberation in groups with heterogeneous stakes.

Conclusion

The most important insight of this dissertation is that diversity may not always manifest as a double-edged sword, in the sense that different identities necessarily challenge the integration of otherwise valuable epistemic potential. Instead, and as the different dissertation chapters show, the picture is much more nuanced: Identity characteristics that point towards underlying *cognitive differences* can improve learning from dissimilar others. Different *professional identities* can induce constructive disagreement, which improves the decisions of teams in the long run. But *ideological differences* in segregated, bipartisan communities can also backfire and deteriorate the accuracy of their decisions. Lastly, identity diversity can be associated with differences in *interest-based stakes*, which complicates how groups can find agreement.

Whether identity diversity hinders or facilitates the integration of epistemic diversity depends on many factors, such as the type of problem, the specific identities at stake, and the way information is distributed in a group. These factors influence how individuals exchange information with one another, leading to interactional dynamics and decision outcomes that can be challenging to anticipate intuitively.

As my findings illustrate, there are no one-size-fits-all prescriptions for leveraging diversity in group settings. Yet, in an increasingly interconnected world, interactions among individuals with different identities are a constant, and our reliance on others' input continues to grow. Therefore, an understanding of the specific mechanisms through which identity and epistemic diversity impact individual and collective decision-making remains crucial. This dissertation takes us one step closer to reaching that understanding.

Nederlandse samenvatting

Groepsdiversiteit wordt in de wetenschap vaak beschreven als een 'tweesnijdend zwaard' voor collectieve beslissingen. Epistemische diversiteit – oftewel verschillen in perspectieven, vaardigheden en kennis – kan de nauwkeurigheid, creativiteit en effectiviteit van groepsbesluiten verbeteren. Tegelijkertijd gaan epistemische verschillen tussen groepsleden vaak gepaard met identiteitsverschillen. Verschillende achtergronden – cultureel, ideologisch, of professioneel – kunnen het lastig maken om het wederzijdse vertrouwen en de open communicatie op te bouwen die nodig zijn om diversiteit effectief te benutten. Diversiteit binnen groepen biedt dus potentieel, maar dat potentieel wordt niet automatisch benut.

Verschillende wetenschappelijke disciplines hebben geprobeerd manieren te identificeren om de voordelen van epistemische diversiteit te benutten, terwijl de valkuilen van identiteitsdiversiteit worden vermeden. Organisatieonderzoek heeft veel empirische data verzameld over de demografische samenstelling van teams en hun prestaties. Sociologische simulatiemodellen laten zien hoe kleinschalige interactiepatronen tussen individuen grootschalige fenomenen zoals polarisatie kunnen veroorzaken. En formeel-filosofisch werk in de sociale epistemologie heeft onderzocht onder welke voorwaarden groepen epistemisch gerechtvaardigde en waarheidsgetrouwe beslissingen kunnen nemen.

Hoewel onderzoek binnen deze disciplines een gedetailleerde analyse van specifieke aspecten van het probleem mogelijk maakt, bestaat tegelijkertijd het risico dat inzichten uit het ene vakgebied niet standhouden wanneer factoren uit andere domeinen worden meegenomen. Een sterke focus op grootschalige enquêtes in organisatieonderzoek maakt het lastig om precies vast te stellen hoe specifieke interactiepatronen op microniveau prestaties op groepsniveau beïnvloeden. Sociologische modellen bestuderen vooral opinieverspreiding, waardoor het moeilijk is uitspraken te doen over de kwaliteit van de uiteindelijke opinies. Filosofische benaderingen ten slotte houden niet altijd rekening met de invloed van niet-epistemische factoren die centraal staan in menselijk gedrag.

Dit proefschrift beoogt een brug te slaan tussen filosofische en sociologische modellen enerzijds, en empirische inzichten uit organisatieonderzoek en de sociale psychologie anderzijds. Daarmee onderzoekt het *hoe identiteitsdiversiteit en epistemische diversiteit met elkaar verenigd kunnen worden, om zo besluitvorming in groepen te verbeteren*.

Hoofdstukoverzicht

Het eerste hoofdstuk van het proefschrift biedt een overzicht van de probleemstelling. Het brengt verschillende theoretische perspectieven samen en introduceert een integratief theoretisch kader: besluitvorming in diverse groepen is een complex probleem. Onderzoekers moeten rekening houden met de context, groepseigenschappen van identiteits- en epistemische diversiteit, deliberatiedynamieken binnen groepen en de epistemische gevolgen van deze deliberaties. Dit kader vormt de leidraad waarmee de hoofdstukken 2 t/m 5 met elkaar worden verbonden. Elk hoofdstuk zoomt in op specifieke aspecten van groepsdiversiteit en de gevolgen daarvan voor informatie-uitwisseling en besluitvorming. Dit eerste hoofdstuk vat de belangrijkste bevindingen samen, plaatst ze in onderlinge samenhang en werpt een blik op toekomstige onderzoeksvragen.

Hoofdstuk 2 benadert groepsdiversiteit en besluitvorming vanuit een netwerkperspectief. Het onderzoekt hoe gebruikers van sociale media elkaar onderling beïnvloeden in het beoordelen van het waarheidsgehalte van berichten, en hoe dit zich ontwikkelt in online gemeenschappen met twee ideologische kampen. Simulaties en een grootschalig empirisch experiment tonen aan dat ideologisch gesegregeerde netwerken het vermogen van gebruikers ondermijnen om accuraat te oordelen: segregatie leidt tot een dynamiek waarin vroege bevooroordeelde beoordelingen die van latere gebruikers beïnvloeden, met als gevolg een kettingreactie van onjuiste beoordelingen. Ideologisch geïntegreerde netwerken creëren daarentegen een positieve feedbacklus die de algehele nauwkeurigheid verbetert. De bevindingen laten zien dat ideologische verschillen op zichzelf geen probleem vormen voor misinformatieherkenning, maar eerder de netwerksegregatie die er vaak mee gepaard gaat.

Hoofdstuk 3 simuleert discussies in kleine teams met subgroepen van verschillende professionele identiteiten. Het onderzoekt hoe de neiging van individuen om vooral met anderen uit hun subgroep om te gaan, de kwaliteit van collectieve besluitvorming beïnvloedt. Een voorkeur voor individuen met dezelfde professionele identiteit leidt aanvankelijk tot meningsverschillen tussen subgroepen, wat de deliberatieduur verlengt en ruimte geeft voor kritische discussie. Daardoor nemen deze teams betere besluiten, terwijl teams waarin iedereen even vaak met elkaar omgaat sneller tot consensus komen rond inferieure opties. De bevindingen sluiten aan bij theorieën over ‘transiënte diversiteit’, die suggereren dat collectieve besluitvorming kan profiteren van begrensde wederzijdse beïnvloeding.

Hoofdstuk 4 onderzoekt hoe identiteit het onderlinge leren beïnvloedt bij probleemoplossingstaken. Een gedragsexperiment laat zien dat bij verschillende politieke wereldbeschouwingen (in tegenstelling tot gedeelde) deelnemers elkaars competentie lager inschatten. Echter, wanneer identiteiten verschillen in ‘cognitieve

stijl' (vastgesteld door een fictieve persoonlijkheidstest), leidt waargenomen verschil juist tot verwachtingen van nieuwe inzichten, hetgeen onderlinge aandacht en invloed versterkt. Cognitieve verschillen blijken als gevolg daarvan de onderlinge leereffectiviteit te vergroten, wat de gangbare opvatting ter discussie stelt dat men vooral leert van gelijkgestemden.

Hoofdstuk 5 onderzoekt hoe groepen consensus kunnen bereiken wanneer subgroepleden belangen hebben bij verschillende besluitopties. Een simulatiemodel laat zien dat consensus afhankelijk is van zowel discussiestijl als de mate van belangentegenstelling: coöperatieve, bemiddelende argumentatiestijlen bevorderen consensus bij sterke voorkeursverschillen, terwijl persuasieve argumentatie de kans op consensus verhoogt wanneer de belangen van subgroepen gelijkwaardig zijn. De bevindingen bieden genuanceerde inzichten in de complexe dynamiek van discussies binnen groepen met uiteenlopende belangen.

Conclusie

De belangrijkste bevinding van dit proefschrift is dat diversiteit niet per definitie een tweesnijdend zwaard vormt, waarbij verschillende identiteiten de benutting van waardevol epistemisch potentieel noodzakelijkerwijs belemmeren. Zoals de individuele hoofdstukken laten zien, is het beeld genuanceerder: verschillen in identiteit gebaseerd op cognitieve stijl kunnen leiden tot verwachtingen van nieuwheid en daarmee het leren van andersdenkenden bevorderen. Verschillende professionele identiteiten kunnen constructieve meningsverschillen creëren, wat op de lange termijn tot betere besluiten leidt. Daarentegen kunnen ideologische verschillen tussen gesegregeerde en gepolariseerde kampen juist contraproductief zijn en de kwaliteit van besluiten ondermijnen. Tot slot kunnen identiteitsverschillen ook van invloed zijn op de manier waarop argumenten en keuzemogelijkheden worden beoordeeld, wat het bereiken van consensus in een groep bemoeilijkt.

Of identiteitsdiversiteit de benutting van epistemische diversiteit belemmert of juist bevordert, hangt af van tal van factoren: het soort probleem, de context waarin communicatie plaatsvindt, de specifieke identiteiten in kwestie, en hoe informatie binnen een groep is verdeeld. Al deze factoren beïnvloeden hoe individuen informatie met elkaar uitwisselen, en leiden tot interactiedynamieken en besluituitkomsten die niet eenvoudig te voorspellen zijn.

Zoals deze bevindingen illustreren, bestaan er geen universele recepten om diversiteit in groepsverband optimaal te benutten. Toch zijn interacties tussen mensen met verschillende identiteiten in een steeds meer verbonden wereld een constante realiteit en zijn we in toenemende mate afhankelijk van de inbreng van anderen. De specifieke mechanismen begrijpen waardoor identiteits- en epistemische diversiteit individuele en collectieve besluitvorming beïnvloeden is daarom van groot belang. Dit proefschrift brengt ons een stap dichter bij dat begrip.

References

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36), eabf4393. https://doi.org/10.1126/sciadv.abf4393

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), eaay3539. https://doi.org/10.1126/sciadv.aay3539

Allen, J., Martel, C., & Rand, D. G. (2022). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. *CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3502040

Aminpour, P., Gray, S. A., Singer, A., Scyphers, S. B., Jetter, A. J., Jordan, R., Murphy, R., & Grabowski, J. H. (2021). The diversity bonus in pooling local knowledge about complex problems. *Proceedings of the National Academy of Sciences*, *118*(5), e2016887118. https://doi.org/10.1073/pnas.2016887118

ANES. (2020). *2020 Time Series Study*. American National Election Studies. https://electionstudies.org/data-center/2020-time-series-study/

Antonio, A. L., Chang, M. J., Hakuta, K., Kenny, D. A., Levin, S., & Milem, J. F. (2004). Effects of Racial Diversity on Complex Thinking in College Students. *Psychological Science*, *15*(8), 507–510. https://doi.org/10.1111/j.0956-7976.2004.00710.x

Aoki, K., & Feldman, M. W. (2014). Evolution of learning strategies in temporally and spatially variable environments: A review of theory. *Theoretical Population Biology*, *91*, 3–19. https://doi.org/10.1016/j.tpb.2013.10.004

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70. https://doi.org/10.1037/h0093718

Assaad, L., Fuchs, R., Jalalimanesh, A., Phillips, K., Schoeppl, L., & Hahn, U. (2023). *A Bayesian Agent-Based Framework for Argument Exchange Across Networks* (No. arXiv:2311.09254). arXiv. https://doi.org/10.48550/arXiv.2311.09254

Axelrod, R. (1997). *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press. https://doi.org/10.1515/9781400822300

Axtell, R., Axelrod, R., Epstein, J. M., & Cohen, M. D. (1996). Aligning simulation models: A case study and results. *Computational & Mathematical Organization Theory*, *1*(2), 123–141. https://doi.org/10.1007/BF01299065

Bail, C. A. (2016). Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences*, *113*(42), 11823–11828. https://doi.org/10.1073/pnas.1607151113

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *PNAS*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., & Volfovsky, A. (2020). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the National Academy of Sciences*, *117*(1), 243–250. https://doi.org/10.1073/pnas.1906420116

Baker, K. M. (1975). *Condorcet, From Natural Philosophy to Social Mathematics*. University of Chicago Press.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.

Baldassarri, D., & Bearman, P. (2007). Dynamics of political polarization. *American Sociological Review*, *72*(5), 784–811.

Baldini, R. (2013). Two success-biased social learning strategies. *Theoretical Population Biology*, *86*, 43–49. https://doi.org/10.1016/j.tpb.2013.03.005

Balietti, S., Getoor, L., Goldstein, D. G., & Watts, D. J. (2021). Reducing opinion polarization: Effects of exposure to similar people with differing political views. *Proceedings of the National Academy of Sciences*, *118*(52), e2112552118. https://doi.org/10.1073/pnas.2112552118

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, *114*(26), E5070–E5076. https://doi.org/10.1073/pnas.1615978114

Becker, J., Guilbeault, D., & Smith, E. B. (2022). The crowd classification problem: Social dynamics of binary-choice accuracy. *Management Science*, *68*(5), 3949–3965. https://doi.org/10.1287/mnsc.2021.4127

Bell, S. T., Villado, A. J., Lukasik, M. A., Belau, L., & Briggs, A. L. (2011). Getting Specific about Demographic Diversity Variable and Team Performance Relationships: A Meta-Analysis. *Journal of Management*, *37*(3), 709–743. https://doi.org/10.1177/0149206310365001

Bernstein, E., Shore, J., & Lazer, D. (2018). How intermittent breaks in interaction improve collective intelligence. *Proceedings of the National Academy of Sciences*, *115*(35), 8734–8739. https://doi.org/10.1073/pnas.1802407115

Betz, G. (2022). Natural-Language Multi-Agent Simulations of Argumentative Opinion Dynamics. *Journal of Artificial Societies and Social Simulation*, *25*(1), 2.

Bianchi, F., & Squazzoni, F. (2015). Agent-based models in sociology. *WIREs Computational Statistics*, *7*(4), 284–306. https://doi.org/10.1002/wics.1356

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, *100*(5), 992–1026. https://doi.org/10.1086/261849

Blume, L. E., Brock, W. A., Durlauf, S. N., & Ioannides, Y. M. (2011). Identification of Social Interactions. In J. Benhabib, A. Bisin, & M. O. Jackson (Eds.), *Handbook of Social Economics* (Vol. 1, pp. 853–964). North-Holland. https://doi.org/10.1016/B978-0-444-53707-2.00001-3

Bol, T., de Vaan, M., & van de Rijt, A. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences*, *115*(19), 4887–4890.

Borah, P. (2022). The moderating role of political ideology: Need for cognition, media locus of control, misinformation efficacy, and misperceptions about COVID-19. *International Journal of Communication*, *16*, 26.

Boutyline, A., & Willer, R. (2017). The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology*, *38*(3), 551–569.

Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago press.

Brandom, R. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard university press.

Brechwald, W. A., & Prinstein, M. J. (2011). Beyond Homophily: A Decade of Advances in Understanding Peer Influence Processes. *Journal of Research on Adolescence*, *21*(1), 166–179. https://doi.org/10.1111/j.1532-7795.2010.00721.x

Camerer, C. F. (2011). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Carter, A. B., & Phillips, K. W. (2017). The double-edged sword of diversity: Toward a dual pathway model. *Social and Personality Psychology Compass*, *11*(5), e12313. https://doi.org/10.1111/spc3.12313

Centola, D. (2010). The Spread of Behavior in an Online Social Network Experiment. *Science*, *329*(5996), 1194–1197. https://doi.org/10.1126/science.1185231

Chartrand, T. L., & Lakin, J. L. (2013). The Antecedents and Consequences of Human Behavioral Mimicry. *Annual Review of Psychology*, *64*, 285–308. https://doi.org/10.1146/annurev-psych-113011-143754

Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, *55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9). https://doi.org/10.1073/pnas.2023301118

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, *64*(2), 317–332. https://doi.org/10.1111/jcom.12084

Condorcet, M. J. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix* (Vol. 252). American Mathematical Soc.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. *Proceedings of the International Aaai Conference on Web and Social Media*, *5*(1), 89–96.

Cooper, R. (1999). *Coordination games*. Cambridge University Press.

Da, Z. & Xing Huang. (2020). Harnessing the Wisdom of Crowds. *Management Science*, *66*(5), 1847–1867. https://doi.org/10.1287/mnsc.2019.3294

Davis, D. D., & Holt, C. A. (1993). *Experimental Economics*. Princeton University Press.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, *03*(01n04), 87–98. https://doi.org/10.1142/S0219525900000078

Degroot, M. H. (1974). Reaching a Consensus. *Journal of the American Statistical Association*, *69*(345), 118–121. https://doi.org/10.1080/01621459.1974.10480137

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

Del Vicario, M., Quattrociocchi, W., Scala, A., & Zollo, F. (2019). Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web*, *13*(2), 1–22. https://doi.org/10.1145/3316809

DellaPosta, D., Shi, Y., & Macy, M. (2015). Why do liberals drink lattes? *American Journal of Sociology*, *120*(5), 1473–1511.

Derex, M., & Boyd, R. (2016). Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*, *113*(11), 2982–2987. https://doi.org/10.1073/pnas.1518798113

Dessel, A., & Rogge, M. E. (2008). Evaluation of intergroup dialogue: A review of the empirical literature. *Conflict Resolution Quarterly*, *26*(2), 199–238. https://doi.org/10.1002/crq.230

Diekmann, A. (2023). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen*. Rowohlt Verlag GmbH.

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021). How Affective Polarization Shapes Americans' Political Beliefs: A Study of Response to the COVID-19 Pandemic. *Journal of Experimental Political Science*, *8*(3), 223–234. https://doi.org/10.1017/XPS.2020.28

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). *Improving Factuality and Reasoning in Language Models through Multiagent Debate* (No. arXiv:2305.14325). arXiv. https://doi.org/10.48550/arXiv.2305.14325

Eady, G., Nagler, J., Guess, A., Zilinsky, J., & Tucker, J. A. (2019). How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *Sage Open*, *9*(1), 2158244019832705.

Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory and Cognition*, *38*(8), 1087–1100.

Efferson, C., Lalive, R., & Fehr, E. (2008). The Coevolution of Cultural Groups and Ingroup Favoritism. *Science*, *321*(5897), 1844–1849. https://doi.org/10.1126/science.1155805

Ehret, S., Constantino, S. M., Weber, E. U., Efferson, C., & Vogt, S. (2022). Group identities can undermine social tipping after intervention. *Nature Human Behaviour*, *6*(12), 1669–1679. https://doi.org/10.1038/s41562-022-01440-5

Ellemers, N. (2012). The Group Self. *Science*. https://doi.org/10.1126/science.1220987

Ellemers, N., Spears, R., & Doosje, B. (2002). Self and Social Identity. *Annual Review of Psychology*, *53*(Volume 53, 2002), 161–186. https://doi.org/10.1146/annurev.psych.53.100901.135228

Ellison, G. (1993). Learning, local interaction, and coordination. *Econometrica: Journal of the Econometric Society*, 1047–1071.

Epstein, J. M. (2012). *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press. https://www.degruyterbrill.com/document/doi/10.1515/9781400842872/html

Epstein, Z., Lin, H., Pennycook, G., & Rand, D. (2022). *How many others have shared this? Experimentally investigating the effects of social cues on engagement, misinformation, and unpredictability on social media* (No. arXiv:2207.07562). arXiv. https://doi.org/10.48550/arXiv.2207.07562

Ertug, G., Brennecke, J., Kovács, B., & Zou, T. (2022). What Does Homophily Do? A Review of the Consequences of Homophily. *Academy of Management Annals*, *16*(1), 38–69. https://doi.org/10.5465/annals.2020.0230

Estévez-Mujica, C. P., Acero, A., Jiménez-Leal, W., & García-Díaz, C. (2018). The Influence of Homophilous Interactions on Diversity Effects in Group Problem-Solving. *Nonlinear Dynamics, Psychology & Life Sciences*, *22*(1).

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1–20. https://doi.org/10.3758/s13428-021-01694-3

Fehr, E., & Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, *90*(4), 980–994. https://doi.org/10.1257/aer.90.4.980

Feld, S. L. (1981). The Focused Organization of Social Ties. *American Journal of Sociology*, *86*(5), 1015–1035. https://doi.org/10.1086/227352

Feliciani, T. (2025). *Divided spaces and divided opinions: Modeling the impact of residential segregation on opinion polarization.* ICS dissertation.

Feliciani, T., Flache, A., & Mäs, M. (2021). Persuasion without polarization? Modelling persuasive argument communication in teams with strong faultlines. *Computational and Mathematical Organization Theory*, *27*(1), 61–92. https://doi.org/10.1007/s10588-020-09315-8

Flache, A. (2018). Between Monoculture and Cultural Polarization: Agent-based Models of the Interplay of Social Influence and Cultural Diversity. *Journal of Archaeological Method and Theory*, *25*(4), 996–1023. https://doi.org/10.1007/s10816-018-9391-1

Flache, A., & Macy, M. W. (2011). Small worlds and cultural polarization. *The Journal of Mathematical Sociology*, *35*(1–3), 146–176.

Flache, A., & Mäs, M. (2008). How to get the timing right. A computational model of the effects of the timing of contacts on team cohesion in demographically diverse teams. *Computational and Mathematical Organization Theory*, *14*(1), 23–51.

Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of Social Influence: Towards the Next Frontiers. *Journal of Artificial Societies and Social Simulation*, *20*(4), 2.

Flache, A., Mäs, M., & Keijzer, M. (2022). Computational approaches in rigorous sociology: Agent-based computational modeling and computational social science. *Handbook of Sociological Science*, 57–72.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*(S1), 298–320.

Fossett, M. (2006). Ethnic Preferences, Social Distance Dynamics, and Residential Segregation: Theoretical Explorations Using Simulation Analysis∗. *The Journal of Mathematical Sociology*, *30*(3–4), 185–273. https://doi.org/10.1080/00222500500544052

Frey, D., & Šešelja, D. (2020). Robustness and Idealizations in Agent-Based Models of Scientific Interaction. *The British Journal for the Philosophy of Science*, *71*(4), 1411–1437. https://doi.org/10.1093/bjps/axy039

Frey, V., Flache, A., Bakker, D., & Mäs, M. (2024). Who influences lower-status individuals more: People of higher-status outgroups or people of their lower-status ingroup? Examining the difference between matters of opinion and matters of fact. *Social Science Research*, *123*, 103060. https://doi.org/10.1016/j.ssresearch.2024.103060

Frey, V., & van de Rijt, A. (2021). Social Influence Undermines the Wisdom of the Crowd in Sequential Decision Making. *Management Science*, *67*(7), 4273–4286. https://doi.org/10.1287/mnsc.2020.3713

Friedkin, N. E., & Bullo, F. (2017). How truth wins in opinion dynamics along issue sequences. *Proceedings of the National Academy of Sciences*, *114*(43), 11380–11385.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*(1), 35–42. https://doi.org/10.1016/j.socec.2010.10.008

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

Gërxhani, K., & Miller, L. (2022). Experimental sociology. In K. Gërxhani, N. de Graf, & R. Werner (Eds.), *Handbook of Sociological Science* (pp. 309–323). Edward Elgar Publishing. https://www.elgaronline.com/edcollchap-oa/book/9781789909432/book-part-9781789909432-26.xml

Gilbert, N., & Troitzsch, K. (2005). *Simulation for the social scientist*. McGraw-Hill Education (UK).

Goeree, J. K., Palfrey, T. R., Rogers, B. W., & McKelvey, R. D. (2007). Self-correcting information cascades. *Review of Economic Studies*, *74*(3), 733–762.

González-Bailón, S., Lazer, D., Barberá, P., Zhang, M., Allcott, H., Brown, T., Crespo-Tenorio, A., Freelon, D., Gentzkow, M., Guess, A. M., Iyengar, S., Kim, Y. M., Malhotra, N., Moehler, D., Nyhan, B., Pan, J., Rivera, C. V., Settle, J., Thorson, E., … Tucker, J. A. (2023). Asymmetric ideological segregation in exposure to political news on Facebook. *Science*, *381*(6656), 392–398. https://doi.org/10.1126/science.ade7138

Greene, W. (2009). Discrete Choice Modeling. In T. C. Mills & K. Patterson (Eds.), *Palgrave Handbook of Econometrics: Volume 2: Applied Econometrics* (pp. 473–556). Palgrave Macmillan UK. https://doi.org/10.1057/9780230244405_11

Gross, J., Méder, Z. Z., De Dreu, C. K. W., Romano, A., Molenmaker, W. E., & Hoenig, L. C. (2023). The evolution of universal cooperation. *Science Advances*, *9*(7), eadd8289. https://doi.org/10.1126/sciadv.add8289

Grossi, D., Hahn, U., Mäs, M., Nitsche, A., Behrens, J., Boehmer, N., Brill, M., Endriss, U., Grandi, U., Haret, A., Heitzig, J., Janssens, N., Jonker, C. M., Keijzer, M., Kistner, A., Lackner, M., Lieben, A., Mikhaylovskaya, A., Murukannaiah, P. K., … Putte, F. V. D. (2024). *Enabling the Digital Democratic Revival: A Research Program for Digital Democracy* (No. arXiv:2401.16863). arXiv. https://doi.org/10.48550/arXiv.2401.16863

Grow, A., & Flache, A. (2019). Models of Reputation and Status Dynamics. In F. Giardini & R. Wittek (Eds.), *The Oxford handbook of gossip and reputation* (pp. 230–249). Oxford University Press.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Sci. Adv.*, *5*(1), eaau4586. https://doi.org/10.1126/sciadv.aau4586

Guilbeault, D., Becker, J., & Centola, D. (2018). Social learning and partisan bias in the interpretation of climate trends. *Proceedings of the National Academy of Sciences*, *115*(39), 9714–9719. https://doi.org/10.1073/pnas.1722664115

Guilbeault, D., van Loon, A., Lix, K., Goldberg, A., & Srivastava, S. B. (2024). Exposure to the Views of Opposing Others with Latent Cognitive Differences Results in Social Influence—But Only When Those Differences Remain Obscured. *Management Science*, *70*(10), 6669–6684. https://doi.org/10.1287/mnsc.2022.00895

Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. (2019). *The future of misinformation detection: New perspectives and trends*. http://arxiv.org/abs/1909.03654

Guo, D., & Yu, A. J. (2019). Human learning and decision-making in the bandit task: Three wrongs make a right. *Conference on Cognitive Computational Neuroscience*. https://pdfs.semanticscholar.org/decc/a47db4e33ad70ffadeadf622108aa9ec69f8.pdf

Habermas, J. (1985a). *The theory of communicative action: Volume 1: Reason and the rationalization of society* (Vol. 1). Beacon press.

Habermas, J. (1985b). *The theory of communicative action: Volume 2: Lifeword and system: A critique of functionalist reason* (Vol. 2). Beacon press.

Hahn, U., & Harris, A. J. (2014). What does it mean to be biased: Motivated reasoning and rationality. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 41–102). Elsevier.

Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Knopf Doubleday Publishing Group.

Hall, E. T. (1976). *Beyond culture*. Doubleday.

Harlé, K. M., Zhang, S., Schiff, M., Mackey, S., Paulus, M. P., & Yu, A. J. (2015). Altered Statistical Learning and Decision-Making in Methamphetamine Dependence: Evidence from a Two-Armed Bandit Task. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01910

Harrison, D. A., Price, K. H., & Bell, M. P. (1998). Beyond Relational Demography: Time and the Effects of Surface- and Deep-Level Diversity on Work Group Cohesion. *Academy of Management Journal*, *41*(1), 96–107. https://doi.org/10.5465/256901

Harrison, D. A., Price, K. H., Gavin, J. H., & Florey, A. T. (2002). Time, Teams, and Task Performance: Changing Effects of Surface- and Deep-Level Diversity on Group Functioning. *Academy of Management Journal*, *45*(5), 1029–1045. https://doi.org/10.5465/3069328

Hedström, P. (2006). Experimental Macro Sociology: Predicting the Next Best Seller. *Science*, *311*(5762), 786–787. https://doi.org/10.1126/science.1124707

Hedström, P., & Bearman, P. (2011). *The Oxford Handbook of Analytical Sociology*. OUP Oxford.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, *5*(3).

Hegselmann, R., & Krause, U. (2006). Truth and cognitive division of labor: First steps towards a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation*, *9*(3), 10.

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *12*(3), 123–135. https://doi.org/10.1002/evan.10110

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54. https://doi.org/10.1016/j.tics.2014.10.004

Hogg, M. A. (2006). Social identity theory. *Contemporary Social Psychological Theories*, *13*, 111–1369.

Hogg, M. A. (2016). Social Identity Theory. In S. McKeown, R. Haji, & N. Ferguson (Eds.), *Understanding Peace and Conflict Through Social Identity Theory: Contemporary Global Perspectives* (pp. 3–17). Springer International Publishing. https://doi.org/10.1007/978-3-319-29869-6_1

Hogg, M. A., Turner, J. C., & Davidson, B. (1990). Polarized Norms and Social Frames of Reference: A Test of the Self-Categorization Theory of Group Polarization. *Basic and Applied Social Psychology*, *11*(1), 77–100. https://doi.org/10.1207/s15324834basp1101_6

Homan, A. C., Van Knippenberg, D., Van Kleef, G. A., & De Dreu, C. K. (2007). Bridging faultlines by valuing diversity: Diversity beliefs, information elaboration, and performance in diverse work groups. *Journal of Applied Psychology*, *92*(5), 1189.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

Horwitz, S. K., & Horwitz, I. B. (2007). The Effects of Team Diversity on Team Outcomes: A Meta-Analytic Review of Team Demography. *Journal of Management*, *33*(6), 987–1015. https://doi.org/10.1177/0149206307308587

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Jackson, M., & Cox, D. R. (2013). The Principles of Experimental Design and Their Application in Sociology. *Annual Review of Sociology*, *39*(Volume 39, 2013), 27–49. https://doi.org/10.1146/annurev-soc-071811-145443

Jackson, S. E., May, K. E., Whitney, K., Guzzo, R. A., & Salas, E. (1995). Understanding the dynamics of diversity in decision-making teams. *Team Effectiveness and Decision Making in Organizations*, *204*, 261.

Joshi, A., & Roh, H. (2009). The Role Of Context In Work Team Diversity Research: A Meta-Analytic Review. *Academy of Management Journal*, *52*(3), 599–627. https://doi.org/10.5465/amj.2009.41331491

Jun, Y., Meng, R., & Johar, G. V. (2017). Perceived social presence reduces fact-checking. *PNAS*, *114*(23), 5976–5981. https://doi.org/10.1073/pnas.1700175114

Kahneman, D., & Tversky, A. (2012). Prospect Theory: An Analysis of Decision Under Risk. In *Handbook of the Fundamentals of Financial Decision Making: Vol. Volume 4* (pp. 99–127). World Scientific. https://doi.org/10.1142/9789814417358_0006

Keijzer, M. (2022). *Opinion Dynamics in Online Social Media. Chapter 1 "The Complex Dynamics of Polarization in Online Social Media."* ICS dissertation.

Keijzer, M., Mäs, M., & Flache, A. (2018). Communication in Online Social Networks Fosters Cultural Isolation. *Complexity*, *2018*, e9502872. https://doi.org/10.1155/2018/9502872

Keusch, F., & Kreuter, F. (2021). Digital trace data. In U. Engel, A. Quan-Haase, S. X. Liu, & L. Lyberg (Eds.), *Handbook of Computational Social Science, Volume 1*. Taylor & Francis. https://doi.org/10.4324/9781003024583-8

Keuschnigg, M., & Ganser, C. (2017). Crowd wisdom relies on agents' ability in small groups with a voting aggregation rule. *Management Science*, *63*(3), 818–828.

Kim, A., Moravec, P. L., & Dennis, A. R. (2019). Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems*, *36*(3), 931–968. https://doi.org/10.1080/07421222.2019.1628921

Kinzler, K. D. (2021). Language as a Social Cue. *Annual Review of Psychology*, *72*(1), 241–264. https://doi.org/10.1146/annurev-psych-010418-103034

Kirvan, P. (2025, January 24). *What is red teaming? | Definition from TechTarget*. WhatIs. https://www.techtarget.com/whatis/definition/red-teaming

Kossinets, G., & Watts, D. J. (2009). Origins of Homophily in an Evolving Social Network. *American Journal of Sociology*, *115*(2), 405–450. https://doi.org/10.1086/599247

Kretschmer, D., Leszczensky, L., & McMillan, C. (2024). Strong ties, strong homophily? Variation in homophily on sociodemographic characteristics by relationship strength. *Social Forces*, soae169. https://doi.org/10.1093/sf/soae169

Lau, D. C., & Murnighan, J. K. (1998). Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of Management Review*, *23*(2), 325–340.

Lazer, D., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lazer, D., & Friedman, A. (2007). The Network Structure of Exploration and Exploitation. *Administrative Science Quarterly*, *52*(4), 667–694. https://doi.org/10.2189/asqu.52.4.667

Levine, S. S., Apfelbaum, E. P., Bernard, M., Bartelt, V. L., Zajac, E. J., & Stark, D. (2014). Ethnic diversity deflates price bubbles. *Proceedings of the National Academy of Sciences*, *111*(52), 18524–18529. https://doi.org/10.1073/pnas.1407301111

Levinthal, D. A. (1997). Adaptation on Rugged Landscapes. *Management Science*, *43*(7), 934–950. https://doi.org/10.1287/mnsc.43.7.934

Levy, G. (2007). Decision making in committees: Transparency, reputation, and voting rules. *American Economic Review*, *97*(1), 150–168.

Levy, G., & Razin, R. (2019). Echo Chambers and Their Effects on Economic and Political Outcomes. *Annual Review of Economics*, *11*(1), 303–328. https://doi.org/10.1146/annurev-economics-080218-030343

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Lipatov, M., Illari, L., Johnson, N. F., & Gavrilets, S. (2025). Coevolution of network and attitudes under competing propaganda machines. *Npj Complexity*, *2*(1), 1–14. https://doi.org/10.1038/s44260-025-00033-3

Liquin, E. G., & Gopnik, A. (2022). Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, *218*, 104940. https://doi.org/10.1016/j.cognition.2021.104940

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

Lu, L., Yuan, Y. C., & McLeod, P. L. (2012). Twenty-Five Years of Hidden Profiles in Group Decision Making: A Meta-Analysis. *Personality and Social Psychology Review*, *16*(1), 54–75. https://doi.org/10.1177/1088868311417243

Macy, M. W., & Willer, R. (2002). From Factors to Actors: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology*, *28*, 143–166.

Madsen, J. K., Bailey, R. M., & Pilditch, T. D. (2018). Large networks of rational agents form persistent echo chambers. *Scientific Reports*, *8*(1), Article 1. https://doi.org/10.1038/s41598-018-25558-7

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*, 1–16. https://doi.org/10.1037/xap0000315

March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, *2*(1), 71–87. https://doi.org/10.1287/orsc.2.1.71

Marchi, S. de, & Page, S. E. (2014). Agent-Based Models. *Annual Review of Political Science*, *17*(Volume 17, 2014), 1–20. https://doi.org/10.1146/annurev-polisci-080812-191558

Mark, N. P. (2003). Culture and Competition: Homophily and Distancing Explanations for Cultural Niches. *American Sociological Review*, *68*(3), 319–345. https://doi.org/10.2307/1519727

Mäs, M., & Flache, A. (2013). Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PloS One*, *8*(11), e74516.

Mäs, M., Flache, A., Takács, K., & Jehn, K. A. (2013). In the short term we divide, in the long term we unite: Demographic crisscrossing and the effects of faultlines on subgroup polarization. *Organization Science*, *24*(3), 716–736.

Mason, L. (2018). *Uncivil Agreement: How Politics Became Our Identity*. University of Chicago Press.

Mason, W., & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, *109*(3), 764–769. https://doi.org/10.1073/pnas.1110069108

McElreath, R., Wallin, A., & Fasolo, B. (2013). The evolutionary rationality of social learning. In R. Hertwig & U. Hoffrage (Eds.), *Simple heuristics in a social world* (pp. 381–408). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195388435.003.0014

McLeod, P. L., Lobel, S. A., & Cox, T. H. (1996). Ethnic Diversity and Creativity in Small Groups. *Small Group Research*, *27*(2), 248–264. https://doi.org/10.1177/1046496496272003

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968

Merton, R. K. (1949). *Social Theory and Social Structure*. Simon and Schuster.

Meta. (2025, March 13). Testing Begins for Community Notes on Facebook, Instagram and Threads. *Meta*. https://about.fb.com/news/2025/03/testing-begins-community-notes-facebook-instagram-threads/

Milliken, F. J., & Martins, L. L. (1996). Searching for Common Threads: Understanding the Multiple Effects of Diversity in Organizational Groups. *Academy of Management Review*, *21*(2), 402–433. https://doi.org/10.5465/amr.1996.9605060217

Mouffe, C. (1999). Deliberative democracy or agonistic pluralism? *Social Research*, 745–758.

Muise, D., Hosseinmardi, H., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2022). Quantifying partisan news diets in Web and TV audiences. *Science Advances*, *8*(28), eabn0083. https://doi.org/10.1126/sciadv.abn0083

Newman, M. E. J., & Watts, D. J. (1999). Scaling and percolation in the small-world network model. *Physical Review E*, *60*(6), 7332–7342. https://doi.org/10.1103/PhysRevE.60.7332

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Northcraft, G. B., Polzer, J. T., Neale, M. A., & Kramer, R. M. (1995). B. In *Diversity in work teams: Research paradigms for a changing workplace* (pp. 69–96). American Psychological Association. https://doi.org/10.1037/10189-003

Nowak, A., & Vallacher, R. R. (2019). Nonlinear societal change: The perspective of dynamical systems. *British Journal of Social Psychology*, *58*(1), 105–128. https://doi.org/10.1111/bjso.12271

Nowak, M. A. (2006). Five Rules for the Evolution of Cooperation. *Science*, *314*(5805), 1560–1563. https://doi.org/10.1126/science.1133755

O'Connor, C., Goldberg, S., & Goldman, A. (2024). Social Epistemology. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2024). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2024/entries/epistemology-social/

O'Connor, C., & Weatherall, J. O. (2018). Scientific polarization. *European Journal for Philosophy of Science*, *8*(3), 855–875. https://doi.org/10.1007/s13194-018-0213-9

Olsson, E. J. (2013). A Bayesian Simulation Model of Group Deliberation and Polarization. In F. Zenker (Ed.), *Bayesian Argumentation: The practical side of probability* (pp. 113–133). Springer. https://doi.org/10.1007/978-94-007-5357-0_6

Page, S. (2019). *The Diversity Bonus: How Great Teams Pay Off in the Knowledge Economy*. Princeton University Press. https://doi.org/10.1515/9780691193823

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), Article 7855. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019a). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526. https://doi.org/10.1073/pnas.1806781116

Pennycook, G., & Rand, D. G. (2019b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *J Pers*, *88*(2), 185–200. https://doi.org/10.1111/jopy.12476

Peters, U. (2021). Hidden figures: Epistemic costs and benefits of detecting (invisible) diversity in science. *European Journal for Philosophy of Science*, *11*(1), 33. https://doi.org/10.1007/s13194-021-00349-6

Phillips, K. W. (2003). The Effects of Categorically Based Expectations on Minority Influence: The Importance of Congruence. *Personality and Social Psychology Bulletin*, *29*(1), 3–13. https://doi.org/10.1177/0146167202238367

Phillips, K. W., & Loyd, D. L. (2006). When surface and deep-level diversity collide: The effects on dissenting group members. *Organizational Behavior and Human Decision Processes*, *99*(2), 143–160. https://doi.org/10.1016/j.obhdp.2005.12.001

Phillips, K. W., Northcraft, G. B., & Neale, M. A. (2006). Surface-Level Diversity and Decision-Making in Groups: When Does Deep-Level Similarity Help? *Group Processes & Intergroup Relations*, *9*(4), 467–482. https://doi.org/10.1177/1368430206067557

Phillips, K. W., & O'Reilly, C. (1998). Demography and Diversity in Organizations: A Review of 40 Years of Research. In *Research in Organizational Behavior* (Vol. 20, pp. 77–140).

Pretus, C., Javeed, A., Hughes, D. R., Hackenburg, K., Tsakiris, M., Vilarroya, O., & Bavel, J. J. V. (2022). *The Misleading count: An identity-based intervention to mitigate the spread of partisan misinformation*. PsyArXiv. https://doi.org/10.31234/osf.io/7j26y

Priem, R. L., Harrison, D. A., & Muir, N. K. (1995). Structured Conflict and Consensus Outcomes in Group Decision Making. *Journal of Management*, *21*(4), 691–710. https://doi.org/10.1177/014920639502100406

Prior, M., Sood, G., Khanna, K., Prior, M., Sood, G., & Khanna, K. (2015). You cannot be serious: The impact of accuracy incentives on partisan bias in reports of economic perceptions. *Quart J Polit Sci*, *10*, 489–518.

Pröllochs, N. (2021). *Community-Based Fact-Checking on Twitter's Birdwatch Platform*. https://doi.org/10.48550/arXiv.2104.07175

Rambaran, J. A., Dijkstra, J. K., Munniksma, A., & Cillessen, A. H. N. (2015). The development of adolescents' friendships and antipathies: A longitudinal multivariate network test of balance theory. *Social Networks*, *43*, 162–176.

Reagans, R. (2013). Demographic diversity as network connections: Homophily and the diversity-performance debate. *The Oxford Handbook of Diversity and Work*, 192–206.

Reverdy, P., & Leonard, N. E. (2015). Parameter estimation in softmax decision-making models with linear objective functions. *IEEE Transactions on Automation Science and Engineering*, *13*(1), 54–67.

Reyes-García, V., Gallois, S., & Demps, K. (2016). A Multistage Learning Model for Cultural Transmission: Evidence from Three Indigenous Societies. In H. Terashima & B. S. Hewlett (Eds.), *Social Learning and Innovation in Contemporary Hunter-Gatherers: Evolutionary and Ethnographic Perspectives* (pp. 47–60). Springer Japan. https://doi.org/10.1007/978-4-431-55997-9_4

Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General*, *147*(11), 1553–1570. https://doi.org/10.1037/xge0000466

Roux, V., Bril, B., Cauliez, J., Goujon, A.-L., Lara, C., Manen, C., de Saulieu, G., & Zangato, E. (2017). Persisting technological boundaries: Social interactions, cognitive correlations and polarization. *Journal of Anthropological Archaeology*, *48*, 320–335. https://doi.org/10.1016/j.jaa.2017.09.004

Sakoda, J. M. (1971). The checkerboard model of social interaction. *The Journal of Mathematical Sociology*, *1*(1), 119–132. https://doi.org/10.1080/0022250X.1971.9989791

Sander, R. H., Kucheva, Y. A., & Zasloff, J. M. (2018). *Moving toward Integration: The Past and Future of Fair Housing*. Harvard University Press.

Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*, *1*(2), 143–186. https://doi.org/10.1080/0022250X.1971.9989794

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *PNAS*, *116*(16), 7662–7669. https://doi.org/10.1073/pnas.1805871115

Schimmelpfennig, R., Razek, L., Schnell, E., & Muthukrishna, M. (2022). Paradox of diversity in the collective brain. *Philosophical Transactions of the Royal Society B*. https://doi.org/10.1098/rstb.2020.0316

Schultner, D. T., Stillerman, B. S., Lindström, B. R., Hackel, L. M., Hagen, D. R., Jostmann, N. B., & Amodio, D. M. (2024). Transmission of societal stereotypes to individual-level prejudice through instrumental learning. *Proceedings of the National Academy of Sciences*, *121*(45), e2414518121. https://doi.org/10.1073/pnas.2414518121

Schulz-Hardt, S., & Mojzisch, A. (2012). How to achieve synergy in group decision making: Lessons to be learned from the hidden profile paradigm. *European Review of Social Psychology*, *23*(1), 305–343. https://doi.org/10.1080/10463283.2012.744440

Sherif, M., & Hovland, C. I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change* (pp. xii, 218). Yale University Press.

Shi, F., Teplitskiy, M., Duede, E., & Evans, J. A. (2019). The wisdom of polarized crowds. *Nature Human Behaviour*, *3*(4), Article 4. https://doi.org/10.1038/s41562-019-0541-6

Shore, J., Bernstein, E., & Lazer, D. (2015). Facts and Figuring: An Experimental Investigation of Network Structure and Performance in Information and Solution Spaces. *Organization Science*, *26*(5), 1432–1446. https://doi.org/10.1287/orsc.2015.0980

Smaldino, P. E., & Velilla, A. P. (2025). The evolution of similarity-biased social learning. *Evolutionary Human Sciences*, *7*, e4. https://doi.org/10.1017/ehs.2024.46

Snijder, L. L., Gross, J., Stallen, M., & De Dreu, C. K. W. (2024). Prosocial preferences can escalate intergroup conflicts by countering selfish motivations to leave. *Nature Communications*, *15*(1), 9009. https://doi.org/10.1038/s41467-024-53409-9

Sohrab, S. G., Waller, M. J., & Kaplan, S. (2015). Exploring the Hidden-Profile Paradigm: A Literature Review and Analysis. *Small Group Research*, *46*(5), 489–535. https://doi.org/10.1177/1046496415599068

Solomon, M. (2006). Norms of Epistemic Diversity. *Episteme*, *3*(1–2), 23–36. https://doi.org/10.3366/epi.2006.3.1-2.23

Sommers, S. R. (2006). On racial diversity and group decision making: Identifying multiple effects of racial composition on jury deliberations. *Journal of Personality and Social Psychology*, *90*(4), 597–612. https://doi.org/10.1037/0022-3514.90.4.597

Spreng, R. N., & Turner, G. R. (2021). From exploration to exploitation: A shifting mental mode in late life development. *Trends in Cognitive Sciences*, *25*(12), 1058–1071. https://doi.org/10.1016/j.tics.2021.09.001

Stark, T. H., & Flache, A. (2012). The Double Edge of Common Interest: Ethnic Segregation as an Unintended Byproduct of Opinion Homophily. *Sociology of Education*, *85*(2), 179–199. https://doi.org/10.1177/0038040711427314

Stasser, G., & Stewart, D. (1992). Discovery of hidden profiles by decision-making groups: Solving a problem versus making a judgment. *Journal of Personality and Social Psychology*, *63*, 426–434. https://doi.org/10.1037/0022-3514.63.3.426

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, *48*, 1467–1478. https://doi.org/10.1037/0022-3514.48.6.1467

Stasser, G., & Titus, W. (2003). Hidden Profiles: A Brief History. *Psychological Inquiry*, *14*(3–4), 304–313. https://doi.org/10.1080/1047840X.2003.9682897

Stein, J., Frey, V., & Flache, A. (2024). Talk Less to Strangers: How Homophily Can Improve Collective Decision-Making in Diverse Teams. *Journal of Artificial Societies and Social Simulation*, *27*(1), 14. https://doi.org/10.18564/jasss.5224

Stein, J., Frey, V., & van de Rijt, A. (2023). Realtime user ratings as a strategy for combatting misinformation: An experimental study. *Scientific Reports*, *13*(1), 1626.

Stein, J., Keuschnigg, M., & van de Rijt, A. (2023). Network segregation and the propagation of misinformation. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-022-26913-5

Stein, J., Romeijn, J.-W., & Mäs, M. (2025). Ill-informed Consensus or Truthful Disagreement? How Argumentation Styles and Preference Perceptions Affect Deliberation Outcomes in Groups with Conflicting Stakes. *Erkenntnis*. https://doi.org/10.1007/s10670-024-00913-5

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field. *Social Science Computer Review*, *38*(5), 503–516. https://doi.org/10.1177/0894439319843669

Stoica, V. I., & Flache, A. (2014). From Schelling to Schools: A Comparison of a Model of Residential Segregation with a Model of School Segregation. *Journal of Artificial Societies and Social Simulation*, *17*(1), 5.

Stovel, K., & Shaw, L. (2012). Brokerage. *Annual Review of Sociology*, *38*, 139–158. https://doi.org/10.1146/annurev-soc-081309-150054

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations* (1st ed). Doubleday. http://site.ebrary.com/id/10064804

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). *Some like it hoax: Automated fake news detection in social networks*. http://arxiv.org/abs/1704.07506

Tajfel, H. (1972). Social categorization, English manuscript of La catégorization sociale. In S. Moscovici (Ed.), *Introduction a la psychologie sociale* (Vol. 1, pp. 272–302). Larousse. https://cir.nii.ac.jp/crid/1573387450700299392

Tajfel, H. (2010). *Social Identity and Intergroup Relations*. Cambridge University Press.

Thomas, R. J. (2019). Sources of Friendship and Structurally Induced Homophily across the Life Course. *Sociological Perspectives*, *62*(6), 822–843. https://doi.org/10.1177/0731121419828399

Trepte, S., & Loy, L. S. (2017). Social Identity Theory and Self-Categorization Theory. In P. Rössler, C. A. Hoffner, & L. Zoonen (Eds.), *The International Encyclopedia of Media Effects* (1st ed., pp. 1–13). Wiley. https://doi.org/10.1002/9781118783764.wbieme0088

Tubergen, F. van. (2020). Chapter 3: Methods. In *Introduction to Sociology*. Routledge.

Van de Rijt, A. (2019). Self-correcting dynamics in social influence processes. *American Journal of Sociology*, *124*(5), 1468–1495.

van der Does, T., Galesic, M., Dunivin, Z. O., & Smaldino, P. E. (2022). Strategic identity signaling in heterogeneous networks. *Proceedings of the National Academy of Sciences*, *119*(10), e2117898119. https://doi.org/10.1073/pnas.2117898119

Van Dijk, H., Van Engen, M. L., & Van Knippenberg, D. (2012). Defying conventional wisdom: A meta-analytical examination of the differences between demographic and job-related diversity relationships with performance. *Organizational Behavior and Human Decision Processes*, *119*(1), 38–53.

van Veen, D.-J., Kudesia, R. S., & Heinimann, H. R. (2020). An Agent-Based Model of Collective Decision-Making: How Information Sharing Strategies Scale With Information Overload. *IEEE Transactions on Computational Social Systems*, *7*(3), 751–767. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2020.2986161

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

West, E. A., & Iyengar, S. (2022). Partisanship as a Social Identity: Implications for Polarization. *Political Behavior*, *44*(2), 807–838. https://doi.org/10.1007/s11109-020-09637-y

Westfall, J., Van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving Political Polarization in the United States: Party Identity Strength and Attitude Extremity Exacerbate the Perceived Partisan Divide. *Perspectives on Psychological Science*, *10*(2), 145–158. https://doi.org/10.1177/1745691615569849

Witt, A., Toyokawa, W., Lala, K. N., Gaissmaier, W., & Wu, C. M. (2024). Humans flexibly integrate social information despite interindividual differences in reward. *Proceedings of the National Academy of Sciences*, *121*(39), e2404928121. https://doi.org/10.1073/pnas.2404928121

Wittenbaum, G. M., Hollingshead, A. B., & Botero, I. C. (2004). From cooperative to motivated information sharing in groups: Moving beyond the hidden profile paradigm. *Communication Monographs*, *71*(3), 286–310. https://doi.org/10.1080/0363452042000299894

Wood, J. T. (2004). *Communication Theories in Action: An Introduction*. Wadsworth.

Wu, C. M., Deffner, D., Kahl, B., Meder, B., Ho, M. H., & Kurvers, R. H. J. M. (2024). *Visual-spatial dynamics drive adaptive social learning in immersive environments* (p. 2023.06.28.546887). bioRxiv. https://doi.org/10.1101/2023.06.28.546887

Wu, J., & O'Connor, C. (2021). *How Should We Promote Transient Diversity in Science?* [Preprint]. http://philsci-archive.pitt.edu/19428/

Zmigrod, L. (2022). A Psychology of Ideology: Unpacking the Psychological Structure of Ideological Thinking. *Perspectives on Psychological Science*, *17*(4), 1072–1092. https://doi.org/10.1177/17456916211044140

Zollman, K. J. (2010). The epistemic benefit of transient diversity. *Erkenntnis*, *72*(1), 17–35.

Zou, W., & Xu, X. (2023). Ingroup bias in a social learning experiment. *Experimental Economics*, *26*(1), 27–54. https://doi.org/10.1007/s10683-022-09788-1

Acknowledgements

For those who know me personally, it may be no surprise that writing this dissertation reminds me of a voyage on a sailing ship. Throughout the past five years, I crossed calm seas, unpredictable squalls, and stretches of dense fog. Fortunately, I was not alone on deck, and this journey would never have reached its harbor without the skill and dedication of the many people who sailed with me.

First, my deepest thanks to the senior officers I was allowed to learn from, my supervisors. Each of you brought your own essential skills to the bridge. Andreas, thank you for providing your advice and expertise whenever I needed it, for being a reliable source of support, and for your kindness. Michael, you provided excitement and vision. With you, what was looming beyond the horizon never seemed too far away. Jan-Willem, you introduced me to the novel waters of social epistemology and added an interdisciplinary perspective that expanded the map of my voyage. Vincenz, you took the time to examine things closely with me and were an indispensable sparring partner to refine ideas with. Because of your sharp eye for detail (I think it was the right one?), I could spot shoals ahead of time and safely circumnavigate them instead of running aground.

bridge: a section above deck housing a command center

Ships are accompanied by local navigators when they sail in unfamiliar waters. In my case, I was blessed to take on two outstanding pilots for a leg of the passage. Maxime, thank you for hosting me at IAST, introducing me to your lab group, and for opening my eyes to your research. Arnout – at this point more of a longstanding mentor than a passing acquaintance - thank you for your continuous support, and for being a huge source of inspiration to my work. Your enthusiasm and curiosity made me appreciate that research, just like sailing, is a dynamic process and not a linear one.

pilot: a specially knowledgeable person qualified to navigate a vessel through difficult waters

leg: a segment of a voyage between two waypoints

During my passage, I crossed paths with many other ships and their crews. Mentioning all of them lies beyond the possible for this section, but I would like to acknowledge a few of those scholars who shared a part of my journey with me, even if it was not strictly related to my dissertation: I thank Davide, Martina, and Shannon for conspiring with me. Marc, for pushing along all matters fake news. Jornt, for bringing family sociology into my network research. Wieteke, for our shared work on multiple jobholders. The reading committee, Leah Henderson, Scott E Page, and Heiko Rauhut I owe thanks for inspecting my dissertation from bowsprit to taffrail. And I would like to whistle a salute to my cursing, weather tanned and wiry boatswain: Casper 'Popeye' Kaandorp – your self-proclaimed title of code-god is well-deserved.

taffrail: a rail at the stern of a ship

boatswain: petty officer who 'pipes' commands to able seamen with a whistle

To the funding body SCOOP (and Rafael Wittek and Liesbet Heyse in particular), I owe gratitude for acting as the ship's pursers. You ensured that the hold was stocked and that the voyage could continue without fear of running short on stores. The SCOOPies, the ICS community and the Norms and Networks cluster represented the dockyards: Here, I could inspect my (incomplete) ship, have it checked by experts and collect ideas on how to make it seaworthy. Thank you for providing a platform to present my work in progress, and for dedicating countless hours to thinking along on how to improve it.

stores: supplies and equipment required for the navigation, operation, and upkeep of a ship

The department of sociology at the University of Groningen served as my home port – a safe and welcoming place where I could weather storms and gather fresh supplies. Special thanks go to Anita, the harbormaster, who worked ceaselessly to ensure that there was always a spot at the wharf for me to moor at.

harbormaster: person in charge of the operation of a harbor

moor: to dock a ship

My family provided the sailor's home where I could rest my sea legs while on shore leave. Frida, I am deeply appreciative of the many times you checked in about my state of affairs and made an effort to understand what I was up to. Christiane, thank you for providing me with a home I could return to, mentally and physically, whenever I needed to. Nothing is more comforting than a galley stocked with cheese after a long night's drive. Wolfgang, it was amazing to have an experienced captain like you to keep an eye on my journey,

galley: ship's kitchen

provide external advice and spin yarn with. And Maggy, you are the sweetest ship's dog one could imagine.

spin yarn: tell a tale or story

Perhaps, the biggest gratitude I owe to my fellow mariners in training, my closest PhD friends. I was blessed to be accompanied by a couple of midship(wo)men with a special fondness for lobsters: Elsje, in many ways you were like an anchor to me, steadfast, reliable, and fiercely loyal. Anna, you were my ship's rigging: standing tough and tall, mastering any storm that would be thrown upon you yet never expecting gratitude for it. Julian, like the wind, full of wits and surprises, sometimes mysterious but always indispensable to the passage. Sofie L, you were the ship's doctor, discreetly and empathetically lending your ear to those who needed it.[14]

midshipman: naval officer in training

I must not forget my other messmates, whom I am deeply grateful to have shared many jolly days at sea with: Amber, who practiced Dutch with me from the start and made me appreciate the North and its inhabitants like no one else could. Marie, with her admirable fight for mid's rights. Ramona, whose sharp insights continue to enlighten me. Sofie W, my kindred spirit in trekking adventures and Polish delicacies. Rowan, who gets The Lighthouse like no other. Stjoppa, with his curiosity and unmistakable way of thinking outside the box.

mess: a space where part of a crew eat and sometimes live together

That said, there are many other midship(wo)men who deserve mentioning: Dan, who was always happy to share his sailor's pipe when taking a break. Dennis, an overflowing cornucopia of unconventional projects. Alla, whose waters are silent but deeper than any sounding can reach. Fernanda, the real sailor of the two of us. Sören, lover of coffee, cats, and cinema. Danica, explorer and connoisseur of fine spirits. Klee, the most able top(hu)man in the rig. Hendrik, chief engineer in the netlogo department.

sounding: measuring the depth of the water

topman: a sailor working sails and rigging

doldrums: monotone, windless weather near the thermal equator

A sailor's life can be dull and dreary when the doldrums strike. Thankfully, they never lasted long because my shanty choir, Aquilo, would fill my sails again. Arjen, Ilse, Jasper, Mareike, Rose, Suzanne, and Veerle, I am unspeakably happy to have found such a fantastic vocal ensemble in you. Special

shanty: work song to accompany rhythmical labor

Aquilo: roman god of northern winds

[14] A part of a ship's doctor's responsibilities is to oversee pest control. To this end, it is not surprising that you also made sure to keep the (male) ship rats in check.

thanks go to Mareike and Veerle, whose knowledge of the local tongue proved invaluable for my Dutch summary.

With my Friesday friends, Lisa and Kimberley, I was excited to have shared many sticks of gold and numerous pleasant evenings. My Duitsers in Diaspora, Nico, Inga, Timo, Alex, Lena, and Stefan S kept Settlers of Catan close and homesickness at bay. To my marine biologists, Alenya, Anni, Maite, and Nino I am grateful to have been told much yarn about the fascinating creatures that lurk in the depths of the seas. Stefan P, thank you for the bike rides, bouldering sessions, and intriguing conversations. Robert, chief steward, my diet would be hard tack and scurvy upon me if it was not for your creativity and skill in the galley. Friederike, expert of Europe's Eastern lands and beacon of hope for Thuringia, thank you for our longstanding friendship. And I would like to thank Florian, my prized hiking buddy, passionate sociologist, and trusting confidante.

Finally, to my partner Ann: You are the lighthouse keeper who keeps the flame lit and the foghorn sounding. Your light guides me in the dark and helps me navigate uncertain waters. Thank you for your love, your companionship, and your faith in me. I could not imagine how to reach shore without you.

And with this, I will close the logbook of this passage. The equator is near, first dog watch begins, and the time for my crossing-the-line ceremony is nigh. May Neptun be with me.

steward: ship's cook

hard tack: dense cracker made from flour, water, and salt

scurvy: vitamin deficiency caused by lack of access to fresh produce

first dog watch: 4 - 6pm in the traditional maritime watch-keeping system

crossing-the-line ceremony: rite of passage, celebration of a sailor's first crossing of the equator

ICS dissertation series

The ICS series presents dissertations of the Interuniversity Center for Social Science Theory and Methodology. Each of these studies aims at integrating explicit theory formation with state-of-the-art empirical research or at the development of advanced methods for empirical research. The ICS was founded in 1986 as a cooperative effort of the universities of Groningen and Utrecht. Since 1992, the ICS has expanded to the University of Nijmegen and, since 2017, to the University of Amsterdam. Most of the projects are financed by the participating universities or by the Dutch Research Council. The international composition of the ICS graduate students is mirrored in the increasing international orientation of the projects and thus of the ICS series itself.

1. Kees van Liere (1990), *"Lastige leerlingen: Een empirisch onderzoek naar sociale oorzaken van probleemgedrag op basisscholen."* Amsterdam: Thesis Publishers
2. Marco van Leeuwen (1990), *"Bijstand in Amsterdam, ca. 1800-1850: Armenzorg als beheersings en overlevingsstrategie."* ICS dissertation, Utrecht
3. Ineke Maas (1990), *"Deelname aan podiumkunsten via de podia, de media en actieve beoefening: Substitutie of leereffecten?"* Amsterdam: Thesis Publishers
4. Marjolein Broese van Groenou (1991), *"Gescheiden netwerken: De relaties met vrienden en verwanten na echtscheiding."* Amsterdam: Thesis Publishers
5. Jan van den Bos (1991), *"Dutch EC policy making: A model guided approach to coordination and negotiation."* Amsterdam: Thesis Publishers
6. Karin Sanders (1991), *"Vrouwelijke pioniers: Vrouwen en mannen met een 'mannelijke' hogere beroepsopleiding aan het begin van hun loopbaan."* Amsterdam: Thesis Publishers
7. Sjerp de Vries (1991), *"Egoism, altruism, and social justice: Theory and experiments on cooperation in social dilemmas."* Amsterdam: Thesis Publishers
8. Ronald Batenburg (1991), *"Automatisering in bedrijf."* Amsterdam: Thesis Publishers
9. Rudi Wielers (1991), *"Selectie en allocatie op de arbeidsmarkt. Een uitwerking voor de informele en geïnstitutionaliseerde kinderopvang."* Amsterdam: Thesis Publishers
10. Gert Westert (1991), *"Verschillen in ziekenhuisgebruik."* ICS dissertation, Groningen
11. Hanneke Hermsen (1992), *"Votes and policy preferences: Equilibria in party systems."* Amsterdam: Thesis Publishers

12. Cora Maas (1992), *"Probleemleerlingen in het basisonderwijs."* Amsterdam: Thesis Publishers
13. Ed Boxman (1992), *"Contacten en carrière : Een empirisch theoretisch onderzoek naar de relatie tussen sociale netwerken en arbeidsmarktposities"* Amsterdam: Thesis Publishers
14. Conny Taes (1992), *"Kijken naar banen: Een onderzoek naar de inschatting van arbeidsmarktkansen bij schoolverlaters uit het middelbaar beroepsonderwijs."* Amsterdam: Thesis Publishers
15. Peter van Roozendaal (1992), *"Cabinets in multi party democracies: The effect of dominant and central parties on cabinet composition and durability."* Amsterdam: Thesis Publishers
16. Marcel van Dam (1992), *"Regio zonder regie: Verschillen in en effectiviteit van gemeentelijk arbeidsmarktbeleid."* Amsterdam: Thesis Publishers
17. Tanja van der Lippe (1993), *"Arbeidsverdeling tussen mannen en vrouwen."* Amsterdam: Thesis Publishers
18. Marc Jacobs (1993), *"Software: Kopen of kopiëren? Een sociaal wetenschappelijk onderzoek onder PC gebruikers."* Amsterdam: Thesis Publishers
19. Peter van der Meer (1993), *"Verdringing op de Nederlandse arbeidsmarkt: Sector- en sekseverschillen."* Amsterdam: Thesis Publishers
20. Gerbert Kraaykamp (1993), *"Over lezen gesproken: Een studie naar sociale differentiatie in leesgedrag."* Amsterdam: Thesis Publishers
21. Evelien Zeggelink (1993), *"Strangers into friends: The evolution of friendship networks using an individual oriented modeling approach."* Amsterdam: Thesis Publishers
22. Jaco Berveling (1994), *"Het stempel op de besluitvorming: Macht, invloed en besluitvorming op twee Amsterdamse beleidsterreinen."* Amsterdam: Thesis Publishers
23. Wim Bernasco (1994), *"Coupled careers: The effects of spouse's resources on success at work."* Amsterdam: Thesis Publishers
24. Liset van Dijk (1994), *"Choices in child care: The distribution of child care among mothers, fathers and non parental care providers."* Amsterdam: Thesis Publishers
25. Jos de Haan (1994), *"Research groups in Dutch sociology."* Amsterdam: Thesis Publishers
26. Kwasi Boahene (1995), *"Innovation adoption as a socioeconomic process: The case of the Ghanaian cocoa industry."* Amsterdam: Thesis Publishers
27. Paul Ligthart (1995), *"Solidarity in economic transactions: An experimental study of framing effects in bargaining and contracting."* Amsterdam: Thesis Publishers
28. Roger Leenders (1995), *"Structure and influence: Statistical models for the dynamics of actor attributes, network structure, and their interdependence."* Amsterdam: Thesis Publishers
29. Beate Volker (1995), *"Should auld acquaintance be forgot...? Institutions of communism, the transition to capitalism and personal networks: The case of East Germany."* Amsterdam: Thesis Publishers

30. Anneke Cancrinus-Matthijsse (1995), *"Tussen hulpverlening en ondernemerschap: Beroepsuitoefening en taakopvattingen van openbare apothekers in een aantal West-Europese landen."* Amsterdam: Thesis Publishers
31. Nardi Steverink (1996), *"Zo lang mogelijk zelfstandig: Naar een verklaring van verschillen in oriëntatie ten aanzien van opname in een verzorgingstehuis onder fysiek kwetsbare ouderen."* Amsterdam: Thesis Publishers
32. Ellen Lindeman (1996), *"Participatie in vrijwilligerswerk."* Amsterdam: Thesis Publishers
33. Chris Snijders (1996), *"Trust and commitments."* Amsterdam: Thesis Publishers
34. Koos Postma (1996), *"Changing prejudice in Hungary. A study on the collapse of state socialism and its impact on prejudice against Gypsies and Jews."* Amsterdam: Thesis Publishers
35. Jooske van Busschbach (1996), *"Uit het oog, uit het hart? Stabiliteit en verandering in persoonlijke relaties."* Amsterdam: Thesis Publishers
36. René Torenvlied (1996), *"Besluiten in uitvoering: Theorieën over beleidsuitvoering modelmatig getoetst op sociale vernieuwing in drie gemeenten."* Amsterdam: Thesis Publishers
37. Andreas Flache (1996), *"The double edge of networks: An analysis of the effect of informal networks on cooperation in social dilemmas."* Amsterdam: Thesis Publishers
38. Kees van Veen (1997), *"Inside an internal labor market: Formal rules, flexibility and career lines in a Dutch manufacturing company."* Amsterdam: Thesis Publishers
39. Lucienne van Eijk (1997), *"Activity and wellbeing in the elderly."* Amsterdam: Thesis Publishers
40. Róbert Gál (1997), *"Unreliability: Contract discipline and contract governance under economic transition."* Amsterdam: Thesis Publishers
41. Anne-Geerte van de Goor (1997), *"Effects of regulation on disability duration."* ICS dissertation, Utrecht
42. Boris Blumberg (1997), *"Das Management von Technologiekooperationen: Partnersuche und Verhandlungen mit dem Partner aus Empirisch Theoretischer Perspektive."* ICS dissertation, Utrecht
43. Marijke von Bergh (1997), *"Loopbanen van oudere werknemers."* Amsterdam: Thesis Publishers
44. Anna Petra Nieboer (1997), *"Life events and well-being: A prospective study on changes in well-being of elderly people due to a serious illness event or death of the spouse."* Amsterdam: Thesis Publishers
45. Jacques Niehof (1997), *"Resources and social reproduction: The effects of cultural and material resources on educational and occupational careers in industrial nations at the end of the twentieth century."* ICS dissertation, Nijmegen
46. Ariana Need (1997), *"The kindred vote: Individual and family effects of social class and religion on electoral change in the Netherlands, 1956 1994."* ICS dissertation, Nijmegen
47. Jim Allen (1997), *"Sector composition and the effect of education on Wages: An international comparison."* Amsterdam: Thesis Publishers

48. Jack Hutten (1998), *"Workload and provision of care in general practice: An empirical study of the relation between workload of Dutch general practitioners and the content and quality of their care."* ICS dissertation, Utrecht
49. Per Kropp (1998), *"Berufserfolg im Transformationsprozeß: Eine theoretisch empirische Studie über die Gewinner und Verlierer der Wende in Ostdeutschland."* ICS dissertation, Utrecht
50. Maarten Wolbers (1998), *"Diploma-inflatie en verdringing op de arbeidsmarkt: Een studie naar ontwikkelingen in de opbrengsten van diploma's in Nederland."* ICS dissertation, Nijmegen
51. Wilma Smeenk (1998), *"Opportunity and marriage: The impact of individual resources and marriage market structure on first marriage timing and partner choice in the Netherlands."* ICS dissertation, Nijmegen
52. Marinus Spreen (1999), *"Sampling personal network structures: Statistical inference in ego graphs."* ICS dissertation, Groningen
53. Vincent Buskens (1999), *"Social networks and trust."* ICS dissertation, Utrecht
54. Susanne Rijken (1999), *"Educational expansion and status attainment: A cross-national and over-time comparison."* ICS dissertation, Utrecht
55. Mérove Gijsberts (1999), *"The legitimation of inequality in state-socialist and market societies, 1987-1996."* ICS dissertation, Utrecht
56. Gerhard van de Bunt (1999), *"Friends by choice: An actor-oriented statistical network model for friendship networks through time."* ICS dissertation, Groningen
57. Robert Thomson (1999), *"The party mandate: Election pledges and government actions in the Netherlands, 1986 1998."* Amsterdam: Thela Thesis
58. Corine Baarda (1999), *"Politieke besluiten en boeren beslissingen: Het draagvlak van het mestbeleid tot 2000."* ICS dissertation, Groningen
59. Rafael Wittek (1999), *"Interdependence and informal control in organizations."* ICS dissertation, Groningen
60. Diane Payne (1999), *"Policy Making in the European Union: An analysis of the impact of the reform of the structural funds in Ireland."* ICS dissertation, Groningen
61. René Veenstra (1999), *"Leerlingen, Klassen, Scholen: Prestaties en vorderingen van leerlingen in het voortgezet onderwijs."* Amsterdam, Thela Thesis
62. Marjolein Achterkamp (1999), *"Influence strategies in collective decision making: A comparison of two models."* ICS dissertation, Groningen
63. Peter Mühlau (2000), *"The governance of the employment relation: Arelational signaling perspective."* ICS dissertation, Groningen
64. Agnes Akkerman (2000), *"Verdeelde vakbeweging en stakingen: Concurrentie om leden."* ICS dissertation, Groningen
65. Sandra van Thiel (2000), *"Quangocratization: Trends, causes and consequences."* ICS dissertation, Utrecht
66. Sylvia Peacock-Korupp (2000), *"Mothers and the process of social stratification."* ICS dissertation, Utrecht
67. Rudi Turksema (2000), *"Supply of day care."* ICS dissertation, Utrecht
68. Bernard Nijstad (2000), *"How the group affects the mind: Effects of communication in idea generating groups."* ICS dissertation, Utrecht

69. Inge de Wolf (2000), *"Opleidingsspecialisatie en arbeidsmarktsucces van sociale wetenschappers."* ICS dissertation, Utrecht
70. Jan Kratzer (2001), *"Communication and performance: An empirical study in innovation teams."* ICS dissertation, Groningen
71. Madelon Kroneman (2001), *"Healthcare systems and hospital bed use."* ICS/NIVEL-dissertation, Utrecht
72. Herman van de Werfhorst (2001), *"Field of study and social inequality: Four types of educational resources in the process of stratification in the Netherlands."* ICS dissertation, Nijmegen
73. Tamás Bartus (2001), *"Social capital and earnings inequalities: The role of informal job search in Hungary."* ICS dissertation Groningen
74. Hester Moerbeek (2001), *"Friends and foes in the occupational career: The influence of sweet and sour social capital on the labour market."* ICS dissertation, Nijmegen
75. Marcel van Assen (2001), *"Essays on actor perspectives in exchange networks and social dilemmas."* ICS dissertation, Groningen
76. Inge Sieben (2001), *"Sibling similarities and social stratification: The impact of family background across countries and cohorts."* ICS dissertation, Nijmegen
77. Alinda van Bruggen (2001), *"Individual production of social well-being: An exploratory study."* ICS dissertation, Groningen
78. Marcel Coenders (2001), *"Nationalistic attitudes and ethnic exclusionism in a comparative perspective: An empirical study of attitudes toward the country and ethnic immigrants in 22 countries."* ICS dissertation, Nijmegen
79. Marcel Lubbers (2001), *"Exclusionistic electorates: Extreme right-wing voting in Western Europe."* ICS dissertation, Nijmegen
80. Uwe Matzat (2001), *"Social networks and cooperation in electronic communities: A theoretical-empirical analysis of academic communication and internet discussion groups."* ICS dissertation, Groningen
81. Jacques Janssen (2002), *"Do opposites attract divorce? Dimensions of mixed marriage and the risk of divorce in the Netherlands."* ICS dissertation, Nijmegen
82. Miranda Jansen (2002), *"Waardenoriëntaties en partnerrelaties: Een panelstudie naar wederzijdse invloeden."* ICS dissertation, Utrecht
83. Anne-Rigt Poortman (2002), *"Socioeconomic causes and consequences of divorce."* ICS dissertation, Utrecht
84. Alexander Gattig (2002), *"Intertemporal decision making."* ICS dissertation, Groningen
85. Gerrit Rooks (2000), *"Contract en conflict: Strategisch management van inkooptransacties."* ICS dissertation, Utrecht
86. Károly Takács (2002), *"Social networks and intergroup conflict."* ICS dissertation, Groningen
87. Thomas Gautschi (2002), *"Trust and exchange, effects of temporal embeddedness and network embeddedness on providing and dividing a surplus."* ICS dissertation, Utrecht
88. Hilde Bras (2002), *"Zeeuwse meiden: Dienen in de levensloop van vrouwen, ca. 1850 – 1950."* Aksant Academic Publishers, Amsterdam

89. Merijn Rengers (2002), *"Economic lives of artists: Studies into careers and the labour market in the cultural sector."* ICS dissertation, Utrecht
90. Annelies Kassenberg (2002), *"Wat scholieren bindt: Sociale gemeenschap in scholen."* ICS dissertation, Groningen
91. Marc Verboord (2003), *"Moet de meester dalen of de leerling klimmen? De invloed van literatuuronderwijs en ouders op het lezen van boeken tussen 1975 en 2000."* ICS dissertation, Utrecht
92. Marcel van Egmond (2003), *"Rain falls on all of us (but some manage to get more wet than others): Political context and electoral participation."* ICS dissertation, Nijmegen
93. Justine Horgan (2003), *"High-performance human resource management in Ireland and the Netherlands: Adoption and effectiveness."* ICS dissertation, Groningen
94. Corine Hoeben (2003), *"LETS' be a community: Community in local exchange trading systems."* ICS dissertation, Groningen
95. Christian Steglich (2003), *"The framing of decision situations: Automatic goal selection and rational goal pursuit."* ICS dissertation, Groningen
96. Johan van Wilsem (2003), *"Crime and context: The impact of individual, neighborhood, city and country characteristics on victimization."* ICS dissertation, Nijmegen
97. Christiaan Monden (2003), *"Education, inequality and health: The impact of partners and life course."* ICS dissertation, Nijmegen
98. Evelyn Hello (2003), *"Educational attainment and ethnic attitudes: How to explain their relationship."* ICS dissertation, Nijmegen
99. Marnix Croes en Peter Tammes (2004). *"Gif laten wij niet voortbestaan: Een onderzoek naar de overlevingskansen van Joden in de Nederlandse gemeenten, 1940-1945."* Aksant Academic Publishers, Amsterdam.
100. Ineke Nagel (2004), *"Cultuurdeelname in de levensloop."* ICS dissertation, Utrecht
101. Marieke van der Wal (2004), *"Competencies to participate in life: Measurement and the impact of school."* ICS dissertation, Groningen
102. Vivian Meertens (2004), *"Depressive symptoms in the general population: A multifactorial social approach."* ICS dissertation, Nijmegen
103. Hanneke Schuurmans (2004), *"Promoting well-being in frail elderly people: Theory and intervention."* ICS dissertation, Groningen
104. Javier Arregui Moreno (2004), *"Negotiation in legislative decision-making in the European Union."* ICS dissertation, Groningen
105. Tamar Fischer (2004), *"Parental divorce, conflict and resources: The effects on children's behaviour problems, socioeconomic attainment, and transitions in the demographic career."* ICS dissertation, Nijmegen
106. René Bekkers (2004), *"Giving and volunteering in the Netherlands: Sociological and psychological perspectives."* ICS dissertation, Utrecht
107. Renée van der Hulst (2004), *"Gender differences in workplace authority: An empirical study on social networks."* ICS dissertation, Groningen

108. Rita Smaniotto (2004), *"'You scratch my back and I scratch yours' versus 'love thy neighbour': Two proximate mechanisms of reciprocal altruism."* ICS dissertation, Groningen
109. Maurice Gesthuizen (2004), *"The life course of the low-educated in the Netherlands: Social and economic risks."* ICS dissertation, Nijmegen
110. Carlijne Philips (2005), *"Vakantiegemeenschappen: Kwalitatief en kwantitatief onderzoek naar gelegenheid en refreshergemeenschap tijdens de vakantie."* ICS dissertation, Groningen
111. Esther de Ruijter (2005), *"Household outsourcing."* ICS dissertation, Utrecht
112. Frank van Tubergen (2005), *"The integration of immigrants in cross-national perspective: Origin, destination, and community effects."* ICS dissertation, Utrecht
113. Ferry Koster (2005), *"For the time being: Accounting for inconclusive findings concerning the effects of temporary employment relationships on solidary behavior of employees."* ICS dissertation, Groningen
114. Carolien Klein Haarhuis (2005), *"Promoting anti-corruption reforms: Evaluating the implementation of a World Bank anti-corruption program in seven African countries (1999-2001)."* ICS dissertation, Utrecht
115. Martin van der Gaag (2005), *"Measurement of individual social capital."* ICS dissertation, Groningen
116. Johan Hansen (2005), *"Shaping careers of men and women in organizational contexts."* ICS dissertation, Utrecht
117. Davide Barrera (2005), *"Trust in embedded settings."* ICS dissertation, Utrecht
118. Mattijs Lambooij (2005), *"Promoting cooperation: Studies into the effects of long-term and short-term rewards on cooperation of employees."* ICS dissertation, Utrecht
119. Lotte Vermeij (2006), *"What's cooking? Cultural boundaries among Dutch teenagers of different ethnic origins in the context of school."* ICS dissertation, Utrecht
120. Mathilde Strating (2006), *"Facing the challenge of rheumatoid arthritis: A 13-year prospective study among patients and cross-sectional study among their partners."* ICS dissertation, Groningen
121. Jannes de Vries (2006), *"Measurement error in family background variables: The bias in the intergenerational transmission of status, cultural consumption, party preference, and religiosity."* ICS dissertation, Nijmegen
122. Stefan Thau (2006), *"Workplace deviance: Four studies on employee motives and self-regulation."* ICS dissertation, Groningen
123. Mirjam Plantinga (2006), *"Employee motivation and employee performance in child care: The effects of the introduction of market forces on employees in the Dutch child-care sector."* ICS dissertation, Groningen
124. Helga de Valk (2006), *"Pathways into adulthood: A comparative study on family life transitions among migrant and Dutch youth."* ICS dissertation, Utrecht
125. Henrike Elzen (2006), *"Self-Management for chronically ill older people."* ICS dissertation, Groningen
126. Ayse Güveli (2007), *"New social classes within the service class in the Netherlands and Britain: Adjusting the EGP class schema for the technocrats and the social and cultural specialists."* ICS dissertation, Nijmegen

127. Willem-Jan Verhoeven (2007), *"Income attainment in post-communist societies."* ICS dissertation, Utrecht
128. Marieke Voorpostel (2007), *"Sibling support: The exchange of help among brothers and sisters in the Netherlands."* ICS dissertation, Utrecht
129. Jacob Dijkstra (2007), *"The effects of externalities on partner choice and payoffs in exchange networks."* ICS dissertation, Groningen
130. Patricia van Echtelt (2007), *"Time-greedy employment relationships: Four studies on the time claims of post-Fordist work."* ICS dissertation, Groningen
131. Sonja Vogt (2007), *"Heterogeneity in social dilemmas: The case of social support."* ICS dissertation, Utrecht
132. Michael Schweinberger (2007), *"Statistical methods for studying the evolution of networks and behavior."* ICS dissertation, Groningen
133. István Henrik Back (2007), *"Commitment and evolution: Connecting emotion and reason in long-term relationships."* ICS dissertation, Groningen
134. Ruben van Gaalen (2007), *"Solidarity and ambivalence in parent-child relationships."* ICS dissertation, Utrecht
135. Jan Reitsma (2007), *"Religiosity and solidarity: Dsimensions and relationships disentangled and tested."* ICS dissertation, Nijmegen
136. Jan Kornelis Dijkstra (2007), *"Status and affection among (pre)adolescents and their relation with antisocial and prosocial behavior."* ICS dissertation, Groningen
137. Wouter van Gils (2007), *"Full-time working couples in the Netherlands: Causes and consequences."* ICS dissertation, Nijmegen
138. Djamila Schans (2007), *"Ethnic diversity in intergenerational solidarity."* ICS dissertation, Utrecht
139. Ruud van der Meulen (2007), *"Brug over woelig water: Lidmaatschap van sportverenigingen, vriendschappen, kennissenkringen en veralgemeend vertrouwen."* ICS dissertation, Nijmegen
140. Andrea Knecht (2008), *"Friendship selection and friends' influence: Dynamics of networks and actor attributes in early adolescence."* ICS dissertation, Utrecht
141. Ingrid Doorten (2008), *"The division of unpaid work in the household: A stubborn pattern?"* ICS dissertation, Utrecht
142. Stijn Ruiter (2008), *"Association in context and association as context: Causes and consequences of voluntary association involvement."* ICS dissertation, Nijmegen
143. Janneke Joly (2008), *"People on our minds: When humanized contexts activate social norms."* ICS dissertation, Groningen
144. Margreet Frieling (2008), *"'Joint production' als motor voor actief burgerschap in de buurt."* ICS dissertation, Groningen
145. Ellen Verbakel (2008), *"The partner as resource or restriction? Labour market careers of husbands and wives and the consequences for inequality between couples."* ICS dissertation, Nijmegen
146. Gijs van Houten (2008), *"Beleidsuitvoering in gelaagde stelsels: De doorwerking van aanbevelingen van de Stichting van de Arbeid in het CAO-overleg."* ICS dissertation, Utrecht

147. Eva Jaspers (2008), *"Intolerance over time: Macro- and micro-level questions on attitudes towards euthanasia, homosexuality and ethnic minorities."* ICS dissertation, Nijmegen
148. Gijs Weijters (2008), *"Youth delinquency in Dutch cities and schools: A multilevel approach."* ICS dissertation, Nijmegen
149. Jessica Nooij (2009), *"The self in social rejection."* ICS dissertation, Groningen
150. Gerald Mollenhorst (2009), *"Networks in contexts: How meeting opportunities affect personal relationships."* ICS dissertation, Utrecht
151. Tom van der Meer (2009), *"States of freely associating citizens: Comparative studies into the impact of state institutions on social, civic and political participation."* ICS dissertation, Nijmegen
152. Manuela Vieth (2009), *"Commitments and reciprocity in trust situations: Experimental studies on obligation, indignation, and self-consistency."* ICS dissertation, Utrecht
153. Rense Corten (2009), *"Co-evolution of social networks and behavior in social dilemmas: Theoretical and empirical perspectives."* ICS dissertation, Utrecht
154. Arieke Rijken (2009), *"Happy families, high fertility? Childbearing choices in the context of family and partner relationships."* ICS dissertation, Utrecht
155. Jochem Tolsma (2009), *"Ethnic hostility among ethnic majority and minority groups in the Netherlands: An investigation into the impact of social mobility experiences, the local living environment and educational attainment on ethnic hostility."* ICS dissertation, Nijmegen
156. Freek Bucx (2009), *"Linked lives: Young adults' life course and relations with parents."* ICS dissertation, Utrecht
157. Philip Wotschack (2009), *"Household governance and time allocation: Four studies on the combination of work and care."* ICS dissertation, Groningen
158. Nienke Moor (2009), *"Explaining worldwide religious diversity: The relationship between subsistence technologies and ideas about the unknown in pre-industrial and (post-)industrial societies."* ICS dissertation, Nijmegen
159. Lieke ten Brummelhuis (2009), *"Family matters at work: Depleting and enriching effects of employees' family lives on work outcomes."* ICS dissertation, Utrecht
160. Renske Keizer (2010), *"Remaining childless: Causes and consequences from a life-course perspective."* ICS dissertation, Utrecht
161. Miranda Sentse (2010), *"Bridging contexts: The interplay between family, child, and peers in explaining problem behavior in early adolescence."* ICS dissertation, Groningen
162. Nicole Tieben (2010), *"Transitions, tracks and transformations: Social inequality in transitions into, through and out of secondary education in the Netherlands for cohorts born between 1914 and 1985."* ICS dissertation, Nijmegen
163. Birgit Pauksztat (2010), *"Speaking up in organizations: Four studies on employee voice."* ICS dissertation, Groningen
164. Richard Zijdeman (2010), *"Status attainment in the Netherlands, 1811-1941: Spatial and temporal variation before and during industrialization."* ICS dissertation, Utrecht

165. Rianne Kloosterman (2010), *"Social background and children's educational careers: The primary and secondary effects of social background over transitions and over time in the Netherlands."* ICS dissertation, Nijmegen
166. Olav Aarts (2010), *"Religious diversity and religious involvement: A study of religious markets in Western societies at the end of the twentieth century."* ICS dissertation, Nijmegen
167. Stephanie Wiesmann (2010), *"24/7 Negotiation in couples' transition to parenthood."* ICS dissertation, Utrecht
168. Borja Martinovic (2010), *"Interethnic contacts: A dynamic analysis of interaction between immigrants and natives in Western countries."* ICS dissertation, Utrecht
169. Anne Roeters (2010), *"Family life under pressure? Parents' paid work and the quantity and quality of parent-child and family time."* ICS dissertation, Utrecht
170. Jelle Sijtsema (2010), *"Adolescent aggressive behavior: Status and stimulation goals in relation to the peer context."* ICS dissertation, Groningen
171. Kees Keizer (2010), *"The spreading of disorder."* ICS dissertation, Groningen
172. Michael Mäs (2010), *"The diversity puzzle. explaining clustering and polarization of opinions."* ICS dissertation, Groningen
173. Marie-Louise Damen (2010), *"Cultuurdeelname en CKV: Studies naar effecten van kunsteducatie op de cultuurdeelname van leerlingen tijdens en na het voortgezet onderwijs."* ICS dissertation, Utrecht
174. Marieke van de Rakt (2011), *"Two generations of crime: The intergenerational transmission of convictions over the life course."* ICS dissertation, Nijmegen
175. Willem Huijnk (2011), *"Family life and ethnic attitudes: The role of the family for attitudes towards intermarriage and acculturation among minority and majority groups."* ICS dissertation, Utrecht
176. Tim Huijts (2011), *"Social ties and health in Europe: Individual associations, cross-national variations, and contextual explanations."* ICS dissertation, Nijmegen
177. Wouter Steenbeek (2011), *"Social and physical disorder: How community, business presence and entrepreneurs influence disorder in Dutch neighborhoods."* ICS dissertation, Utrecht
178. Miranda Vervoort (2011), *"Living together apart? Ethnic concentration in the neighborhood and ethnic minorities' social contacts and language practices."* ICS dissertation, Utrecht
179. Agnieszka Kanas (2011), *"The economic performance of immigrants: The role of human and social capital."* ICS dissertation, Utrecht
180. Lea Ellwardt (2011), *"Gossip in organizations: A social network study."* ICS dissertation, Groningen
181. Annemarije Oosterwaal (2011), *"The gap between decision and implementation: Decision making, delegation and compliance in governmental and organizational settings."* ICS dissertation, Utrecht
182. Natascha Notten (2011), *"Parents and the media: Causes and consequences of parental media socialization."* ICS dissertation, Nijmegen
183. Tobias Stark (2011), *"Integration in schools: A process perspective on students' interethnic attitudes and interpersonal relationships."* ICS dissertation, Groningen

184. Giedo Jansen (2011), *"Social cleavages and political choices: Large-scale comparisons of social class, religion and voting behavior in Western democracies."* ICS dissertation, Nijmegen
185. Ruud van der Horst (2011), *"Network effects on treatment results in a closed forensic psychiatric setting."* ICS dissertation, Groningen
186. Mark Levels (2011), *"Abortion laws in European Countries between 1960 and 2010: Legislative developments and their consequences for women's reproductive decision making."* ICS dissertation, Nijmegen
187. Marieke van Londen (2012), *"Exclusion of ethnic minorities in the Netherlands: The effects of individual and situational characteristics on opposition to ethnic policy and ethnically mixed neighbourhoods."* ICS dissertation, Nijmegen
188. Sigrid Mohnen (2012), *"Neighborhood context and health: How neighborhood social capital affects individual health."* ICS dissertation, Utrecht
189. Asya Zhelyazkova (2012), *"Compliance under controversy: Analysis of the transposition of European directives and their provisions."* ICS dissertation, Utrecht
190. Valeska Korff (2012), *"Between cause and control: Management in a humanitarian organization."* ICS dissertation, Groningen
191. Maike Gieling (2012), *"Dealing with diversity: Adolescents' support for civil liberties and immigrant rights."* ICS dissertation, Utrecht
192. Katya Ivanova (2012), *"From parents to partners: The impact of family on romantic relationships in adolescence and emerging adulthood."* ICS dissertation, Groningen
193. Jelmer Schalk (2012), *"The performance of public corporate actors: Essays on effects of institutional and network embeddedness in supranational, national, and local collaborative contexts."* ICS dissertation, Utrecht
194. Alona Labun (2012), *"Social networks and informal power in organizations."* ICS dissertation, Groningen
195. Michal Bojanowski (2012), *"Essays on social network formation in heterogeneous populations: Models, methods, and empirical analyses."* ICS dissertation, Utrecht
196. Anca Minescu (2012), *"Relative group position and intergroup attitudes in Russia."* ICS dissertation, Utrecht
197. Marieke van Schellen (2012), *"Marriage and crime over the life course: The criminal careers of convicts and their spouses."* ICS dissertation, Utrecht
198. Mieke Maliepaard (2012), *"Religious trends and social integration: Muslim minorities in the Netherlands."* ICS dissertation, Utrecht
199. Fransje Smits (2012), *"Turks and Moroccans in the Low Countries around the year 2000: Determinants of religiosity, trend in religiosity and determinants of the trend."* ICS dissertation, Nijmegen
200. Roderick Sluiter (2012), *"The diffusion of morality policies among Western European countries between 1960 and 2010: A comparison of temporal and spatial diffusion patterns of six morality and eleven non-morality policies."* ICS dissertation, Nijmegen
201. Nicoletta Balbo (2012), *"Family, friends and fertility."* ICS dissertation, Groningen
202. Anke Munniksma (2013), *"Crossing ethnic boundaries: Parental resistance to and consequences of adolescents' cross-ethnic peer relations"* ICS dissertation, Groningen

203. Anja-Kristin Abendroth (2013), *"Working women in Europe: How the country, workplace, and family context matter."* ICS dissertation, Utrecht
204. Katia Begall (2013), *"Occupational hazard? The relationship between working conditions and fertility."* ICS dissertation, Groningen
205. Hidde Bekhuis (2013), *"The popularity of domestic cultural products: Cross-national differences and the relation to globalization."* ICS dissertation, Utrecht
206. Lieselotte Blommaert (2013), *"Are Joris and Renske more employable than Rashid and Samira? A study on the prevalence and sources of ethnic discrimination in recruitment in the Netherlands using experimental and survey data."* ICS dissertation, Utrecht
207. Wiebke Schulz (2013), *"Careers of men and women in the 19th and 20th centuries."* ICS dissertation, Utrecht
208. Ozan Aksoy (2013), *"Essays on social preferences and beliefs in non-embedded social dilemmas."* ICS dissertation, Utrecht
209. Dominik Morbitzer (2013), *"Limited farsightedness in network formation."* ICS dissertation, Utrecht
210. Thomas de Vroome (2013), *"Earning your place: The relation between immigrants' economic and psychological integration in the Netherlands."* ICS dissertation, Utrecht
211. Marloes de Lange (2013), *"Causes and consequences of employment flexibility among young people: Recent developments in the Netherlands and Europe."* ICS dissertation, Nijmegen
212. Roza Meuleman (2014), *"Consuming the Nation: Domestic cultural consumption: Its stratification and relation with nationalist attitudes."* ICS dissertation, Utrecht
213. Esther Havekes (2014), *"Putting interethnic attitudes in context: The relationship between neighbourhood characteristics, interethnic attitudes and residential behaviour."* ICS dissertation, Utrecht
214. Zoltán Lippényi (2014), *"Transitions toward an open society? Intergenerational occupational mobility in Hungary in the 19th and 20th centuries."* ICS dissertation, Utrecht
215. Anouk Smeekes (2014), *"The presence of the past: Historical rooting of national identity and current group dynamics."* ICS dissertation, Utrecht
216. Michael Savelkoul (2014), *"Ethnic diversity and social capital: Testing underlying explanations derived from conflict and contact theories in Europe and the United States."* ICS dissertation, Nijmegen
217. Martijn Hogerbrugge (2014), *"Misfortune and family: How negative events, family ties, and lives are linked."* ICS dissertation, Utrecht
218. Gina-Felicia Potarca (2014), *"Modern love: Comparative insights in online dating preferences and assortative mating."* ICS dissertation, Groningen
219. Mariska van der Horst (2014), *"Gender, aspirations, and achievements: Relating work and family aspirations to occupational outcomes."* ICS dissertation, Utrecht
220. Gijs Huitsing (2014), *"A social network perspective on bullying"* ICS dissertation, Groningen
221. Thomas Kowalewski (2015), *"Personal growth in organizational contexts."* ICS dissertation, Groningen

222. Manuel Muñoz-Herrera (2015), *"The impact of individual differences on network relations: Social exclusion and inequality in productive exchange and coordination games."* ICS dissertation, Groningen
223. Tim Immerzeel (2015), *"Voting for a change: The democratic lure of populist radical right parties in voting behavior."* ICS dissertation, Utrecht
224. Fernando Nieto Morales (2015), *"The control imperative: Studies on reorganization in the public and private sectors."* ICS dissertation, Groningen
225. Jellie Sierksma (2015), *"Bounded helping: How morality and intergroup relations shape children's reasoning about helping."* ICS dissertation, Utrecht
226. Tinka Veldhuis (2015), *"Captivated by fear: An evaluation of terrorism detention policy."* ICS dissertation, Groningen
227. Miranda Visser (2015), *"Loyality in humanity: Turnover among expatriate humanitarian aid workers."* ICS dissertation, Groningen
228. Sarah Westphal (2015), *"Are the kids alright? Essays on postdivorce residence arrangements and children's well-being."* ICS dissertation, Utrecht
229. Britta Rüschoff (2015), *"Peers in careers: Peer relationships in the transition from school to work."* ICS dissertation, Groningen
230. Nynke van Miltenburg (2015), *"Cooperation under peer sanctioning institutions: Collective decisions, noise, and endogenous implementation."* ICS dissertation, Utrecht
231. Antonie Knigge (2015), *"Sources of sibling similarity: Status attainment in the Netherlands during modernization."* ICS dissertation, Utrecht
232. Sanne Smith (2015), *"Ethnic segregation in friendship networks: Studies of its determinants in English, German, Dutch, and Swedish school classes."* ICS dissertation, Utrecht
233. Patrick Präg (2015), *"Social stratification and health: Four essays on social determinants of health and wellbeing."* ICS dissertation, Groningen
234. Wike Been (2015), *"European top managers' support for work-life arrangements"* ICS dissertation, Utrecht
235. André Grow (2016), *"Status differentiation: New insights from agent-based modeling and social network analysis."* ICS dissertation, Groningen
236. Jesper Rözer (2016), *"Family and personal networks: How a partner and children affect social relationships."* ICS dissertation, Utrecht
237. Kim Pattiselanno (2016), *"At your own risk: The importance of group dynamics and peer processes in adolescent peer groups for adolescents' involvement in risk behaviors."* ICS- dissertation, Groningen
238. Vincenz Frey (2016), *"Network formation and trust."* ICS dissertation, Utrecht
239. Rozemarijn van der Ploeg (2016), *"Be a buddy, not a bully? Four studies on social and emotional processes related to bullying, defending, and victimization."* ICS dissertation, Groningen
240. Tali Spiegel (2016), *"Identity, career trajectories and wellbeing: A closer look at individuals with degenerative eye conditions."* ICS- dissertation, Groningen
241. Felix Christian Tropf (2016), *"Social science genetics and fertility."* ICS dissertation, Groningen

242. Sara Geven (2016), *"Adolescent problem behavior in school: The role of peer networks."* ICS dissertation, Utrecht
243. Josja Rokven (2016), *"The victimization-offending relationship from a longitudinal perspective."* ICS dissertation, Nijmegen
244. Maja Djundeva (2016), *"Healthy ageing in context: Family welfare state and the life course."* ICS dissertation, Groningen
245. Mark Visser (2017), *"Inequality between older workers and older couples in the Netherlands: A dynamic life course perspective on educational and social class differences in the late career."* ICS dissertation, Nijmegen
246. Beau Oldenburg (2017), *"Bullying in schools: The role of teachers and classmates."* ICS dissertation, Groningen
247. Tatang Muttaqin (2017), *"The education divide in Indonesia: Four essays on determinants of unequal access to and quality of education."* ICS dissertation, Groningen
248. Margriet van Hek (2017), *"Gender inequality in educational attainment and reading performance: A contextual approach."* ICS dissertation, Nijmegen
249. Melissa Verhoef-van Dorp (2017), *"Work schedules, childcare and well-being: Essays on the associations between modern-day job characteristics, childcare arrangements and the well-being of parents and children."* ICS dissertation, Utrecht
250. Timo Septer (2017), *"Goal priorities, cognition and conflict: Analyses of cognitive maps concerning organizational change."* ICS dissertation, Groningen
251. Bas Hofstra (2017), *"Online social networks: Essays on membership, privacy, and structure."* ICS dissertation, Utrecht
252. Yassine Khoudja (2018), *"Women's labor market participation across ethnic groups: The role of household conditions, gender role attitudes, and religiosity in different national contexts."* ICS dissertation, Utrecht
253. Joran Laméris (2018), *"Living together in diversity: Whether, why and where ethnic diversity affects social cohesion."* ICS dissertation, Nijmegen
254. Maaike van der Vleuten (2018), *"Gendered Choices: Fields of study of adolescents in the Netherlands."* ICS dissertation, Utrecht
255. Mala Sondang Silitonga (2018), *"Corruption in Indonesia: The impact of institutional change, norms, and networks."* ICS dissertation, Groningen
256. Manja Coopmans (2018), *"Rituals of the past in the context of the present: The role of Remembrance Day and Liberation Day in Dutch society."* ICS dissertation, Utrecht
257. Paul Hindriks (2018), *"The struggle for power: Attitudes towards the political participation of ethnic minorities."* ICS dissertation, Utrecht
258. Nynke Niezink (2018), *"Modeling the dynamics of networks and continuous behavior."* ICS dissertation, Groningen
259. Simon de Bruijn (2018), *"Reaching agreement after divorce and separation: Essays on the effectiveness of parenting plans and divorce mediation."* ICS dissertation, Utrecht
260. Susanne van 't Hoff-de Goede (2018), *"While you were locked up: An empirical study on the characteristics, social surroundings and wellbeing of partners of prisoners in The Netherlands."* ICS dissertation, Utrecht

261. Loes van Rijsewijk (2018), *"Antecedents and consequences of helping among adolescents."* ICS dissertation, Groningen
262. Mariola Gremmen (2018), *"Social network processes and academic functioning: The role of peers in students' school well-being, academic engagement, and academic achievement."* ICS dissertation, Groningen
263. Jeanette Renema (2018), *"Immigrants' support for welfare spending: The causes and consequences of welfare usage and welfare knowledgeability."* ICS dissertation, Nijmegen
264. Suwatin Miharti (2018), *"Community health centers in Indonesia in the era of decentralization: The impact of structure, staff composition and management on health outcomes."* ICS dissertation, Groningen
265. Chaïm la Roi (2019), *"Stigma and stress: Studies on attitudes towards sexual minority orientations and the association between sexual orientation and mental health."* ICS dissertation, Groningen
266. Jelle Lössbroek (2019), *"Turning grey into gold: Employer-employee interplay in an ageing workforce."* ICS dissertation, Utrecht
267. Nikki van Gerwen (2019), *"Employee cooperation through training: A multi-method approach."* ICS dissertation, Utrecht
268. Paula Thijs (2019), *"Trends in cultural conservatism: The role of educational expansion, secularisation, and changing national contexts."* ICS dissertation, Nijmegen
269. Renske Verweij (2019), *"Understanding childlessness: Unravelling the link with genes and the socio-environment."* ICS dissertation, Groningen
270. Niels Blom (2019), *"Partner relationship quality under pressing work conditions: Longitudinal and cross-national investigation."* ICS dissertation, Nijmegen
271. Müge Simsek (2019), *"The dynamics of religion among native and immigrant youth in Western Europe."* ICS dissertation, Utrecht
272. Leonie van Breeschoten (2019), *"Combining a career and childcare: The use and usefulness of work-family policies in European organizations."* ICS dissertation, Utrecht
273. Roos van der Zwan (2019), *"The political representation of ethnic minorities and their vote choice."* ICS dissertation, Nijmegen
274. Ashwin Rambaran (2019), *"The classroom as context for bullying: A social network approach."* ICS dissertation, Groningen
275. Dieko Bakker (2019), *"Cooperation and social control: Effects of preferences, institutions, and social structure."* ICS dissertation, Groningen
276. Femke van der Werf (2019), *"Shadow of a rainbow? National and ethnic belonging in Mauritius."* ICS dissertation, Utrecht
277. Robert Krause (2019), *"Multiple imputation for missing network data."* ICS dissertation, Groningen
278. Take Sipma (2020), *"Economic insecurity and populist radical right voting."* ICS dissertation, Nijmegen
279. Mathijs Kros (2020), *"The nature of negative contact: Studies on interethnic relations in Western societies."* ICS dissertation, Utrecht

280. Lonneke van den Berg (2020), *"Time to leave: Individual and contextual explanations for the timing of leaving home."* ICS dissertation, Amsterdam
281. Marianne Hooijsma (2020), *"Clashrooms: Interethnic peer relationships in schools."* ICS dissertation, Groningen
282. Marina Tulin (2020*), "Blind spots in social resource theory: Essays on the creation, maintenance and returns of social capital."* ICS dissertation, Amsterdam
283. Tessa Kaufman (2020), *"Toward tailored interventions: Explaining, assessing, and preventing persistent victimization of bullying."* ICS dissertation, Groningen
284. Lex Thijssen (2020), *"Racial and ethnic discrimination in western labor markets: Empirical evidence from field experiments."* ICS dissertation, Utrecht
285. Lukas Norbutas (2020), *"Trust on the dark web: An analysis of illegal online drug markets."* ICS dissertation, Utrecht
286. Tomáš Diviák (2020), *"Criminal networks: Actors, mechanisms, and structures."* ICS dissertation, Groningen
287. Tery Setiawan (2020), *"Support for interreligious conflict in Indonesia*." ICS dissertation, Nijmegen
288. Vera de Bel (2020), *"The ripple effect in family networks: Relational structures and well-being in divorced and non-divorced families*." ICS dissertation, Groningen
289. Diego Palacios (2020), *"How context and the perception of peers' behaviors shape relationships in adolescence: A multiplex social network perspective."* ICS dissertation, Groningen
290. Saskia Glas (2020), *"Where are the Muslim Feminists? Religiosity and support for gender equality in the Arab region*." ICS dissertation, Nijmegen
291. Tomas Turner-Zwinkels (2020), *"A new macro-micro approach to the study of political careers: Theoretical, Methodological and empirical challenges and solutions."* ICS dissertation, Groningen
292. Lotte Scheeren (2020), *"Not on the same track? Tracking age and gender inequality in education."* ICS dissertation, Amsterdam
293. Joris Broere (2020), *"Essays on how social network structure affects asymmetric coordination and trust*." ICS dissertation, Utrecht
294. Marcus Kristiansen (2021*), "Contact with benefits: How social networks affect benefit receipt dynamics in the Netherlands."* ICS dissertation, Utrecht
295. Judith Kas (2021), *"Trust and reputation in the peer-to-peer platform economy*." ICS dissertation, Utrecht
296. Andrea Forster (2021), *"Navigating educational institutions: Mechanisms of educational inequality and social mobility in different educational systems."* ICS dissertation, Amsterdam
297. Jannes ten Berge (2021), *"Technological change and work: The relation between technology implementation within organizations and changes in workers' employment."* ICS dissertation, Utrecht
298. Jolien Geerlings (2021), "*Teaching in culturally diverse classrooms: The importance of dyadic relations between teachers and children*." ICS dissertation, Utrecht
299. Kirsten van Houdt (2021), *"Stepfamilies in adulthood: Solidarity between parents and adult children*." ICS dissertation, Amsterdam

300. Suzanne de Leeuw (2021), *"The intergenerational transmission of educational attainment after divorce and remarriage."* ICS dissertation, Amsterdam
301. Fleur Goedkoop (2021), *"Involvement in bottom-up energy transitions: The role of local and contextual embeddedness."* ICS dissertation, Groningen
302. Eva Vriens (2021), *"Mutualism in the 21st century: The why, when, and how behind successful risk-sharing institutions."* ICS dissertation, Utrecht
303. Ardita Muja (2021), *"From school to work: The role of the vocational specificity of education in young people's labor market integration."* ICS dissertation, Nijmegen
304. Siyang Kong (2021), *"Women and work in contemporary China: The effect of market transition on women's employment, earnings, and status attainment."* ICS dissertation, Utrecht
305. Marijn Keijzer (2022), *"Opinion dynamics in online social media."* ICS dissertation, Groningen
306. Sander Kunst (2022), *"The educational divide in openness towards globalisation in Western Europe."* ICS dissertation, Amsterdam
307. Nella Geurts (2022), *"Puzzling pathways: The integration paradox among migrants in Western Europe."* ICS dissertation, Nijmegen
308. Dragana Stojmenovska (2022), *"Men's place: The incomplete integration of women in workplace authority."* ICS dissertation, Amsterdam
309. Bram Hogendoorn (2022), *"Divorce and inequality: Stratification in the risk and consequences of union dissolution."* ICS dissertation, Amsterdam
310. Tom Nijs (2022), *"This place is ours: Collective psychological ownership and its social consequences."* ICS dissertation, Utrecht
311. Nora Storz (2022), *"'This land is ours—but is it also theirs?' Collective ownership beliefs and reconciliation in territorial conflict regions."* ICS dissertation, Utrecht
312. Tara Koster (2022), *"Parenting and fairness in diverse families."* ICS dissertation, Utrecht
313. Danelien van Aalst (2022), *"Elements Contributing to Teachers' Role in Bullying."* ICS dissertation, Groningen
314. Wybren Nooitgedagt (2022), *"Who owns the country? Collective psychological ownership and intergroup relations in settler societies."* ICS dissertation, Utrecht
315. Marija Dangubić (2022), *"Rejecting Muslim minority practices: Principles and prejudices."* ICS dissertation, Utrecht
316. Sara Cvetkovska (2022), *"Lines in the shifting sand: The implications of being tolerated."* ICS dissertation, Utrecht
317. Maaike Hornstra (2022), *"Going beyond the dyad: Adult intergenerational closeness after divorce and remarriage."* ICS dissertation, Amsterdam
318. Wouter Kiekens (2022), *"Sexual and gender minority youth's mental health and substance use: Disparities, mechanisms, and protective factors."* ICS dissertation, Groningen
319. Carlijn Bussemakers (2022), *"Adversity and educational inequality: The interplay between adverse experiences and parental resources for children's educational attainment."* ICS dissertation, Nijmegen

320. Evi Velthuis (2022), "'*To tolerate or not to tolerate?*' *Reasons for tolerance of minority group practices among majority members in the Netherlands and Germany."* ICS dissertation, Utrecht
321. Hendrik Nunner (2023), *"Co-evolution of social networks and infectious diseases."* ICS dissertation, Utrecht
322. Carly van Mensvoort (2023), *"Inspiring leaders: An empirical study of female supervisors at work."* ICS dissertation, Nijmegen
323. Anne van der Put (2023), *"Healthy at work: The role of the work environment in worksite health promotion."* ICS dissertation, Utrecht
324. Rowan ten Kate (2023), *"Understanding loneliness among older migrants*." ICS dissertation, Groningen
325. Sanne Kellij (2023), *"I see, I see what you don't see: Neural and behavioral social-cognitive processes underlying (persistent) victimization."* ICS dissertation, Groningen
326. Inge Hendriks (2023), *"Changes and contrasts in attitudes towards ethnic minorities."* ICS dissertation, Nijmegen
327. Eleonora Marucci (2023), *"Antecedents and consequences of teacher attunement in primary and secondary school classrooms."* ICS dissertation, Groningen
328. Kasper Otten (2023), *"Cooperation in changing groups: How newcomers and norms shape public good provision in the lab, online games, and the field."* ICS dissertation, Utrecht
329. Carlos de Matos Fernandes (2023), *"In or out? The paradox of exclusionary mechanisms in keeping cooperation going."* ICS dissertation, Groningen
330. Ruohuang Jiao (2023), *"Reputation effects in peer-to-peer online markets: meta-analyses and laboratory experiments."* ICS dissertation, Utrecht
331. Marlou Ramaekers (2024), *"Informal helping. Insights from a dyadic, family and societal perspective."* ICS dissertation, Nijmegen
332. Kim Stienstra (2024), *"Educational quality and inequality: The interplay between schools, families, and genes."* ICS dissertation, Utrecht
333. Ece Arat (2024), *"Diverse stepfamilies: Parenting and children's well-being."* ICS dissertation, Utrecht
334. Christian Fang (2024), *"Family life in postdivorce families."* ICS dissertation, Utrecht
335. Vera Buijs (2024), *"Close relationships and subjective well-being: A life course perspective on social needs, relations, and interactions."* ICS dissertation, Groningen
336. Twan Huijsmans (2024), *"Our place in politics: Urban-rural divergence and how place affects political attitudes."* ICS dissertation, Amsterdam
337. Thomas Teekens (2024), *"Sustainable collaboration in care: Joint production motivation and interprofessional learning in an interorganizational network."* ICS dissertation, Groningen
338. Ana Macanović (2024), *"Trust in the shadows: The role of communication in extra-legal contexts*." ICS dissertation, Utrecht
339. Christoph Janietz (2024), *"Inequality at work: Occupations, organizations, and the wage distribution."* ICS dissertation, Amsterdam

340. Philipp Schneider (2024), *"Social Influence and the Energy Transition: Leveraging social networks and norms."* ICS dissertation, Utrecht
341. Nick Wuestenenk (2024), "*Support for sexual liberalization among ethnic majorities and minorities in Europe: The role of social norms in the public expression of opinions.*" ICS dissertation, Utrecht
342. Renae Sze Ming Loh (2024), *"Copy Paste? Digital skills, social reproduction, and social mobility in education."* ICS dissertation, Nijmegen
343. Klara Raiber (2024), *"Beyond sacrifice? Long-term employment consequences of providing unpaid care."* ICS dissertation, Nijmegen
344. Sofie Lorijn (2024), "*Peer relationships in the transition from primary to secondary education.*" ICS dissertation, Groningen
345. Xingna Qin (2024), "*In connection – The role of peers, parents, and teachers in adolescent friendship dynamics."* ICS dissertation, Groningen
346. Echo Teng Li (2024), *"Simulation models of the collective consequences of bounded rationality in opinion formation in networks: Cases of market concentration and vaccination opinion polarization."* ICS dissertation, Groningen
347. Thomas Feliciani (2025), *"Divided spaces and divided opinions: Modeling the impact of residential segregation on opinion polarization."* ICS dissertation, Groningen
348. Thijmen Jeroense (2025), *"Social network segregation: Studies on network homogeneity in sociodemographic characteristics and political attitudes."* ICS dissertation, Nijmegen
349. Dieuwke Zwier (2025), "*Rhythms of class: Socio-economic disparities and peer dynamics in secondary school choice*." ICS dissertation, Amsterdam
350. Lian van Vemde (2025), *"Creating inclusive societies: Fostering belonging and positive intergroup relations in culturally diverse classrooms."* ICS dissertation, Utrecht
351. Rob Franken (2025), *"Social networks and sports participation: The interplay of selection and influence processes."* ICS dissertation, Nijmegen
352. Julian Rengers (2025), *"Selective disclosure at work: Lesbian, gay, and bisexual employees' disclosure decisions, motivations, and approaches."* ICS dissertation, Groningen
353. Sara Wiertsema (2025), "*Young adults in motion: Understanding the impact of the school-to-work transition on sports and physical activity."* ICS dissertation, Nijmegen
354. Sofie Wiersma (2025), *"Imprints at work: How the pasts of organizations and leaders shape workplace precarity and inequality."* ICS dissertation, Groningen
355. Tessa Ubels (2025), *"Stuck in migration: Pathways to social change and well-being for people on the move and the role of psychosocial support."* ICS dissertation, Nijmegen
356. Katrin Müller (2025), *"Shifting perspectives: majority members' perceptions of the prevalence of ethno-racial minority discrimination."* ICS dissertation, Nijmegen
357. Maikel Meijeren (2026), *"Helping those in need: volunteering for humanitarian organizations and for refugees."* ICS dissertation, Nijmegen
358. Jonas Stein (2026), *"Reconciling epistemic and identity diversity: Identifying pathways to better decision-making in social groups."* ICS dissertation, Groningen

Biographical information

Jonas Stein obtained a bachelor's degree in Sociology at the University of Mannheim, Germany in 2017. During his years in Mannheim, he held several teaching assistantship positions and completed an internship at the Institute for Economics and Social Sciences of the Hans-Böckler Foundation. His Bachelor's thesis on social mobility in the German tertiary education system was awarded best of his cohort.

In 2018, Jonas joined the 'Sociology and Social Research' master's program at Utrecht University, where he graduated *cum laude* and as the faculty's valedictorian two years later. He complemented his education with a stay at the Computational Social Science Team at Centre Marc Bloch and at the Institute of Analytical Sociology, Linköping.

In 2020, he began his PhD research at the Norms and Networks cluster of the Sociology department, University of Groningen. From 2022 to 2025, Jonas assisted with the daily operations of *Rationality and Society* as the journal's Managing Editor. As part of his PhD trajectory, he worked at the Institute of Advanced Studies, Toulouse with Prof. Maxime Derex in 2024. His doctoral training was co-hosted by the Interuniversity Center for Social Science Theory and Methodology (ICS) and the research program Sustainable Cooperation – Roadmaps to Resilient Societies (SCOOP).

As of July 2025, Jonas will continue his academic pursuits at the University of Groningen as a member of Prof. Carsten de Dreu's research group.

university of groningen

SCOOP
SUSTAINABLE COOPERATION

Diverse groups encompass a wide range of insights and perspectives, which can enhance the quality of the decisions they make together. Yet, differences in social, cultural, or professional backgrounds can also give rise to friction and misunderstanding, making it difficult for groups to realize their potential. This dissertation examines how groups can mitigate the challenges associated with diversity while harnessing the benefits it affords.

Jonas Stein conducted the research in this dissertation at the Department of Sociology and the ICS at the University of Groningen. His work is part of the transdisciplinary research program Sustainable Cooperation – Roadmaps to Resilient Societies (SCOOP).