CONSTRUCTOR UNIVERSITY

The Complex Effects of Distorted Social Perceptions on Opinions about Climate Change

by

Peter Steiglechner

a thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy
in Sociology

Approved Dissertation Committee

Prof. Dr. Agostino Merico
Constructor University
Leibniz-Centre for Tropical Marine Research (ZMT)

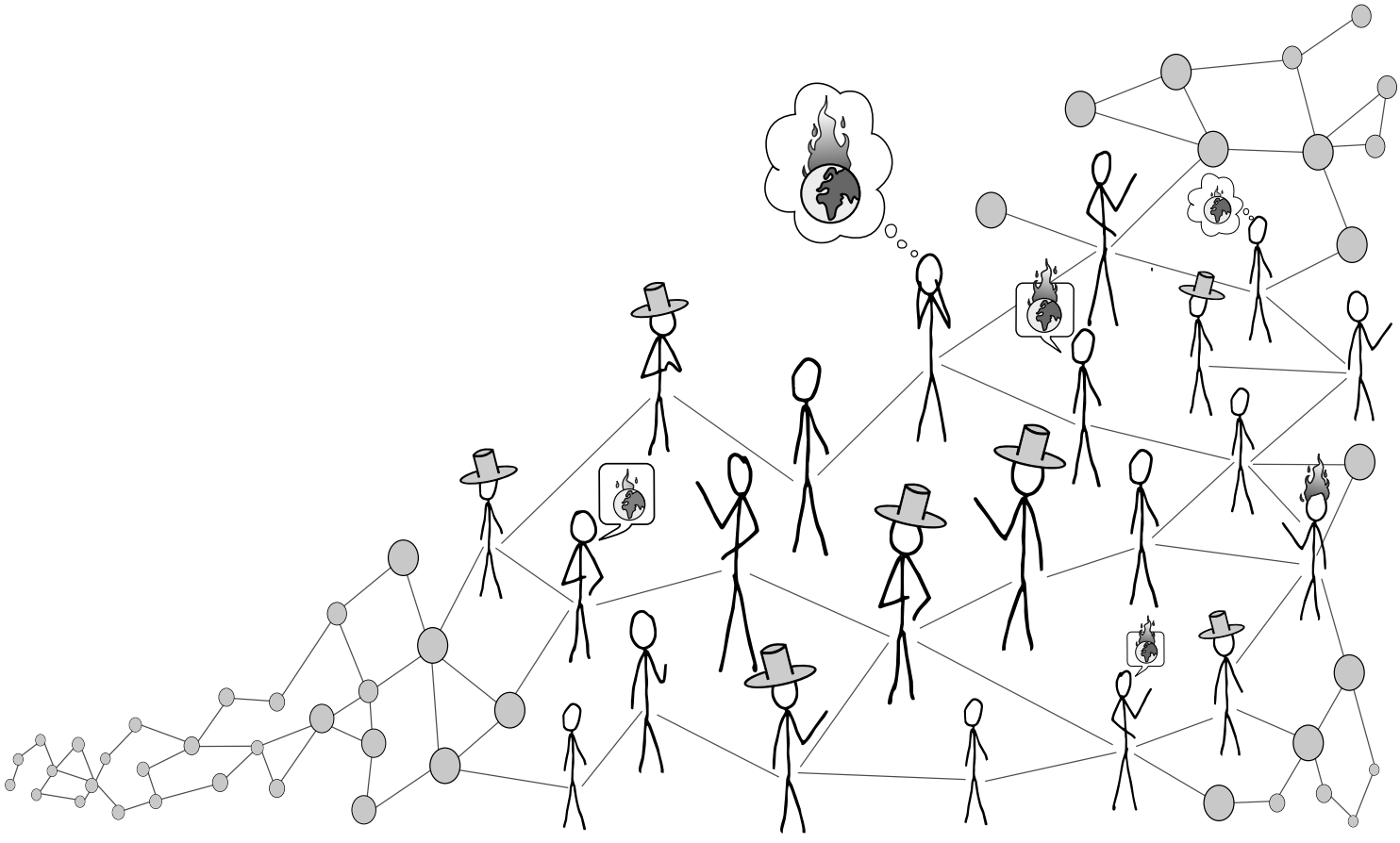
Prof. Dr. Achim Schlüter
Constructor University
Leibniz-Centre for Tropical Marine Research (ZMT)

Prof. Dr. Paul E. Smaldino
University of California, Merced, USA
Santa Fe Institute, USA

Prof. Dr. Jan Lorenz
Constructor University

Date of Defense: 30th May 2024

School of Business, Social & Decision Sciences



Statutory Declaration

Name	Peter Steiglechner
Matriculation Number	20332482
Thesis type	PhD thesis

English: Declaration of Authorship

I hereby declare that the thesis submitted was created and written solely by myself without any external support. Any sources, direct or indirect, are marked as such. I am aware of the fact that the contents of the thesis in digital form may be revised with regard to usage of unauthorized aid as well as whether the whole or parts of it may be identified as plagiarism. I do agree my work to be entered into a database for it to be compared with existing sources, where it will remain in order to enable further comparisons with future theses. This does not grant any rights of reproduction and usage, however. The Thesis has been written independently and has not been submitted at any other university for the conferral of a PhD degree; neither has the thesis been previously published in full.

German: Erklärung der Autorenschaft (Urheberschaft)

Ich erkläre hiermit, dass die vorliegende Arbeit ohne fremde Hilfe ausschließlich von mir erstellt und geschrieben worden ist. Jedwede verwendeten Quellen, direkter oder indirekter Art, sind als solche kenntlich gemacht worden. Mir ist die Tatsache bewusst, dass der Inhalt der Thesis in digitaler Form geprüft werden kann im Hinblick darauf, ob es sich ganz oder in Teilen um ein Plagiat handelt. Ich bin damit einverstanden, dass meine Arbeit in einer Datenbank eingegeben werden kann, um mit bereits bestehenden Quellen verglichen zu werden und dort auch verbleibt, um mit zukünftigen Arbeiten verglichen werden zu können. Dies berechtigt jedoch nicht zur Verwendung oder Vervielfältigung. Diese Arbeit wurde in der vorliegenden Form weder einer anderen Prüfungsbehörde vorgelegt noch wurde das Gesamtdokument bisher veröffentlicht.

Bremen, Tuesday 20th August, 2024

Place, Date Signature

Declaration on the contribution to multi-author sections

Chapter 2	Title	Noise and opinion dynamics: How ambiguity promotes pro-majority consensus in the presence of confirmation bias
	Journal	Royal Society Open Science (*in press*)
	Authors	Peter Steiglechner, Marijn A. Keijzer, Paul E. Smaldino, Deyshawn Moser, Agostino Merico
	Contributions	(i) PS, PES and AM conceptualised the project, (ii) PS designed the model, (iii) PS performed simulation runs and analysed the data, (iv) PS wrote the original draft with contributions from MAK, and (v) PS, MAK, PES, DM and AM reviewed and edited the final draft
Chapter 3	Title	Social identity bias and communication network clustering interact to shape patterns of opinion dynamics
	Journal	Journal of the Royal Society Interface (*published*)
	Authors	Peter Steiglechner, Paul E. Smaldino, Deyshawn Moser, Agostino Merico
	Contributions	(i) PS, PES and AM conceptualised the project, (ii) PS designed the model, (iii) PS performed simulation runs and analysed the data, (iv) PS wrote the original draft, and (v) PS, PES, DM and AM reviewed and edited the final draft
Chapter 4	Title	Perceived and actual opinion polarisation in the German climate change debate
	Journal	to be submitted to a peer-reviewed journal
	Authors	Peter Steiglechner, Paul E. Smaldino, Agostino Merico
	Contributions	(i) PS, PES and AM conceptualised the project, (ii) PS designed the model, (iii) PS performed all the analysis, (iv) PS wrote the original draft, and (v) PS, PES and AM reviewed and edited the draft.

Acknowledgements

Foremost, I am very grateful to my PhD Supervisor, Agostino Merico. It has been a true pleasure working with such an enthusiastic, optimistic, and trusting scientist. I am really happy that I got involved in this project. Beyond your role as a supervisor, you have become an essential scientific mentor to me. I've enjoyed learning from you, whether it was the tedious process of writing, presenting to an audiences with little prior interest or knowledge in my work, or emphasising its value. To be fair, I often enough argued against its value (sorry) and had to be convinced by you. We had many long and intense discussions but I believe that those conversations really made me learn how to improve my research work. I really value also our discussions on politics and society and the fact that you never forgot to emphasise that, ultimately, work should be fun.

I am especially thankful also to Paul Smaldino. You have taken on a role as a co-supervisor in this project without hesitation during a period when I was looking for feedback and direction in the wide field of opinion formation and social modelling. I am not sure where this PhD would have drifted without our discussions and your comments. I am also very glad that you gave me the opportunity to spend an enriching and enjoyable time in California and at SFI, which introduced me to a quite new perspective on the various directions of social modelling.

I also want to thank my further committee members: Achim Schlüter for providing insightful feedback from a cross-disciplinary perspective and Jan Lorenz for providing a very detailed and enriching analysis of the modelling work in my PhD thesis despite joining the committee at an extremely late stage.

There are many more colleagues (now friends) who have made working on this PhD better and more enjoyable. This includes especially Marijn Keijzer, Deyshawn Moser, Fridolin Haag, the systems ecology group at ZMT Bremen and all other researchers and staff there, the Department of Cognitive and Information Sciences at UC Merced, the people I met at the conferences or summer schools (Jan, Bruce, František, Marlene, ...), and many more. I am extremely grateful to have colleagues, friends, and family who eagerly engaged in discussions and (unknowingly) gave me inspiration, who listened to early versions of my presentations, who did not stop me when I got overly excited about the work, or who endured my lengthy complaints about it and still found ways to support me. Thank you! There are too many to name, but those who accompanied this PhD adventure intensively and throughout nearly all four years include Debbie, Nico, Selma, Sarah, Karin, Johannes, Franzi, Christian, Anni, and Jakob, as well as my various flatmates over the past years and all of you that I may have missed at the moment.

Spanning from 2020 to 2024, this PhD saw a whole pandemic, the invasion of Russia into Ukraine (and many more ongoing conflicts in Israel and Palestine, Yemen, Sudan, ...), the end of the Trump presidency with the US democracy in a pivotal phase, and an ever accelerating climate crisis. In the middle of these challenges, I could sit at a desk, study models of social dynamics, and write down my conclusions, hoping that they provide a valuable contribution to the important goal of better understanding the mechanisms driving our social world—what a privilege. So, I thank society for this opportunity. Lastly, I thank the German Research Foundation (DFG) for funding this research project.

Contents

Abstract

Polarisation is a great concern in current social and political debates. A divergence of opinions or, more generally, a lack of societal agreement, for example on fundamental problems like climate change, presents a barrier to rapid action against a looming crisis. There are many theories on why people polarise on certain topics. However, the drivers of polarisation in social environments are multi-faceted and involve complex feedbacks among social, cognitive, and structural processes. While humans require interactions with each other to form shared views and cooperate effectively on many problems, social influence can produce a variety of opinion patterns, such as consensus, persistent disagreement, or polarisation. In this thesis, I develop mathematical models of opinion formation or perception to uncover the conditions underpinning the emergence of different patterns. In particular, I formalise how psychological factors distort the way individuals perceive others into a mathematical language and analyse how these perceptions affect the formation of consensus or the persistence of disagreement in a virtual society. Each chapter focuses on a different factor. In Chapter 2, an interplay of ambiguity in communication and confirmation bias in information processing affects how individuals adapt their opinions to social influence. In Chapter 3, social identities and related in-group bias affect the degree to which individuals are influenced by others. In Chapter 4, peoples' subjective representations of the opinions of others change in time and, thereby, affect the degree of perceived polarisation. The models rely on theoretical insights about social processes and, partly, on empirical data obtained from surveys about climate change opinions. Taken together, the three studies demonstrate that the factors distorting people's perceptions or responses to social influence—noise, bias, or subjective perception—have a non-negligible and sometimes surprising impact on collective opinion patterns. First, there is an optimal combination of moderate confirmation bias and moderate ambiguity in communication for a group to reach pro-environmental agreement (Chapter 2). Second, in-group bias typically strengthens disagreement between different social identity groups. But moderate in-group bias can also facilitate agreement when a society is characterised by a very dispersed interaction network. This surprising effect emerges because the bias increases the alignment within one in-group, thus accelerating the spread of an opinion across the dispersed network (Chapter 3). Third, individuals may perceive more ideological polarisation than actually exists when the way they represent the opinions of others is shaped by political identity groups and those groups become increasingly aligned. My analysis of survey data on climate-related opinions of German citizens shows that ideological polarisation on climate change may indeed be exaggerated. However, perceived polarisation varies greatly across political groups (Chapter 4). This thesis demonstrates the importance of better understanding the mechanisms behind social phenomena and their non-trivial consequences on opinion dynamics. While the models and the conclusions presented in the thesis may not be readily used to predict opinion patterns, owing to the complexity and inherent uncertainty of our society, they contribute to the social sciences by demonstrating counter-intuitive consequences of seemingly obvious theoretical assumptions, highlighting gaps and potentially critical ambiguities in social theories, and suggesting future directions for empirical analysis.

Chapter 1

Introduction

1.1 Climate change and public opinion—a natural and sociopolitical challenge

The climate crises is one of the biggest and most fundamental challenges humanity faces today. Starting in the 19th century, the industrialisation has driven medical, scientific, and economic advances that have led to an exponential growth of the human population with increasingly high average standard of living. However, being powered by steam and internal combustion engines, the industrialisation has also pushed the concentration of greenhouse gases in the atmosphere to levels that are now $50\,\%$ higher than in pre-industrial times. We are already observing the impacts of the increased greenhouse effect on the climate with global average temperature up by more than 1 degree (IPCC, 2023), heat waves of unprecedented severity and frequency (Hansen et al., 2012), and numerous other natural disasters, such as droughts or floods (Ripple et al., 2022). The situation could become much more critical, with several major tipping elements of the climate system, such as the West Antarctic and the Greenland ice-sheets or the Atlantic meridional overturning circulation (the gulf stream), expected to collapse in the near future (Lenton et al., 2019). Crossing such tipping points—and potentially triggering a tipping cascade—will lead to unimaginable consequences for humans and ecosystems in the coming decades and centuries, including the irreversible loss of biodiversity, extreme heat and weather events, increased likelihood of pandemics, or freshwater scarcity (IPCC, 2023; World Economic Forum, 2024). In sum, unabated climate change is a threat to human well-being and planetary health—or, put simply, 'we are in a state of planetary emergency' (Lenton et al., 2019). These facts have long been known to the scientific community, as reflected by the IPCC reports since 1990 or the famous 'Limits to Growth' report by the Club of Rome (Meadows et al., 1972). Scientists across disciplines have virtually reached a consensus about the climate crisis and its anthropogenic causes (Myers et al., 2021) and an overwhelming majority calls for urgent action to mitigate the consequences of a looming catastrophe (Ripple et al., 2022).

While the science agrees on the fundamental questions about climate change, opinions among the general public are much more discordant. There is reason for optimism for climate advocates: in the recent past, peoples' opinions on climate change have shifted remarkably towards a higher awareness about anthropogenic climate impacts and a stronger concern about its consequences, as indicated, for example, in many surveys (Andre et al., 2024; Leiserowitz et al., 2021; Pew Research Center, 2022). In Europe, this development is

probably most evident in the surging popularity of grassroots climate activist movements such as 'Fridays for Future' (FFF)—a recurring global school-strike initiated by Greta Thunberg in 2018—or 'Extinction Rebellion' (XR)—a group that uses civil disobedience to raise public awareness for climate tipping points. At the same time, societies appear deeply polarised (Druckman et al., 2021) and this extends also to the issue of climate change (Dunlap et al., 2016; Falkenberg et al., 2022; Hornsey & Lewandowsky, 2022; Ross et al., 2019; Smith et al., 2024). Disagreement about the fundamental questions of climate change, such as its causes or the severity of its consequences, appears to persist despite the overwhelming and unanimous scientific evidence supporting them.

Tackling the climate crisis depends critically on social cohesion, stable political consensus, and a concerted effort by the public—all which are undermined by polarisation or persistent disagreement (Dunlap et al., 2016). While embracing diversity of political opinions is an indispensable part of the democratic process (Gutmann & Thompson, 2009; Smaldino et al., 2023), high (or even increasing) levels of disagreement among the public can also hamper political action, especially when the public increasingly disagrees on fact-based questions, such as whether climate change is caused by human activity or not and whether rapid and far-reaching mitigation efforts need to be pursued or not. Disagreement is associated with reduced social trust and cohesion, thereby undermines support for policies or established social norms that aim to address the crisis. For example, perceived polarisation may reduce people's willingness to act climate-friendly as people tend to cooperate only to the extent that they believe others will do so too (Andre et al., 2024; Gächter, 2007). Similarly, during the COVID-19 pandemic, the effectiveness of face masks to prevent the spread of the virus depended on people's compliance, that is on a certain level of agreement about the threat posed by a further spreading of the virus. Disagreement may also politicise a topic like climate change such that, even if a government implements long-term climate policies, these can be quickly overruled or undermined, once the government changes (Andre et al., 2024; Dixit & Weibull, 2007). Deteriorated social trust within a polarising society also provides a fertile ground for misinformation and fake-news (Lewandowsky et al., 2017; World Economic Forum, 2024). In sum, continued high levels of disagreement about fact-based questions regarding existential crisis like climate change are undesired because they impose a barrier to effective policy-making. The goal of this thesis is to better understand (i) the reasons why disagreement persists (or even increases) despite the scientific consensus on climate change, (ii) the factors that drive disagreement or polarisation, and (iii) the conditions under which people may be more likely to reach a consensus. While I use climate change as the prime example to study these questions, many conclusions are general and apply to opinion dynamics over a range of political topics.

1.2 Opinion formation in social environments

To understand how opinion patterns, such as consensus or polarisation, emerge on a societal or group-level, we need to understand how individuals form their opinions. The traditional explanation for the lack of public awareness and concern about climate change—known as the 'knowledge/information deficit model'—assumes that the public lacks (high-quality) information about the climate crisis and that the mere dissemination of such information should lead to more awareness of the crisis and its consequences (Suldovsky, 2017). Another popular explanation for the awareness gap is that news media, companies, or elites with vested interests play a vital role in driving polarisation (Bolsen & Shapiro, 2018). However, humans form their opinions in many different ways, involving complex and intertwined processes that act, simultaneously, on the cognitive, the affective, the behavioural, and the social levels. Thus, climate scepticism or indifference is not only the result

of a lack of knowledge in the public or influence from external actors (Cook & Overpeck, 2019; Ecker et al., 2022; Hornsey & Lewandowsky, 2022; Kahan et al., 2012): 'facts [alone] do not change minds' (Toomey, 2023). Climate change, like many other political topics, is a deeply social and psychological issue (Bliuc et al., 2015; Pearson et al., 2016).

Social influence is one of the key factors that does change the opinions of people (Deutsch & Gerard, 1955; Festinger, 1954; Latané, 1981; Turner et al., 1989). Humans are highly sensitive to the beliefs and actions of people in their social circles and tend to compare themselves in relation to others (Festinger, 1954). As such, interpersonal exchanges, such as discussions among individuals, observations of the choices and behaviours of other people and encounters with social norms, provide social cues for individuals to shape and adapt their opinions. Social influence is particularly important for highly complex and uncertain topics. Climate change is a prime example in this respect with its differential and diffuse social causes and impacts (Pearson et al., 2016). A person who, for example, listens to a friend's (climate-related) arguments for adopting a vegan diet, discusses her doubts about the scientific validity of climate change with relatives, receives social recognition for buying an electric car, or observes peers joining the school-strikes 'Fridays For Future', will likely adapt her own perspective on climate change following such social interaction.

According to classic theories of social influence (Deutsch & Gerard, 1955; French, 1956; Friedkin & Johnsen, 1990), interactions should lead to the convergence of opinions. When individuals observe the opinions or choices of other individuals, their own opinions adapt in order to better align with the observed ones. In other words, people strive for consistency within their social circles (Cialdini & Goldstein, 2004; Galesic et al., 2021). The degree of assimilation to social influence depends on the uncertainty about one's own and others' opinions: the more uncertain a person is about her own opinion, the more susceptible she is to social influence and the less impact she has on shaping the opinions of other individuals (Deutsch & Gerard, 1955). While social influence mostly leads to assimilation and convergence of opinions, it may also result in a repulsion from other (very dissimilar) opinions (Carpentras et al., 2022; Mäs et al., 2010). Such a 'negative' effect of social influence, however, is much less evident in empirical studies than the assimilative effect (Keijzer et al., 2024).

Assuming that social influence is predominantly assimilative, one should expect the opinions of people to increasingly align (Abelson, 1967). However, there are many social and psychological factors that undermine or constrain social influence. A major theme in this thesis is that people are subjective observers and do not perceive the opinions of others accurately. In the remainder of this section, I describe two types of mechanisms—systematic and random—that distort how people perceive the opinions of others. Systematic distortions reflect cognitive biases, which influence how humans process information. I focus particularly on cognitive biases related to social identities (Chapters 3 and 4). Random distortions correspond to sources of noise and ambiguity in communication or social influence (Chapter 2). I demonstrate in this thesis that such mechanisms can have relevant and sometimes counter-intuitive effects that need to be taken into account in order to understand the emergence of collective opinion patterns under social influence.

People form opinions and make decisions in situations that are characterised by great uncertainties, a multitude of (contradictory) personal and social goals, and an often overwhelming amount of information. The notion of a perfectly rational *homo oeconomicus* assumes that people navigate such situations by optimising their decisions with respect to their personal preferences. In reality, humans are constrained by limited computational capacities, time, memory, and information uptake (Gigerenzer & Goldstein, 1996). Hence, they use short-cuts and learn heuristics to process information or to arrive at conclusions (Kendal et al., 2018)—humans are bounded rational (Simon, 1955). For example, when people evaluate whether they find a social media post

offensive, they tend to use social cues, such as a hashtag, as an indicator of (dis-)similarity (Powell et al., 2023). There are two opposite perspectives on bounded rationality in the psychological literature: one focuses on how these short-cuts produce systematic errors in human judgement, referred to as cognitive biases, i.e. judgements that deviate from optimal, rational behaviour (Kahneman, 2012); the other focuses on how these short-cuts or heuristics represent adaptive general-purpose processing strategies that are optimised to realistic situations in our uncertain and complex world rather than erroneous compared to an idealised and specialised *homo oeconomicus* (Gigerenzer & Brighton, 2009; Kendal et al., 2018; Nilsson et al., 2016). Throughout this thesis, I mostly use the term bias to describe such behavioural patterns or rules but I do not mean to imply that these behaviours are irrational.

Cognitive biases generally play an important role in the formation and perception of climate change opinions and often they are assumed to have a devastating effect on raising awareness about the crisis (Beattie & McGuire, 2018; Johnson & Levin, 2009; Marshall, 2015) or reaching agreement (Bayes & Druckman, 2021). For example, positive illusions cause people to overestimate the ability to mitigate and adapt to climate change in the future (e.g. through new technological inventions) and thereby to underestimate messages calling for urgent action (Moser et al., 2022). Confirmation bias causes people to pay more attention to information that confirms their prior opinions and to neglect information that contradicts it (Nickerson, 1998). For example, confirmation bias may lead a climate sceptic to simply reject any information about the detrimental consequences of climate change. Cognitive biases, in sum, influence how accurately people perceive or process information. They can, thus, undermine opinion formation.

Given the importance of social influence for opinion formation, especially for topics like climate change, my main focus lies on those biases that are directly related to social processes. When biases distort how humans perceive others' opinions, they can limit or undermine the assimilative effect of social influence (Lord et al., 1979). One of the most dominant factors affecting social influence biases and opinion formation is social identity. People have a tendency to view themselves as part of groups of like-minded or similar individuals (Festinger, 1954). Such self-identification into in- and out-groups is typically based on shared attributes, including demographic (e.g. being part of a young generation), geographic (e.g. urban, Bavarian), cultural (e.g. punks, hippies, favourite sports team), or political (e.g. Democrats or Republicans in the US) characteristics (Tajfel, 1974). While people obviously internalise multiple identities simultaneously, some identities can become particularly salient depending on the context. For example, demographic identities have gained increasing attention in the climate change debate: disagreement on climate change is often framed as a conflict between generations wherein the future freedom of youngsters is being reduced by the current lifestyle of elders (Gonyea & Hudson, 2020; Meleady & Crisp, 2017; Ross et al., 2019; Swim et al., 2022). An identity is often connected to in-group bias (Deutsch & Gerard, 1955; Hewstone et al., 2002). This bias involves that people perceive in-group members—those with the same social identity—more favourably than out-group individuals[1]. Such distorted perception undermines the assimilative effects of social influence. In fact, there is ample evidence that identity plays a critical role for the perception and formation of opinions (DellaPosta et al., 2015; Flores et al., 2022; Macy et al., 2019), including opinions regarding climate change (Bliuc et al., 2015; McCright & Dunlap, 2011b; Pearson & Schuldt, 2018). While it seems intuitive that a systematic in-group bias should foster separations between groups, I demonstrate in Chapters 3 and 4 that its effects on collective opinion patterns or the perception of them are more diverse than what one may think.

[1] In-group bias is a good example for a case in which biased behaviour should not be confused with irrational behaviour. While it is a bias to selectively attend to opinions of people with similar demographics attributes, for example, it may be perfectly rational in the sense of fulfilling social goals such as strengthening the affiliation to a peer group (or one's status therein) and maintaining a positive self-concept (Cialdini & Goldstein, 2004).

Random distortions or noise in perceiving (socially-transmitted) information is another factor that influences opinion formation. To communicate an opinion to others requires a sender, who expresses the opinion, a communication channel, through which the information is transmitted, and a receiver, who perceives and processes the information. Noise represents random or confounding factors affecting any of these components. For example, language can be inherently ambiguous and, therefore, there is often some ambiguity in the verbal expression/perception of opinions (McMahan & Evans, 2018). Moreover, people often do not have made-up opinions in their mind but instead construct them on the spot, such that opinions expressed by the same person vary—people are a 'crowd-within' (Herzog & Hertwig, 2009; Vul & Pashler, 2008). For example, when people express their opinions on climate change, they are affected by random factors such as seasonality or current weather conditions (Borick & Rabe, 2014). In their recent book, Kahneman et al. (2021) present a collection of sources of noise and evidence for how this noise can affect people's opinions, decisions, and behaviours. While systematic distortions often have a direction, such as perceiving opinions of in-group members as more conclusive than opinions from out-group members, random distortions do not (Kahneman et al., 2021). That is, an individual may sometimes perceive the opinions of others as more certain and sometimes as less certain. One may intuitively assume that the effects of such random fluctuations should cancel out in the long run or average out over many individuals. But opinion formation is often non-linear and path dependent and it is not obvious, therefore, how noise affects the emergence of polarisation or consensus (Kahneman et al., 2022). In Chapter 2, I demonstrate that noise—in particular, noise connected to ambiguity in expressed opinions—can indeed have important, non-trivial effects on such collective opinion patterns.

1.3 Why we model opinion formation

In the previous section, I have introduced how social influence shapes opinions and how this process at the level of the individual is undermined by cognitive biases and noise, i.e. systematic and random distortions of perceptions. The goal of this thesis is to better understand how these distortions shape collective opinion patterns, especially in the context of climate change. Each chapter takes on a different perspective related to this goal, but their common thread is that they are all based on a translation of social phenomena into a mathematical language or, in particular, into mathematical or computational models[2]. With the term 'model', I refer to a mechanistic rather than a statistical model, i.e. a model that describes certain parts and processes of a social system within a mathematical framework and allows one to study the system by running simulation experiments. In its broadest sense, a model is a "logical engine for turning assumptions into conclusions" (Smaldino, 2017). That is, it establishes an explicit, transparent, and testable link between theoretical claims about social processes and their consequences for societies. To be able to explain how a certain phenomenon arises, one must at least be able to generate that phenomenon in an artificial society (Epstein, 1999). For example, to explain why differences in climate change opinions persist despite the assimilative effects of social influence, one needs to at least be able to robustly generate such persistence in a model version of a society which encapsulates the theoretical claims made in the suggested explanation. With this core principle in mind, modelling can be a valuable contribution to both empirical analysis and theory development in the social sciences (Edmonds et al., 2019; Epstein, 2008; Nowak et al., 2011; Smaldino, 2017; Smaldino et al., 2015). In the remainder of this section, I discuss three of these contributions that particularly motivated the work

[2]In the following, I will refer to these as 'mathematical models' because also computational models rely on a mathematical rather than a verbal description. The features of mechanistic mathematical modelling described in the following apply generally, although I focus solely on one particular kind of mathematical models, namely agent-based models, in the remainder of this thesis.

in this thesis: mathematical modelling accompanies theory refinement, it offers mechanistic explanations for empirical social macro-scale phenomena (as opposed to statistical models), and it is a tool that can capture most of the inherent complexity of social systems.

The first motivation is that mathematical modelling can guide and support the development and refinement of social theories (Smaldino, 2017; Wimsatt, 1987). In general, theories are often formulated as verbal models leaving space for ambiguous interpretation. Ambiguity can be beneficial especially in the early stages of theory development, but may have adverse consequences if this lack of clarity is not obvious or if precision is discouraged (Frankenhuis et al., 2023). In this sense, mathematical modelling can guide theory development in the following ways: First, the translation of verbal models into precise and explicit mathematical terms, can help to develop and refine mental models because mathematics enforces comprehensiveness, rigour, and consistency (Nowak et al., 2011). Second, it helps to communicate theories between scientists by specifying unambiguously what a theory captures and what it does not capture (Frankenhuis et al., 2023). Third, models can illuminate any gaps in existing theories (Epstein, 2008). Fourth, models can indicate what kind of data is missing to test a theory (Smaldino, 2019). For example, Hewstone et al. (2002) define in-group bias as a "systematic tendency to evaluate one's own membership group (the in-group) [...] more favourably than a non-membership group (the out-group)"[3]. While this is an intuitively appealing description, it allows for different interpretations in the context of opinion formation. What do 'evaluate' and 'favourably' mean in relation to social influence? Do people weigh information from in-group members as more important? Do they pay less attention to out-group opinions? Or do they reject dissimilar out-group opinions more readily than dissimilar in-group opinions? In-group bias likely combines all of these interpretations to some degree and many different interpretations may have a similar (if not the same) effect. Mathematical modelling aims to translate such verbal concepts into a formal language. In Chapter 3, for example, I suggest one way to implement the in-group bias in an opinion formation model. There are certainly different ways to do this with different consequences for the resulting opinion patterns. Modelling allows us to specify such distinct interpretations, to illustrate and communicate their differences, and to show their respective consequences. Similarly, in Chapter 2, I demonstrate that consensus formation in the presence of noise depends on where exactly this noise enters the social influence process. This suggests that we need to measure noise at different stages in the social influence process to understand its impacts. In sum, "the only alternative to using a formal [i.e. mathematical] model is to use a verbal model, or worse, an unspoken mental model" (Smaldino, 2020).

The second motivation for the type of models that I will present in this thesis is that they provide a mechanistic understanding of social processes, which often eludes statistical models driven by empirical data. Traditional empirical methods—experiments, surveys, and data mining—have different strengths and limitations. Experiments allow controlled exploration of social phenomena and the factors that interfere with it (Falk & Heckman, 2009), such as biases in social influence processes. However, ethical concerns often arise when manipulating opinions or behaviours of real individuals on large scales (see the experiments with 'facebook' users by Bond et al., 2012; Kramer et al., 2014). Moreover, one may question the real-world applicability of experimental effects, such as whether an intervention has implications beyond the experimental setting. Surveys aim to capture the distributions of opinions on a certain political topic from a representative sample of the population (e.g. Leiserowitz et al., 2021). Data mining—for example, using topic modelling or sentiment analysis of the content people share on social media platforms to obtain opinion data (e.g. Cinelli et al., 2021; Falkenberg et al., 2022)—offers a very cost-effective possibility to extract vast amounts of data (Sobkowicz et al., 2012).

[3]Note that I use in-group bias, intergroup bias (Hewstone et al., 2002), in-group/out-group bias (Moser et al., 2022), and social identity bias (Steiglechner et al., 2023) interchangeably.

However, both surveyed or mined opinion data typically lack the ability to uncover the mechanisms that lead to opinion change, let alone explain them. In the US, for example, empirical evidence shows that affective polarisation (an increasing dislike between socio-political groups Iyengar et al., 2012) and issue alignment (increasing clustering of opinions within such groups Kozlowski & Murphy, 2021) coincide, but it is not obvious whether and how they are causally related (Armaly & Enders, 2021). Similarly, empirical analysis can identify *who* is sceptical about climate change, but cannot explain *why* (Hornsey et al., 2016). This emphasises the fundamental difference between describing social phenomena and explaining them (Craver, 2006; Elsenbroich & Polhill, 2023). While empirical data are indispensable to identify patterns of social phenomena, mathematical modelling aims at explaining the mechanisms behind them and it achieves this in a very simple, fast, and cost-effective way.

Finally, mathematical modelling provides a language to describe the complex dynamics of social systems driven by inherent non-linear feedback processes (Bookstaber, 2017; Keijzer, 2022; Schill et al., 2019)—an aspect that neither theoretical nor empirical approaches are able to address adequately. Models explore which assumptions about the behaviour of individuals lead to which patterns at the collective level. Complexity arises between these levels, for example, because individuals interact with each other and their opinions aggregate in non-trivial ways. In other words, the link between micro-properties and macro-patterns in social systems is often not obvious, but mathematical models can capture it (Bookstaber, 2017; Keijzer, 2022). A famous example is Schelling's model of segregation emergence (Schelling, 1971). The model assumes a population comprising two ethnic communities who are initially randomly dispersed across a spatial grid (akin to a city). Schelling demonstrated that even small homophilic tendencies, such that individuals relocate when they find themselves in a minority, can drive communities to segregation, albeit unintended by any single individual. Similarly, Granovetter (1978)'s threshold model demonstrates the importance of small changes in individual behaviours. Individuals in the threshold model make binary choices, for example to participate in a climate protest or to abstain from it, based on their personal threshold of participation levels (some individuals participate regardless of prior participation levels, while others join only after a significant proportion does). Even minor shifts in the distribution of these thresholds can amplify through non-linear feedbacks and trigger a cascade of participation. In the present thesis, I also show examples of social phenomena in which small changes in the assumptions about individual-level processes can substantially alter the macro-outcome, such as whether a group of individuals reaches consensus or polarises. With the micro-macro-link being at the core of sociology (Keijzer, 2022), mathematical models provide a tool to demonstrate and explore relations that are hard to observe otherwise.

In sum, mathematical models allow us to design a virtual laboratory and investigate the dynamics produced by artificial societies. In this virtual laboratory, modellers have full control to isolate different drivers, run experiments asking 'what if' or 'how possibly' questions, such as 'what if people did not associate to social identities at all?' or 'how can we possibly obtain polarisation when social influence assimilates the opinions of individuals?'. Being a complex system, slightly different conditions in the social environment can lead to very different macro-scale opinion patterns. Thus, in contrast to statistical models, mechanistic models generate coherent stories of plausible futures—or alternative worlds—that are not necessarily a continuation of the past (Elsenbroich & Badham, 2023). There has been much debate recently (see, for example, Elsenbroich & Polhill, 2023), whether opinion dynamics models should aim at predicting real-world opinion formation (like weather models) rather than provide 'merely' qualitative insights (akin to idealised models of climate change). The purpose of the models that I consider in this thesis are not about prediction. However, as described above, this type of modelling is a valuable contribution to traditional social science tools, like theoretical explorations,

experiments, or surveys, to gain insights about the observed social phenomena, for example, to understand which opinion patterns are likely to dominate in certain social settings and which processes foster these pattern.

1.4 The basic components of an opinion formation model

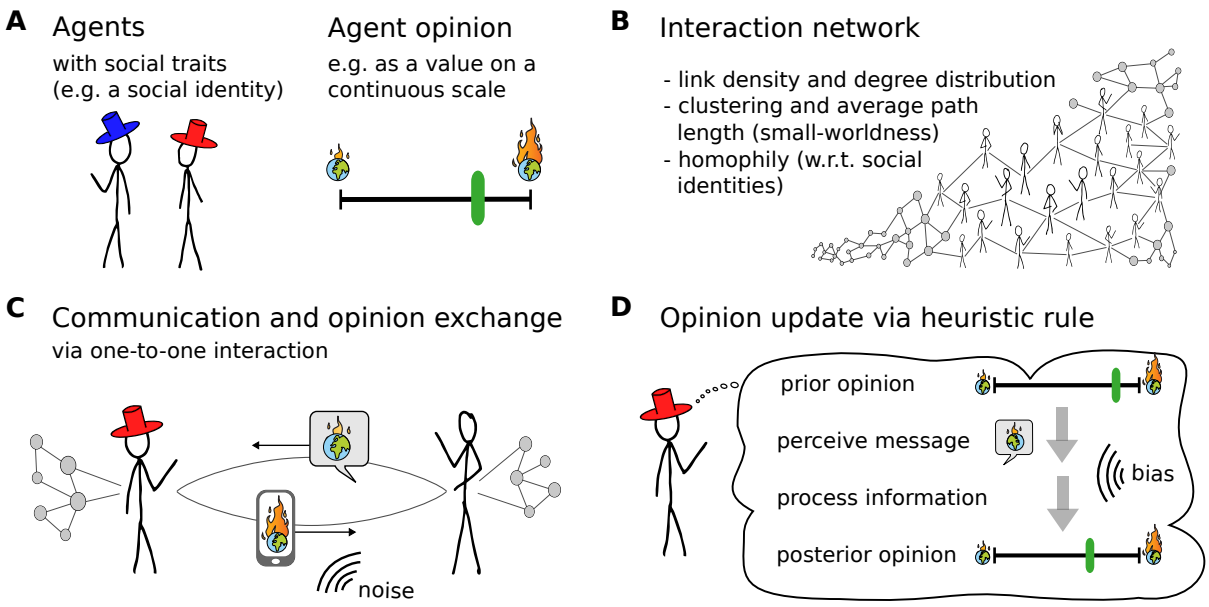


Figure 1.1: Overview of the four components of a typical opinion formation model with specific examples of implementations: (A) an agent with social attributes (here a social identity represented by the colour of the hat) and with an opinion on a specific topic like climate change (represented here as a numerical value between low and high concern about climate change), (B) the way agents are connected, which is represented in a social network with specific properties, such as link density, degree distribution, clustering vs. average path length—or the small-worldness, which implies high clustering and short average path length—, and homophily related to the agents' social identities, (C) the (potentially noisy) process of communicating opinions via one-to-one interaction through verbal or online interaction, and (D) the way individuals update opinions after a social interaction. Typically, this involves that the agent's prior opinion moves in the direction of the message, but the perception and processing of messages is affected by cognitive biases. In (D), for example, the agent's prior high level of concern about climate change is reduced somewhat after receiving a much less concerned message.

Opinion formation (or opinion dynamics) modelling is closely linked to the principles of Agent-Based Modelling (ABM). A system is described as a collection of its parts and the processes acting on these parts (Bonabeau, 2002). This notion is in contrast to equation-based models, which describe a system (and the processes acting on it) from a holistic perspective. ABM can provide valuable insights into the functioning of the investigated system. For example, while it is very difficult to describe a flock of birds from a system level, Reynolds suggested the famous 'boids' model to illustrate that these patterns emerge from very simple rules that describe how each individual bird flies depending on a few closest neighbours (Reynolds, 1987). This agent-based perspective lies also at the core of social modelling. In most opinion formation models, individual people are represented as agents, that is, as independent entities. These agents may have social traits, such as a social identity or a geographic location, and, crucially, hold opinions, which can be somehow expressed as mathematical objects. The agents interact with other agents and thereby influence or learn from each other. Following such interactions, the opinion of an agent changes and adapts. The model typically encapsulates one or multiple (stochastic) rules, which represent theories of human behaviour and information processing.

Such rules account for the bounded rationality of humans, for example, by specifying heuristics and systematic or random mechanisms that distort the way people perceive others and their opinions. By simulating how the opinions of the agents evolve over a number of discrete time-steps, ABMs generate a macro-state from the micro-level interactions. As mentioned above, this generative process is ideal to describe complex systems with emergent and dynamic macro-patterns (de Marchi & Page, 2014; Epstein, 1999), such as consensus or polarisation and the transitions between them. In this section, I describe the key ingredients of opinion formation models: (i) how opinions are represented, (ii) how social structures and interaction patterns are constructed, (iii) how information is exchanged between individual agents, and (iv) how this exchange leads to opinion update.

1.4.1 Opinions

One of the most important aspects of an opinion formation model is the way it quantifies opinions. In the following, x_i denotes the opinion of an individual i on a given issue. There are at least three major representations of opinions: as a binary or discrete value, as a value on a continuous numerical scale, and as a distribution over a continuous range of values (Figure 1.1A). The first option—representing an opinion as a binary or discrete value—builds on the analogy of opinions as magnetic spins (Castellano et al., 2009; Sznajd-Weron & Sznajd, 2000). Opinions are either spin up $x_i = 1$ or spin down $x_i = 0$, reflecting that many opinions in social settings boil down to binary (or very few) choices, such as deciding to vote, joining a protest, refusing a vaccine, or believing in anthropogenic climate change. The second option—representing an opinion as a numerical value on a continuous scale—is probably the most used in opinion formation models (Flache et al., 2017). Opinions can range between two extremes, $x_i \in [-1, 1]$, for example, where -1 represents not being worried about climate change at all and 1 being alarmed about it. Values between those extremes can represent either moderate opinions (being slightly concerned) or uncertainties about the direction (being concerned but less certain about that opinion). This notion of opinions as a point-estimate that can be located between two extremes mirrors the idea of Likert-scale response options in political surveys. The third option—representing an opinion as a (normalised) distribution $x_i = f_i(o)\ \forall\, o \in \mathcal{O} = [-1, 1]$, where $\mathcal{O}$ is the opinion space and $f_i(o)$ represents the support of agent i for all possible beliefs o (Galesic et al., 2021; Martins, 2009; Sobkowicz, 2018)—is an intriguing way to capture the direction, degree, and uncertainty of an opinion at the same time. For example, a narrow shape, such as a tight Gaussian distribution, indicates certainty about an opinion regardless of whether this opinion is moderate or extreme. Irrespective of the representation, many models assume that individuals have opinions on multiple issues. For example, one can represent opinions on climate change and income taxes as a multi-dimensional vector of (binary, continuous, or distributional) objects, which may change independently of each other (Axelrod, 1997) or may be logically related (Friedkin et al., 2016). While other social characteristics, such as social identities (indicated by the colour of the agents' hats in Figure 1.1A), are typically assumed to be fixed in opinion formation models, opinions evolve over time as the agents interact with each other.

1.4.2 Social structure

The second component of opinion formation models is the social structure, which determines the interactions between agents, i.e. who can share information with whom (Figure 1.1B). Most ABMs of social phenomena assume that agents are embedded in networks wherein nodes represent agents and links enable interactions

or flow of information between the agents. This notion is based on the assumption that agents have only access to limited, local information—they observe the opinions of their neighbours (such as friends, family, colleagues). In the context of opinion formation, the most relevant network characteristics are (i) its link density, (ii) clustering, and (iii) average path length between nodes. Additionally, in models including social identities, a network may be characterised also by homophily in relation to such identities.

Link density describes the average number of connections per agent. In reality, people are typically only influenced by a few social contacts, which limits the amount of information that they have access to. For example, Bond et al. (2012) found that an average 'facebook' user has 10 strong ties who exert a noticeable influence on the user's (intended) behaviour. Weak ties, such as acquaintances or distant friends, are often neglected in the context of opinion formation, although their role in the dissemination of information should not be underestimated (Centola & Macy, 2007; Granovetter, 1973). Not all people are equally well-connected in real social networks. The distribution of node degrees (i.e. links per node) is important because a high node degree is often equated with social power (Degroot, 1974; French, 1956)[4]. Moreover, links in real social networks are not randomly distributed (such as in the random networks by Erdos, Rényi, et al., 1960), but their structure typically exhibits small-world properties. Nodes are highly clustered with most links between them remaining relatively local. If, for instance, agent A is connected to B and C, it is likely that B and C are connected as well—also known as triadic closure. However, the average path length, i.e. the network distance between any two nodes, remains relatively short—a phenomenon that has become famous as the 'six degrees of separation' (Watts, 2004). These properties—sparse link density, high clustering, and short path lengths—are captured by small-world network topologies, such as those generated by the Watts-Strogatz mechanism (Watts & Strogatz, 1998), which has become a canonical way to implement small-world networks into social models. Finally, networks can reflect the tendency that humans interact more with those that they consider similar—an aspect known as homophily (Lazarsfeld & Merton, 1954; McPherson et al., 2001). For example, young individuals tend to interact predominantly with their young peers and supporters of a particular football club tend to interact more with other supporters of the same club. In networks, homophily means that the link density within social identity groups exceeds the link density between the groups (see Chapter 3).

The choice of network is an important consideration in opinion formation models because link density, degree distribution, network topology, and homophily all affect how opinions can spread through the population. For example, Schawe et al. (2021) or Centola (2022) describe the importance of agents that bridge between otherwise relatively isolated communities for consensus formation; Moser and Smaldino (2023) demonstrate that higher clustering can foster collective problem solving. Similarly, in Chapter 3, I show that the network topology determines whether in-group bias promotes or impedes consensus among agents of different social identities. Assuming that social relationships are relatively stable, many opinion formation models (including those that I present in Chapters 2 and 3) keep the networks fixed. There is, however, also a thriving research line modelling the co-evolution of networks and opinions (Will et al., 2020). Such models are thus able to depict, for example, the emergence of echo chambers as increasingly aligned and at the same time increasingly isolated groups.

[4] Although more advanced definitions of power typically account for further aspects such as the (complex) centrality of the agent's position in the network (Centola & Macy, 2007).

1.4.3 Information exchange

Humans express their opinions through implicit or explicit messages, for example, when interacting or discussing with others or when they observe others' decisions and behaviours (Figure 1.1C). But expression of opinions and interaction can be formalised in different ways in opinion formation models. A key choice is who sends information and who receives it. In offline or private interactions, influence is typically exerted via one-to-one interactions, i.e. a sender transmits an opinion to a receiver. In social media, a person may post their opinions publicly (one-to-many communication). A person may also be influenced by election polls or news about a protest, where many people express their opinions at once (many-to-one communication). Although often seemingly irrelevant, communication regimes can have an important effect on opinion patterns (Keijzer et al., 2018). In Chapters 2 and 3, I will mainly investigate one-to-one interactions.

Besides specifying who can send or receive messages, models also specify what information can be transmitted via the influence channel. First, socially-transmitted messages may be noisy. For example, when individuals indicate their opinions through well-known 'bumper stickers' like "Vote Trump" or "Atomkraft? Nein Danke!" ("nuclear power? no thanks!") or use clearly identifiable hashtags when they post a statement on a social media platform, the communicated messages are quite unambiguous. But other types of communication, such as political discussions, may involve ambiguous exchange. Second, an opinion message may reveal the social identity of the sender. For example, while online forums sometimes hide social identities, clothes, age, or lifestyle signs often represent clear markers of identity in offline interactions (Keijzer, 2022). Third, information channels may allow only certain kinds of opinion messages to be transmitted. For example, while individuals may have uncertain opinions (represented as values $x_i \in [-1, 1]$ in models), the messages they send are binary ($+$ for approval and $-$ for disapproval; such as in the models by Carpentras et al., 2022; Martins, 2008). These design choices about how agents exchange information in opinion formation models, in particular who can send and receive a message and what such a message entails, reflect different types of social systems and can affect the model behaviour.

1.4.4 Opinion update

The most important component of an opinion formation model is the way opinions evolve. Models comprise rules or heuristics that determine how agents incorporate socially-transmitted information to form their opinions (see Figure 1.1D for an exemplary update rule). Such rules can vary greatly in their complexity. The most simple update rule is that agents imitate the opinions of their neighbours after interacting with them. This rule is implemented, for example, in the voter model (Holley & Liggett, 1975), in which agents have binary spin-like opinions and an opinion flips when the agent is influenced by another agent with opposite opinion. A real-world example of such an opinion imitation is when an individual adopts a vegan diet after being influenced by a vegan person. When opinions are formalised as locations on a continuous scale (Figure 1.1B), opinion updating may involve more nuanced processes. For example, in many models the opinion x_i of an agent i moves in the direction of the opinion x_j of partner j with rate μ: $x_i \leftarrow x_i + \mu \cdot (x_j - x_i)$ (see sketch in Figure 1.1D). This type of opinion updating takes inspiration from early models by French (1956) and Degroot (1974). A real-world example of this type of updating is when a meat enthusiast adopts a slightly more sceptical opinion about her diet after discussing the topic with her vegan friend. More complex update rules implemented in opinion formation models assume that agents (unconsciously) incorporate social cues into their own opinions in a way that is inspired by Bayesian calculus. For example, the models by Martins

(2009) and Sobkowicz (2018)—which provide the basis for Chapter 3—assume that opinions are distributions over a continuous space and that an agent i forms its opinion, x_i, following the Bayesian theorem by taking the opinion distribution x_j of the interaction partner j as the best available approximation of the Bayesian likelihood:

$$x_i \leftarrow \frac{1}{N} \cdot x_i \cdot x_j \quad \text{where } x_i, x_j \text{ are distributions over } [-1,1] \text{ and } N \text{ is a normalisation.} \tag{1.1}$$

A real-world example of such an opinion update is when an insecure meat-eater (i.e. with a relatively broad opinion distribution skewed towards the meat-friendly side) becomes more certain about her diet after being influenced by a relatively secure meat-eater. Here, the broadness of the opinion distribution may represent either an insecurity about meat-eating (i.e. being undecided about whether to support it or not) or a high tolerance towards various choices (i.e. not caring much about meat-based or vegan diets), which are both affected by Bayesian opinion updating.

The update rule determines how individuals process information and is thus the ideal place to implement cognitive biases in models of opinion formation. For example, many models represent confirmation bias as bounded confidence in opinion updating (Deffuant et al., 2000; Hegselmann & Krause, 2002). Agents in these models only accept and adapt to social influences when they perceive the opinion of their interaction partner as sufficiently similar to their own, i.e. if the opinion distance does not exceed the confidence bound ϵ:

$$x_i \leftarrow \begin{cases} x_i + \mu \cdot (x_j - x_i) & \text{if } |x_i - x_j| \leq \epsilon \\ x_i & \text{else} \end{cases} \quad \text{where } x_i, x_j \in [-1,1]. \tag{1.2}$$

Here, $1 - \epsilon/2$ represents the strength of the confirmation bias, ranging from no bias with $\epsilon = 2$—agents accept and adapt to any social influence (from $x_j = -1$ to $x_j = 1$)—to maximum bias with $\epsilon = 0$—agents tolerate no deviation from their own opinion and never update it. A real-world example of the update rule in equation 1.2 with moderate bias is when a meat enthusiast, with $x_i = -1$, rejects any social influence from a vegan enthusiast with $x_j = 1$ but may reduce her enthusiasm when she talks to a more moderate meat eater with $x_j = -0.6$. Biased updating can similarly be implemented in the Bayesian updating heuristic (equation 1.1) by 'flattening' the perceived likelihood, thus making the transmitted information less conclusive:

$$x_i \leftarrow \frac{1}{N} \cdot x_i \cdot (\alpha \cdot x_j + (1-\alpha) \cdot \mathcal{U}) \quad \text{where } x_i \text{ and } x_j \text{ are distributions over } [-1,1]. \tag{1.3}$$

Here, the bias strength is $1 - \alpha$ with $\alpha \in [0,1]$ and the case $\alpha = 0$ implies that the perceived message reduces to $\mathcal{U}$, the fully uninformative flat distribution over the opinion space $[-1,1]$. Cognitive biases in the perception of others' opinions may also involve social characteristics unrelated to opinion similarity, such as social identities (see, for example, Alizadeh et al., 2015; Carpentras et al., 2022, and the model in Chapter 3), although this has been much less studied. For example, a supporter of the football club Werder Bremen will be more influenced by a Bremen player becoming vegan than if the same player played for Hamburger SV (Bremen's rival club), even though supporting a football club and dietary choice are not obviously related. Since I cannot do justice to the rich literature on updating rules and implementations of biases in opinion formation models in this thesis—notable alternative updating rule have been proposed in Axelrod (1997), Dalege et al. (2023), Dandekar et al. (2013), and Geschke et al. (2019), for example—I recommend the reviews by Castellano et al. (2009), Acemoglu and Ozdaglar (2011), Proskurnikov and Tempo (2018), Flache et al. (2017), Noorazar (2020), or Sobkowicz (2020).

1.4.5 Model analysis

To sum up, opinion formation modelling is a very broad field which allows for a variety of ways to implement and formalise theoretical assumptions into an agent-based, mathematical framework, including (i) how the opinions of the agents are represented, (ii) how the societal structure is generated, (iii) how the agents exchange information, and (iv) how they update their opinions following social interaction. Model simulations allow to track the opinions of all agents over time but, usually, modellers are interested in a specific output variable. In this thesis, I focus mainly on the distinction between consensus and disagreement and investigate how assumptions relating to individuals affect the chance of reaching these collective outcomes. Consensus means that the opinions of the agents virtually coincide. This can be measured, for example, as the dispersion, spread, or modality of the opinion distribution, although each of these measures highlights different aspects of consensus/disagreement (Bramson et al., 2017). The model dynamics typically depend on the initialisation of opinions and network topology, the sequence of updates, and any other stochastic factors included in the processes, such as noise in the communication (Carro et al., 2013; Sobkowicz, 2020; Turner & Smaldino, 2018). Hence, to generalise the findings and to test whether the generated patterns are robust, models need to be run multiple times as ensemble simulations and the results need to be aggregated.

1.5 History of opinion formation modelling and its frontiers

Opinion formation modelling has developed quite rapidly over the past decades. In his review of the research field, Sobkowicz (2020) characterised three phases of opinion formation models distinguished by their purposes and perspectives. Below, I describe this categorisation and use it to embed the chapters of this thesis into the broader context. The first generation of opinion formation models, which Sobkowicz calls 'physics of unphysical systems' (or 'socio-physics'), deals mainly with finding analogies between social and natural processes or objects, such as the spin–opinion analogy (see Castellano et al., 2009, for a review). The focus of these studies lies predominantly on analysing how the models themselves behave. The models typically comprise relatively simple representations of opinions (as binary or continuous values) and update rules; their goal is to explore how polarisation and consensus emerge in such 'minimal' settings (Abelson, 1967). For example, many early models investigated the conditions of consensus formation assuming that all agents influence each other in a homogeneously assimilative way (e.g. Degroot, 1974; French, 1956; Friedkin & Johnsen, 1990). The most famous examples of models in the socio-physics phase that were used to uncover the conditions of polarisation are bounded confidence models (Deffuant et al., 2000; Hegselmann & Krause, 2002), which are based on the update rule in equation 1.2, and the numerous subsequent modelling studies these models inspired (see Liu et al., 2023, for a recent review). However, with the purpose of illustrating the complex dynamics in the model rather than in realistic human societies, the 'toy models' (Elsenbroich & Polhill, 2023) of this first generation attracted only limited attention among social scientists. The second generation of opinion formation models—'enhanced models' (Sobkowicz, 2020)—aimed at a closer link between the mathematical models and the social sciences. Such enhanced models accounted, for example, for different roles of the agents (stubborn or contrarian agents or media), multi-dimensionality of opinions, repulsive social influence, or emotions (see Noorazar, 2020, for an overview of the milestones in the field). The focus of these studies lies more on representing the complexity of the social and psychological human systems rather than the model behaviour itself, sometimes at the cost of being able to perform extensive and rigorous analysis of all the relevant processes involved or being able to provide very intuitive explanations of the phenomena observed.

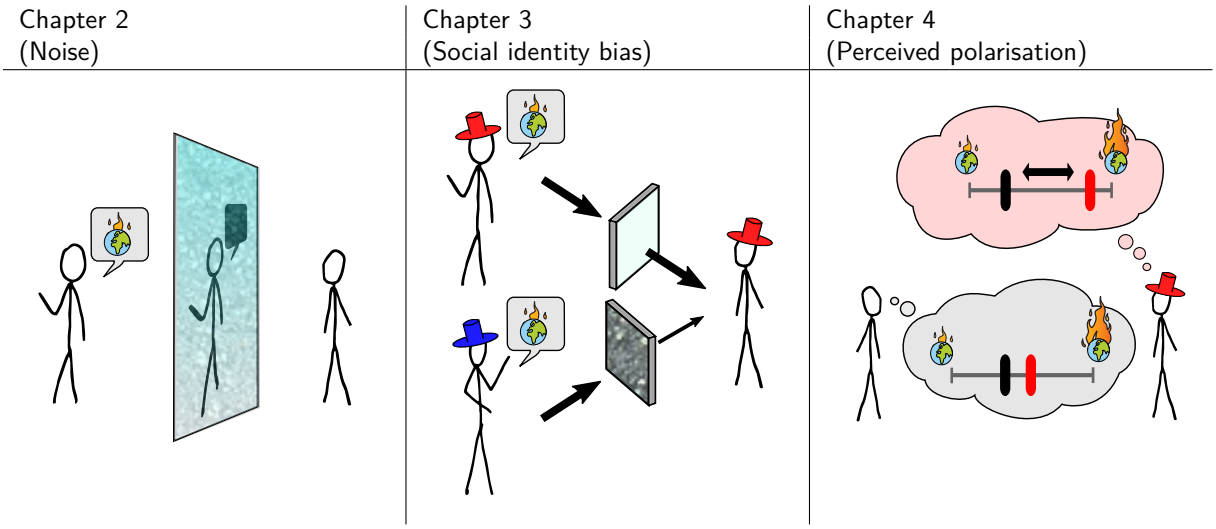


Figure 1.2: Sketches of the themes covered by the three main chapters: noise in communication (Chapter 2), social identity (in-group) bias (Chapter 3), and perceived polarisation (Chapter 4).

According to Flache et al. (2017), the major challenges for future work with current models of opinion dynamics fall into two categories: (i) comparisons between the many alternative implementations of the same theoretical assumptions (e.g. Chapter 2) and between the many models of the same phenomena with alternative underlying theories (e.g. Flache & Mäs, 2008) and (ii) an integration of empirical data to initialise the models (e.g. Chapter 2), support micro-assumptions (e.g. Carpentras et al., 2022), and validate macro-outcomes (e.g. Banisch & Shamon, 2024). Sobkowicz (2020) adds a third frontier: the next-generation of opinion formation models should address 'post-truth' aspects of the social world. Previous models typically assumed that agents have accurate opinions, express such opinions truthfully, change their opinions based on heuristic rules irrespective of underlying socio-political motivations, and have accurate knowledge of others' opinions. In contrast, 'post-truth' models should capture, for example, that (i) opinions are ambiguous concepts, (ii) expressed opinions differ from actual opinions, (iii) agents are subject to motivated reasoning, for example they may disregard personal opinions to maintain their status in a group, (iv) opinion patterns are transient and may never reach equilibrium, and (v) many social influences are manipulation attempts, such as trolling, fake news campaigns, and algorithmic filtering. Sobkowicz describes modelling guidelines to incorporate these aspects. The ones most relevant for this thesis are that models should (i) account for erroneous perceptions of the opinions of others and 'erroneous' expression of opinions, (ii) connect opinions to (ideologically driven) belief systems, and (iii) capture politically motivated reasoning in the formation and perception of opinions. 'Post-truth' aspects touch on the central theme of this thesis—that the way people perceive others and their opinions is subjective and distorted because cognitive biases and noise act as perception filters. This is reflected in each of the three chapters.

1.6 Outline of the thesis

The goal of this thesis is to better understand the formation of opinion patterns in social environments. I focus especially on opinion formation through social influence, given its importance for the topic of climate change. This thesis comprises three manuscripts, each exploring a particular factor, such as noise or cognitive

biases, that distorts peoples' perceptions (Figure 1.2). I investigate how these distortions affect the persistence of disagreement or even cause polarisation and I focus particularly on the conditions under which they foster consensus.

Chapter 2, *Noise and opinion dynamics: How ambiguity promotes pro-majority consensus in the presence of confirmation bias*, focuses on the role of random distortions—noise—in consensus formation. There are many different ways to implement noise and each carries a different meaning. While many models, especially in the 'socio-physics' generation (Sobkowicz, 2020), have studied how noise affects the model behaviour, these studies did often not distinguish between different types of noise and remained agnostic about their real-world interpretations. Moreover, one type of noise has been largely neglected: ambiguity in socially-transmitted message (Figure 1.2A). Ambiguity is inherent to communication and it is particularly important for a hotly debated and complicated issue like climate change. In this chapter, I develop an opinion formation model with different types of noise and confirmation bias, which is represented as bounded confidence. To initialise the model with realistic opinion configurations concerning climate change, I use empirical data about the relative size of opinion segments in the US population—the 'six Americas' (Maibach et al., 2011), ranging from a small dismissive segment to an alarmed one. The main research questions are: which combinations of bias and noise facilitate societal agreement that preserves (or even increases) the pro-climate majority opinion at the same time? Does the answer to this question depend on the type of noise? Therefore, one of the contributions of this study is to provide a systematic comparison of different types of noise, i.e. different implementations of a previously abstract and vague theory, and their interplay with the bias in a rather simple model of opinion formation. The study, thus, stands in the tradition of Sobkowicz's (2020) category of first-generation opinion formation models, but addresses the frontiers proposed by Flache et al. (2017)—comparing different theories and implementations of noise—and Sobkowicz (2020)—accounting for the inherent inaccuracy of expressed and perceived messages.

Chapter 3, *Social identity bias and communication network clustering interact to shape patterns of opinion dynamics*, focuses on the role of social identity in opinion formation. Despite the relevance of identity-driven processes and biases for political topics like climate change (Bliuc et al., 2015; Pearson & Schuldt, 2018), this aspect has been rarely addressed in opinion models in the literature, including the 'enhanced models' (Sobkowicz, 2020). Given the special role of generational identities in the climate change debate, I implement a virtual society composed of youngsters and elders[5]. This generational identity has two effects: (i) homophily in the interaction network, such that youngsters interact more with youngsters and vice versa, and (ii) in-group bias in the perception of individuals. I define opinions as distributions in this model, which indicate opinion direction, strength, and uncertainty at the same time. Although communication is not affected by noise in this model, the way individuals perceive others is distorted through in-group bias. This in-group bias acts as a filter, such that youngsters perceive the opinions of fellow youngsters as more certain or conclusive than the opinions of elders (see equation 1.3 and Figure 1.2B). This aspect addresses directly one of the principles for 'post-truth' models suggested by Sobkowicz (2020): perception is inaccurate and undermined by an external and independent motivation of the individuals to attend to social identities in addition to opinions during communication. One might intuitively assume that a stronger bias in a homophilic society should impede consensus formation as the mechanisms may reinforce the differences between groups, but the conditions under which this intuition holds are not entirely clear. Hence, I explore the following research questions: what are the conditions under which stronger in-group bias impedes consensus? Are there other settings (such as

[5]The model is general and could be similarly applied to any other political debate and a related distinction of social identity groups.

a different network structure) under which in-group bias promotes rather than prevents consensus formation? I furthermore illustrate the mechanism that drives such counterintuitive effects.

Chapter 4, *Perceived and actual opinion polarisation in the German climate change debate*, differs from the traditional opinion formation models presented in Chapters 2 and 3. This study should be viewed as being at the heart of the transition from classic opinion dynamics modelling—wherein artificial societies are generated based on theoretical assumptions—to a more computational social science approach—where empirical, rather than simulated, data are the main object of study, but modelling (and in particular, formalising 'post-truth' aspects, such as erroneous perceptions) helps to draw insights from the data. The study aims to measure ideological polarisation, but it illustrates that such a measure actually requires an underlying model of individual perception. This is because different assumption about individual perception lead to quite different conclusions about perceived and actual ideological polarisation in a society. Using data from the European Social Survey, I extract the distribution of opinions on climate change among Germans in 2016/17 and 2021. Such data could be (and traditionally was) used directly to measure actual opinion polarisation, for example, by studying whether the dispersion of opinions has increased from 2016 to 2021. In this study, I assume instead that the way people perceive the opinions of others, not their actual opinions, determines the level of polarisation they experience. Similar to the study in Chapter 3, I assume further that this perception of opinions depends on the individuals' political or partisan identities and related in-group biases. Based on these principles, I develop a method to measure perceived polarisation, wherein the individuals' perceptions co-evolve with their opinions. Applying this method to the empirical data, I address the following research questions: do Germans perceive more polarisation or less polarisation than their actual climate-change opinions suggest? How does this mismatch between perceived and actual opinion polarisation depend on the level of in-group bias and is it consistent across different partisan groups?

Chapter 2

Noise and opinion dynamics: How ambiguity promotes pro-majority consensus in the presence of confirmation bias

Peter Steiglechner[1,2,*], Marijn A. Keijzer[3], Paul E. Smaldino[4,5], Deyshawn Moser[1,2], Agostino Merico[1,2]

[1]Leibniz Centre for Tropical Marine Research (ZMT), Bremen, Germany

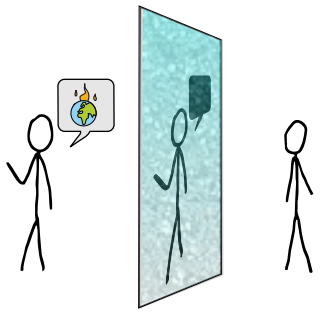
[2]Constructor University, Bremen, Germany

[3]Institute for Advanced Study in Toulouse, Toulouse School of Economics, Toulouse, France

[4]Department of Cognitive and Information Sciences, University of California Merced, Merced, USA

[5]Santa Fe Institute, Santa Fe, USA

[*] *Corresponding author:* peter.steiglechner@leibniz-zmt.de



This chapter contains a manuscript published as

Steiglechner, P., Keijzer, M. A., Smaldino, P. E., Moser, D., & Merico, A. (2024). Noise and opinion dynamics: How ambiguity promotes pro-majority consensus in the presence of confirmation bias. *Royal Society Open Science*, *11*(4), 231071. https://doi.org/10.1098/rsos.231071.

Abstract

Opinion dynamics are affected by cognitive biases and noise. While mathematical models have focused extensively on biases, we still know surprisingly little about how noise shapes opinion patterns. Here, we use an agent-based opinion dynamics model to investigate the interplay between confirmation bias—represented as bounded confidence—and different types of noise. After analysing where noise can enter social interaction, we propose a type of noise that has not been discussed so far, ambiguity noise. While previously considered types of noise acted on agents either before, after or independent of social interaction, ambiguity noise acts on communicated messages, assuming that socially transmitted opinions are inherently noisy. We find that noise can induce agreement when confirmation bias is moderate, but different types of noise require quite different conditions for this effect to occur. An application of our model to the climate change debate shows that at just the right mix of confirmation bias and ambiguity noise, opinions tend to converge to high levels of climate change concern. This result is not observed with the other types. Our findings highlight the importance of considering and distinguishing between the various types of noise and the unique role of ambiguity in opinion formation.

Keywords: *Agent-Based Model, Opinion Formation, Noise, Bounded Confidence, Climate Change, Computational Social Science*

2.1 Introduction

Social influence plays a critical role in how people form opinions (Cialdini & Goldstein, 2004; Kendal et al., 2018; Latané, 1981), including important topics like climate change, public health measures (like wearing face masks to prevent the spread of COVID-19) or government-funded social measures. However, social influence entails more than high-fidelity transmission of information; it is affected by cognitive biases and by noise. The term 'cognitive biases' refers to systematic deviations from classically rational use of information[1]. For example, confirmation bias is a well-documented phenomenon where people disproportionately select the information that confirms their prior beliefs (Lord et al., 1979). The role of biases in opinion formation has been studied extensively and is typically associated with impeding the convergence of opinions (although, under certain conditions, biases may have the opposite effect, e.g. O'Connor & Weatherall, 2018; Steiglechner et al., 2023). Noise represents unsystematic, random factors in the opinion formation processes. Although well studied (Macy & Tsvetkova, 2015; Mäs et al., 2010; Pineda et al., 2009), the role of noise in social influence remains less obvious than the role of biases and has triggered controversies (e.g. the recent exchange between Kahnemann *et al.* and Krakauer *et al.* in Kahneman et al., 2022).

To study what drives the emergence of certain opinion patterns in society, mathematical modelling and, in particular, agent-based modelling have become vital tools (refer to Refs. Flache et al., 2017; Noorazar, 2020; Sobkowicz, 2020, for recent reviews of this field). Models of social influence typically account for cognitive biases by assuming, for instance, that agents perceive information through a biased filter (Sobkowicz, 2018),

[1]though refer to Gigerenzer et al. (2012) for a perspective on why so-called biases may be adaptive responses to uncertainty

that agents make systematic errors when they adapt to social influence (Dandekar et al., 2013; Deffuant et al., 2023) or that agents ignore dissonant opinions (Deffuant et al., 2000; Hegselmann & Krause, 2002). By including biases in the social influence process, the models generate virtual societies in which either disagreement persists or consensus can form (Flache et al., 2017).

Noise is a source of stochasticity in an otherwise deterministic opinion-formation process. If models do not take noise into account, they run the risk of misrepresenting real phenomena on both individual and collective levels (Macy & Tsvetkova, 2015). For example, many models assume that agents always apply the same, predefined heuristics during social interactions. In reality, when interacting with others, people apply a diverse set of social learning strategies depending on the context (Kendal et al., 2018). Models without noise also assume that agents have accurate representations of their own or others' opinions, but in reality, people often construct an opinion on the spot (Galesic et al., 2021; Vul & Pashler, 2008), making social interaction inherently noisy. Moreover, most models that ignore noise produce patterns in which both individual opinions and macro-patterns cease to change when the virtual society has reached an equilibrium. In reality, individual opinions remain dynamic, and although public opinion distributions are often quite stable (Druckman & Leeper, 2012), we do observe drifting or fission–fusion dynamics in real life, for example in the climate change debate (Pew Research Center, 2022). Such macro-dynamics contradict the convergence towards a stable equilibrium obtained in computational models without noise.

Noise can capture a variety of real-world processes, which can all be formalised in different ways. Here, we provide an overview of how different formalisations of noise affect the dynamics of opinion formation. We survey the literature and present a taxonomy of four types of noise underlying the social influence process: (1) selectivity noise, (2) adaptation noise, (3) exogenous noise, and (4) ambiguity noise. The former three have been the object of study in previous models of opinion formation (see section 2.3 for a brief review of the literature on each type of noise), but have not been subject to rigorous comparison. We present this comparison, conceptually and numerically using computational experiments. Previous studies considered noise as acting either on the connection between two agents by varying who tolerates whom as interaction partner (selectivity noise), on the receiver of a message by changing how the opinion of the receiver changes after its influence-response (adaptation noise) or on any agent in general by occasionally perturbing its opinion regardless of social interaction (exogenous noise). The fourth and novel type of noise—ambiguity noise—acts on the communicated message, reflecting the idea that socially transmitted information is often ambiguous by nature. This idea relates to prominent work in psychology (Chater et al., 2020; Galesic et al., 2021; Vul & Pashler, 2008), economics (Kahneman et al., 2021), communication sciences (Eisenberg, 1984; Frankenhuis et al., 2023) or sociology (McMahan & Evans, 2018). The concept of ambiguity noise has appeared previously, for example in a model on interpersonal relations (Deffuant et al., 2013), but its effects on consensus formation in the presence of biased assimilation have not been studied.

Ambiguity noise is a general feature of the social influence process, and we argue that this type of noise is particularly relevant in debates on topics about complex matters where there is a ground truth or objectively superior opinion, such as in the debate on the anthropogenic origin of climate change. Social influence is a critical factor in this debate as people signal their opinions to others—often unconsciously—for example by openly indicating support for public figures on social media or by making everyday decisions, such as dietary choices, which are visible to others (Pearson et al., 2016; Toomey, 2023). When people express opinions in such ways, the signals they send are inherently ambiguous and uncertain and thus prone to create misunderstandings and interpretation errors, especially when dealing with a complex topic like climate change (Pearson & Schuldt, 2018). For example, empirical studies showed that when people were asked about their perception of climate

change, the weather or the season at the time of the interview affected their answers and thus concealed their true opinion on climate change to some extent (Borick & Rabe, 2014; Howe et al., 2019). It is not obvious how ambiguity noise affects opinion patterns in such debates. In fact, as we show later, it can both promote or impede consensus depending on the level of noise. Our study, therefore, aims to clarify the role played by ambiguity noise in opinion dynamics models, a largely overlooked aspect.

Several modelling studies have investigated the impact of noise on opinion dynamics and some of them also considered multiple types of noise e.g. Flache and Macy, 2011; Grauwin and Jensen, 2012. A common feature of these studies is that many prominent findings in social influence models are not robust to the inclusion of noise. For example, the emergence of opinion clustering in societies with biased agents does not occur under even very small levels of noise De Sanctis and Galla, 2009; Klemm et al., 2003a; Kurahashi-Nakamura et al., 2016; Macy and Tsvetkova, 2015; Mäs et al., 2010, despite being a common result in models without noise. However, there has been little effort to systematically and comprehensively compare different types of noise and their effects on the opinion patterns emerging from noisy social influence. Such system comparisons are useful; for example, a recent study compared four types of uncertainty on the evolution of social learning and found systematic differences in how each type influenced evolutionary dynamics (Turner et al., 2023). Similarly, a study by Grauwin and Jensen (Grauwin & Jensen, 2012) investigated two types of noise—exogenous and selectivity noise—finding that they lead to opposite dynamics and interact in non-trivial ways. Although the authors included these two different types of noise, the focus of their work was not to systematically study the interplay of bias and different noises. Our study aims to provide such a systematic comparison of the four different types of noise and to identify their effects on collective opinion patterns in social influence models.

In this study, we present an agent-based model of opinion formation in which the agents are affected by bias and noise. We focus on confirmation bias as one of the most important cognitive biases affecting opinion formation (especially in the debate on climate change (Johnson & Levin, 2009; Moser et al., 2022)). We extend the popular model by Deffuant and Weisbuch (Deffuant et al., 2000) to represent this bias as bounded confidence (BC). We present a taxonomy of types of noise, provide implementations in the framework of the BC model and study the dynamics that these types of noise produce for different degrees of bias. We show that noise in this model generally bolsters agreement when the agents are moderately biased. But different types of noise induce quite different patterns of agreement and disagreement. Finally, we apply the model to the debate on climate change by calibrating agents' opinions to survey data (Maibach et al., 2011). As a majority of the agents hold pro-environmental attitudes at the outset, one would intuitively expect that unambiguous communication and a high degree of openness to other opinions promote pro-environmental consensus. Surprisingly, we find that a mix of ambiguity noise and confirmation bias provides the best conditions to foster convergence on pro-environmental attitudes. This result is unique to ambiguity noise and does not occur with other types of noise.

2.2 Modelling confirmation bias

Confirmation bias reflects the tendency of people to disproportionately attend to information that confirms their prior beliefs. The BC approach (Deffuant et al., 2000; Hegselmann & Krause, 2002) is arguably the most popular in representing this bias in formal models[2]. Here, we adopt a version of the BC model by Deffuant

[2]Other researchers have modelled confirmation bias in the context of collective problem solving using quite different modelling frameworks, but have nevertheless operationalised it similarly as a preference for solutions close to the individual's current solution (Boroomand & Smaldino, 2023; Gabriel & O'Connor, 2024).

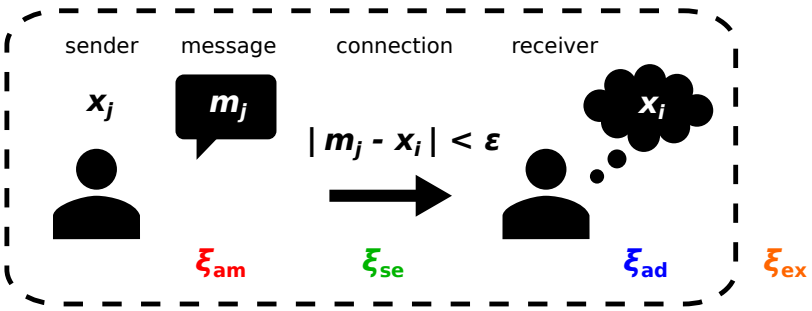


Figure 2.1: Different types of noise and how they act inside or outside the social interaction process: (i) ambiguity noise, ξ_{am}, acts on the message from a sender, (ii) selectivity noise, ξ_{se}, affects whether a receiver is chosen for interaction in light of the difference between message and receiver opinion and the confidence bound, (iii) adaptation noise, ξ_{ad}, affects the receiver's opinion after an interaction, and (iv) exogenous noise, ξ_{ex}, perturbs an agent opinion from outside of the social interaction.

et al. (2000) that considers pairwise interactions and asynchronous updating. In this formulation of the BC model, n agents represent human beings who form their opinions through one-to-one interaction. The opinion of an agent i is represented as a scalar value $x_i \in [0, 1]$, denoting, for example, the agent's concern about climate change, ranging from extremely dismissive ($x_i = 0$) to extremely alarmed ($x_i = 1$). When the two agents i and j interact, their opinions x_i and x_j simultaneously align unless the distance between them exceeds the agents' confidence bounds:

$$x_i \mapsto \begin{cases} x_i + \mu \cdot (m_j - x_i) & \text{if } |m_j - x_i| \leq \epsilon \\ x_i & \text{else} \end{cases} \text{ with the message } m_j = x_j \ , \tag{2.1}$$

where μ is the speed at which opinions converge (Deffuant et al., 2000) and ϵ is the confidence bound. A large confidence bound represents a low bias (and vice versa). The message, m_j, is the perfectly accurate representation of the opinion x_j of agent j.

2.3 Modelling noise

2.3.1 Previously studied types of noise

In this section, we present the types of noise that have been previously studied: selectivity, adaptation and exogenous noise. While the precise formalisation may differ from model to model, we distinguish these three categories because they describe three distinct phases in the social interaction where noise may enter the system. Figure 2.1 visually represents where these types of noise enter the system (as well as where ambiguity noise, described in the next section, is involved).

Selectivity noise acts on the connection between agents. It stochastically induces interactions between agents that would typically ignore each other or inhibits interactions between agents that would typically influence each other positively e.g. De Sanctis and Galla, 2009; Deffuant, 2006; Flache and Macy, 2011; Grauwin and Jensen, 2012; Holyst et al., 2001; Kowalska-Styczeń and Malarz, 2020; Kurahashi-Nakamura

et al., 2016; Mäs et al., 2010. Selectivity noise is sometimes seen as a property of the social system and, in analogy to thermodynamics, referred to as 'social temperature'[3] (Holyst et al., 2001). We implement selectivity noise by adding random (zero-mean Gaussian) fluctuations, ξ_{se}, to the confidence bound, ϵ, of an agent in every interaction with another agent:

$$x_i \mapsto \begin{cases} x_i + \mu \cdot (x_j - x_i) & \text{if } |x_j - x_i| \leq \epsilon \ + \ \xi_{\mathrm{se}} \\ x_i & \text{else} \end{cases} . \tag{2.2}$$

This reflects the fact that even though confirmation bias will normally select information within a certain confidence bound, individuals are sometimes confronted with and influenced by information outside of their confidence bound. It refers to imperfections or errors in the selection of suitable communication partners.

Adaptation noise acts on a receiver by modifying how its opinion changes after a social interaction e.g. Baccelli et al., 2017; Mäs et al., 2010; Su et al., 2017; Turner and Smaldino, 2018; Zhang and Zhao, 2018. We implement adaptation noise by additionally shifting the agent opinion by a random (zero-mean Gaussian) amount, ξ_{ad}, whenever an agent updates its opinion following a successful interaction with another agent, that is when their prior opinions are within the confidence bound:

$$x_i \mapsto \begin{cases} x_i + \mu \cdot (x_j - x_i) \ + \ \xi_{\mathrm{ad}} & \text{if } |x_j - x_i| \leq \epsilon \\ x_i & \text{else} \end{cases} . \tag{2.3}$$

Adaptation noise thus describes changes in opinions that are triggered by a change of mind after an interaction and are unrelated to the transmitted message itself[4]. A socially transmitted message might trigger an individual to autonomously think more about a given topic or may lead to a search for more information. These types of perturbations after a social interaction are captured by adaptation noise.

Exogenous noise acts on opinions in a process entirely separate from the social interactions e.g. Carro et al., 2013; Klemm et al., 2003b; Kurahashi-Nakamura et al., 2016; Maciel and Martins, 2020; Nyczka, 2011; Nyczka et al., 2012; Pineda et al., 2011; Schweighofer et al., 2020; Stern and Livan, 2021; Vieira and Crokidakis, 2016; Zhao et al., 2016. It may capture an individual's autonomous development of thought or insights gained from information obtained independent from social interaction. We implement exogenous noise by shifting the agent opinion by a random (zero-mean Gaussian) amount, ξ_{ex}, (an 'opinion jump') with some small probability ω:

[3] The temperature analogy refers to higher chances of interaction between random particles in hot versus cold systems.

[4] For instance, adaptation noise may lead to the situation where an agent is positively influenced by another agent (because their opinions are within the confidence bound), yet ends up disagreeing more with that agent after their interaction. Assume $x_j > x_i$ and $x_j - x_i < \epsilon$. Although x_i should increase owing to the positive social influence, it may decrease if $\xi_{\mathrm{ad}} \ll 0$. For example, when two individuals already hold very similar opinions, a discussion could lead them to further converge. But if this conversation prompts them to reconsider certain arguments differently, they might end up disagreeing more than they did before their interaction.

$$x_i \mapsto x_i \; + \begin{cases} \xi_{\mathrm{ex}} & \text{with probability } \omega \\ 0 & \text{else} \end{cases} . \tag{2.4}$$

In all of these implementations, which are conceptually in line with the correspondingly cited model literature, we draw the deviations, ξ_{ex}, ξ_{se} and ξ_{ad}, from a Gaussian distribution $\mathcal{N}(0, \nu)$, where $\nu > 0$ defines the fixed level of noise. For exogenous noise, we define $\omega = \nu$ such that the frequency and amplitude of the noisy opinion perturbations are coupled. We ensure that opinions remain within the opinion space, $0 \leq x_i \leq 1$, by truncating the Gaussian noise distribution accordingly. In particular, for exogenous and adaptation noise, we resample the noise, ξ, if the posterior opinion x_i including this noise would fall outside the opinion space. Note that there are various names for these types of noise in the literature (e.g. what we call 'adaptation noise' has been referred to as 'communication noise' (Turner & Smaldino, 2018) or 'individualisation' (Mäs et al., 2010) elsewhere and 'exogenous noise' has been referred to as 'perturbation noise' (Klemm et al., 2003b)). Here, we use names that reflect the different stages of the social interaction process in which noise plays out.

2.3.2 Ambiguity noise

We propose ambiguity in expressed messages as a fourth type of noise. Ambiguity is deeply linked to the conceptualisation of opinions. We assume that opinions are fundamentally uncertain estimates of how one should respond in particular scenarios, such as how a person's concern about climate change may or may not lead to social action depending on how the individual's uncertainty about the opinion interacts with perceptions of the likely costs and benefits of action. It follows that expressed opinions are likely to be ambiguous signals and, as such, cannot be deterministically quantified on a numeric scale. This can lead to ambiguity on the part of a receiver in terms of how to interpret received information. Another source of ambiguity in socially transmitted messages is that people often construct an opinion on the spot rather than holding pre-defined opinions (Galesic et al., 2021). When asked repeatedly, a person may thus produce different opinions on the same topic, a phenomenon called the 'crowd within' effect (Carpentras et al., 2022; Herzog & Hertwig, 2009; Vul & Pashler, 2008). This can lead to ambiguity on the part of a sender.

Ambiguity noise acts on the message conveying the opinion of the sender. We implement ambiguity in a message as a stochastic modification of its true value[5]. Specifically, agent i sees the opinion of agent j as a sample drawn from a Gaussian distribution centred around the opinion x_j and truncated at the bounds of the opinion space. The variance of this noise distribution, ν, determines the typical deviation, ξ_{am}, of a sender's message from its true opinion:

[5]We acknowledge that the term ambiguity is in itself ambiguous. Ambiguity might imply that recipients see multiple coherent interpretations of a socially transmitted message. Here, we define ambiguity noise in the following sense. Although agents receive a message as a single value, this value is coherent with multiple possible inferences of the true opinion of the sender. For readers objecting to our use of ambiguity, we propose substituting 'ambiguity noise' with 'message noise' or 'transmission noise'.

$$x_i \mapsto \begin{cases} x_i + \mu \cdot (m_j - x_i) & \text{if } |x_i - m_j| \leq \epsilon \\ x_i & \text{else} \end{cases} \tag{2.5}$$

with the message $m_j = x_j + \xi_{\text{am}}$ where $\xi_{\text{am}} \sim \mathcal{N}(\mu = 0, \sigma = \nu)$ s.t. $m_j \in [0, 1]$.

When agent i receives the noisy message, m_j, it is unaware of the actual value of x_j but nevertheless applies the update rule using the noisy message as the best available representation of x_j. When noise is negligible, $\nu \to 0$, agents see deterministic and accurate opinions (as in equation 2.1). Note that owing to the truncation of messages, an agent at the boundary with opinion $x_j = 1$, for example, can only send messages with $m_j \leq 1$. That is, in this extreme case, messages m_j are drawn from a half-normal distribution with an average message of $1 - \sqrt{2/\pi} \cdot \nu$ instead of x_j.

Ambiguity noise affects opinion formation directly and indirectly. Through its direct effect, ambiguity noise can temporarily drive agents apart even if their opinions are identical (similar to exogenous or adaptation noise) because the agents are unaware of their actual agreement. Through its indirect effect, ambiguity noise influences whether agents can successfully interact with each other or not (similar to selectivity noise). In particular, when ambiguity noise is strong, $\nu \gg 0$, even agents with very distant opinions may occasionally be positively influenced by each other's opinion when a message deviates so strongly from the sender's opinion that it creates the illusion for the receiver that the sender's true opinion is similar to its own. Ambiguity noise affects the message and, thereby, indirectly also the selection and adaptation of a receiver. In contrast, selectivity or adaptation noises affect the receiver directly and independently of each other. Ambiguity noise thus has a different real-world meaning than these previously studied types of noise and leads to non-trivial results that cannot be inferred by simply adding up selectivity and adaptation noise. We, thus, treat ambiguity as an independent source of noise.

2.3.3 Simulation experiments

We are interested in how ambiguity noise affects the emergence of agreement or disagreement in a virtual society and how the results compare to the other types of noise. As a metric for societal disagreement, we use the dispersion of the agents' opinions refer to Ref. Bramson et al., 2017, for an overview of measures of disagreement. Additionally, we also consider the kurtosis of the opinion distribution in Supplementary Figures A.1 and A.2 with similar qualitative results. The dispersion σ is calculated as the standard deviation of the opinion distribution, $\sigma\left(\{x_i|i\}\right)$, for all agents i at a specific time. A small dispersion indicates that the agents largely agree and the opinion distribution has a narrow (approximately unimodal) shape. A large dispersion indicates disagreement, which can reflect either (i) a polarised society with agents separating into multiple, but internally narrow opinion clusters or (ii) a diffused society with agents forming a single opinion cluster that is dispersed and incohesive.

We perform experiments by fixing the number of agents, $n = 100$, and the maximum speed of convergence, $\mu = 0.5$, and by varying the levels of confirmation bias, which is inversely related to the confidence bound, and noise. For simplicity, we assume that all agents share the same level of bias and noise. Noise is a feature of human behaviour that is independent of bias and, thus, we present our results in terms of the corresponding parameters ν and ϵ. One time step in the model consists of two agents being randomly paired (without

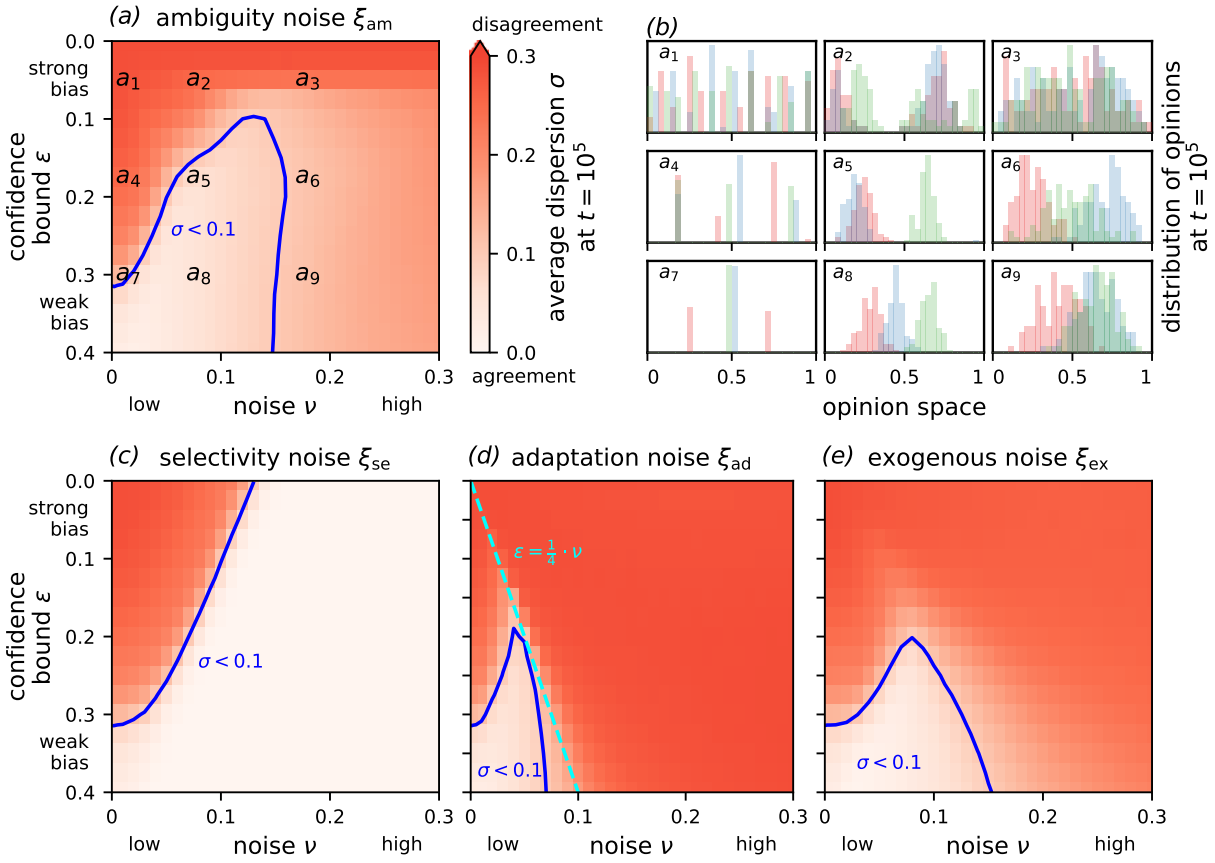


Figure 2.2: Dispersion, σ, averaged over 1000 stochastic simulations with $t = 10^5$ steps, under different levels of confirmation bias as a function of the confidence bound ϵ (y-axis) and noise ν (x-axis), including ambiguity noise ξ_{am} (*a*), selectivity noise ξ_{se} (*c*), adaptation noise ξ_{ad} (*d*) or exogenous noise ξ_{ex} (*e*). In each simulation, the initial opinions of the agents are drawn from a uniform distribution. Low dispersion (light red) indicates strong agreement and high dispersion (deep red) indicates disagreement. For comparison between the panels, the blue line denotes the noise-to-bias ratios for which, on average, $\sigma < 0.1$. The cyan line in panel *(d)* denotes the theoretical transition from agreement to disagreement under adaptation noise derived by Zhang and Zhao (2018). The subpanels in panel *(b)*, corresponding to locations a_1 to a_9 in panel *(a)*, show the distributions of the agents' opinions at $t = 10^5$ over the opinion space for three exemplary simulations (coloured histograms in panel *b*) under different combinations of confirmation bias ($\epsilon = 0.05$, 0.175 or 0.3) and ambiguity noise ($\nu = 10^{-10}$, 0.08 or 0.18).

network structure or systematic selection preferences) and potentially adapting their opinions. We terminate the simulation after $t = 10^5$ time steps, that is when a single agent has engaged, on average, in 2000 one-on-one interactions. In the first part of this study, we analyse hypothetical scenarios in which the agents have uniformly distributed initial opinions with a typical initial dispersion $\sigma(t = 0) = 0.29 \pm 0.01$. In the second part, we apply our model to opinion formation in the climate change debate, where opinions represent the concerns of agents about climate change and initial opinions are sampled from a distribution calibrated to survey data (Maibach et al., 2011).

2.4 Results

2.4.1 Ambiguity noise promotes agreement

Depending on the level of confirmation bias and noise, the opinions of the agents either converge (low dispersion) or remain in disagreement (high dispersion). When noise is negligible, $\nu \to 0$ (locations a_1, a_4, a_7 in figure 2.2*a* and the corresponding subpanels in figure 2.2*b*), agreement emerges only when bias is sufficiently weak, that is when the confidence bound $\epsilon \gtrsim 0.3$ (location a_7). The number of clusters in the BC model without noise is typically around $1/(2\epsilon)$. For a moderate bias, $\epsilon \in [0.1, 0.3]$, ambiguity noise induces agreement, with dispersion $\sigma \to 0$ (location a_5). This reverses when noise is very high (location a_6). A stronger bias in this regime creates more disagreement but can be compensated by stronger noise. When the bias exceeds a threshold, here $\epsilon \lesssim 0.075$, ambiguity noise is not sufficient to create agreement within the simulation time considered and the distribution of opinions remains diffuse with $\sigma \approx 0.3$ (locations a_2 and a_3). This pattern is qualitatively robust for a wide range of parameter choices such as a smaller/larger number of agents n, a shorter/longer simulation time t or a slower speed of convergence μ (see Supplementary Figures A.3–A.8). The range of bias levels within which noise induces agreement among agents widens with an increasing number of agents n or simulation time t. As such, the dispersion is somewhat sensitive to the model parameters in the critical configuration a_5 (see Supplementary Figures A.4, A.6 and A.8) but this does not affect the robustness of our main results. In particular, the existence of a transition between disagreement and agreement for moderate ambiguity noise and moderate bias is independent of the parameters reflecting system size, simulation time and convergence speed.

2.4.2 Ambiguity noise induces group drift

Even when agents are in full agreement about their opinions, the mean opinion of the society, $\overline{x}$, can still change owing to the ambiguity noise in communicated opinions. Opinion distributions can thus become multimodal even after convergence, and clusters of shared opinions can, theoretically, re-emerge. When bias is moderate, the agents even develop more extreme opinions (figure 2.2*b*, a_5 and a_6), that is, the average opinion tends to drift more towards the edges of the opinion space, compared to when bias is weak (a_7–a_9). Supplementary Figure A.9 shows that this drift is a robust feature of the model under moderate ambiguity noise and moderate bias. Drift can be explained by the balance of two opposing forces in the presence of ambiguity noise: repulsion from the bounds of the opinion space and attraction towards extremists. First, agents with extreme opinions are dragged away from the bounds of the opinion space. This is because the opinions of extreme agents can only be pulled in one direction: towards more moderate values. This happens in all variations of the BC model, independent of the type and degree of noise. Second, and acting as a counter force to the effect of extremists becoming more moderate, agents with extreme opinions in a given distribution tend to be more successful in pulling agents with moderate opinions closer to the bounds of the opinion space. Messages from agents with extreme opinions tend to be seen as more moderate (owing to the boundedness of the opinion space and the truncation of the noise distribution when drawing a 'noisy message') and, as a consequence, they are more likely to be accepted by moderate agents. In other words, the agents with relatively extreme opinions benefit from ambiguity, making them more successful in transmitting their messages to receivers than the agents with relatively moderate opinions (see Supplementary Figure A.10). As a consequence, the interplay of a moderate confirmation bias with moderate to high ambiguity noise can pull initially diverse and moderate societies towards more extreme average opinions (see also Supplementary Figure A.11, which

shows how the mean opinion of a Gaussian opinion distribution evolves for different levels of ambiguity noise depending on how extreme the initial opinions are). However, the drift dynamics are non-trivially dependent on the distribution of opinions and drift is a finite-size effect that disappears for larger societies, for example with $n = 1000$ fully connected agents (Supplementary Figure A.3*b*).

2.4.3 Different types of noise induce different opinion patterns

There are some similarities between the results obtained with ambiguity noise (figure 2.2*a*) and the other three types of noise—selectivity (2.2*c*), adaptation (2.2*d*) and exogenous noise (2.2*e*). In particular, noise can induce agreement for a range of bias levels and, within this range, more bias always requires higher noise to achieve agreement. This result depends on the type of noise considered, especially under relatively strong bias, $\epsilon \lesssim 0.2$. Selectivity noise (figure 2.2*c*) induces agreement over the full range of bias levels as long as noise is sufficiently high. However, increasing the level of adaptation and exogenous noise (figure 2.2*d* and 2.2*e*) beyond some threshold inhibits the emergence of agreement, causing an abrupt transition from a narrow opinion distribution to a broader one. Thus, reaching an agreement under a strong bias is inhibited regardless of noise. For example, for adaptation noise, agreements break when $\nu > 0.25 \cdot \epsilon$ (which was also analytically derived in Zhang & Zhao, 2018). That is, with a stronger bias (smaller ϵ), this transition is triggered at lower levels of noise. Agreement emerges only within a narrow range of low adaptation noise and this range shrinks as bias increases. In comparison to these three types of noise, ambiguity noise generates a different pattern (figure 2.2*a*): agreement emerges only within a range of moderate noise levels (similar to adaptation and exogenous noise), but this range does not shrink as bias increases somewhat (in contrast to adaptation and exogenous noise and more similar to selectivity noise) such that even a relatively strong bias, $0.2 \gtrsim \epsilon \gtrsim 0.1$, can be counteracted by ambiguity noise.

2.4.4 Bias and ambiguity noise foster pro-environmental agreement

For a realistic representation of opinions on climate change in society, we calibrate the initial opinions of the agents to empirical data of Maibach et al. (2011) (see Supplementary Section A.1 for details). A high value x_i reflects a strong concern about climate change (and vice versa). Following the six distinct categories obtained in the survey by Maibach et al., we define agents as 'concerned' if $x_i > 4/6$ and 'alarmed' if $x_i > 5/6$. The initial distribution (see subpanel $t = 0$ in figure 2.3*b*) is skewed towards high concerns, but there is also a significant fraction of dismissive or neutral agents. In the context of climate change, we are interested in the ability of a society to reach a pro-environmental agreement (PEA). We define this PEA as a state in which (i) the agents are, on average, at least concerned (if not alarmed) about climate change, $\overline{x} \geq 4/6$, and (ii) the agents largely agree on this level of concern, $\sigma \leq 0.1$. At $t = 0$, both requirements for a PEA are typically not fulfilled with $\overline{x}(t = 0) = 0.60\,(\pm 0.03) < 4/6$ and $\sigma(t = 0) = 0.25\,(\pm 0.01) > 0.1$.

Figure 2.3*a* shows the frequency of a society-wide PEA resulting from 1000 independent simulations for different bias and ambiguity noise levels. The requirement for reaching an agreement, $\sigma \leq 0.1$, prevents the formation of a PEA when (i) bias is strong (see locations a_1–a_3 in figure 2.3*a* and corresponding subpanels in figure 2.3*b*), (ii) bias is moderate and ambiguity noise is low (location a_4), and (iii) ambiguity noise is high (locations a_3, a_6 and a_9). The requirement for collectively being at least concerned about climate change, $\overline{x} > 4/6$, prevents the formation of a PEA when bias is weak (locations a_7–a_9). In particular, in societies with a weak bias, agents tend to reach an agreement, but this agreement comes at the cost of reduced

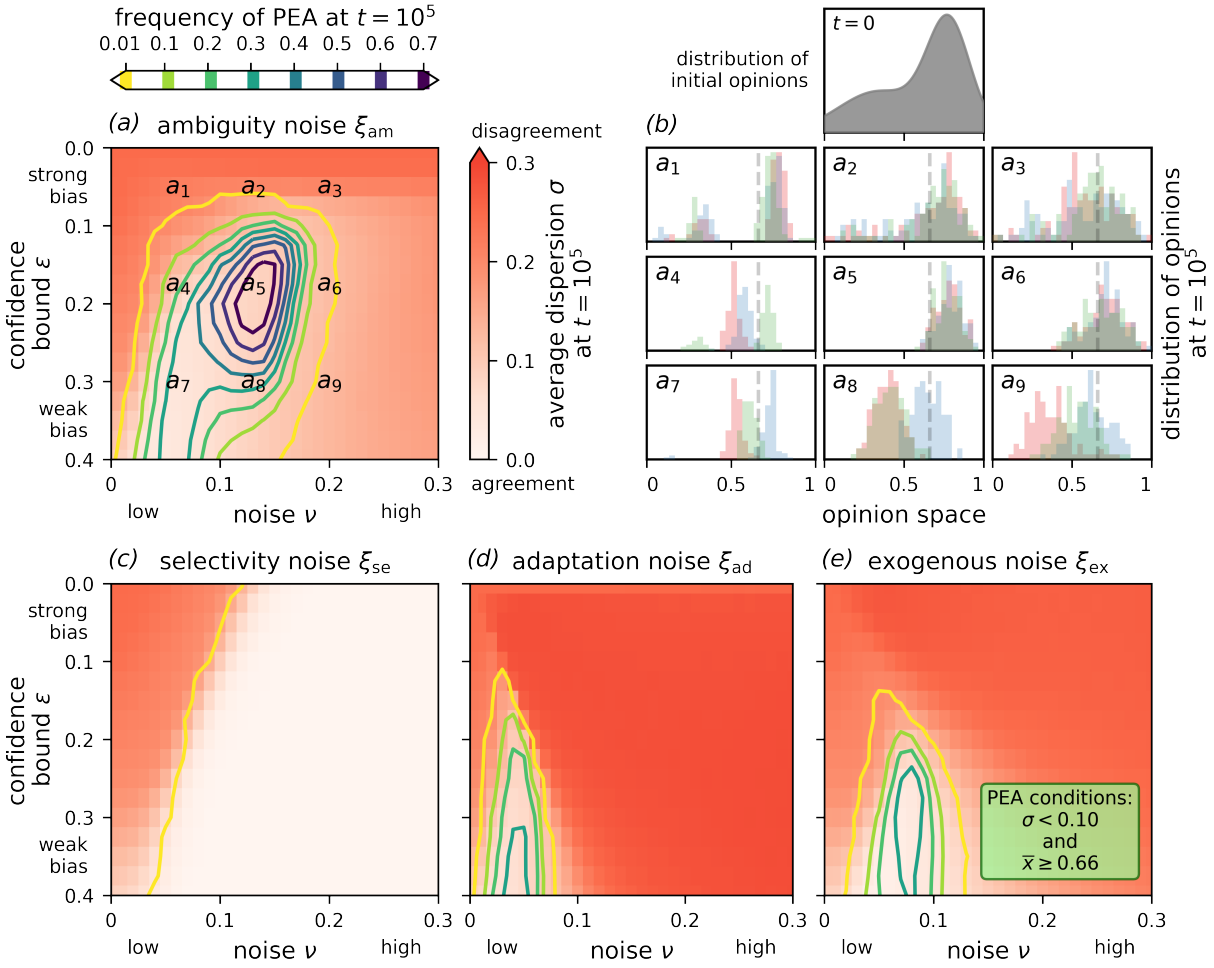


Figure 2.3: Frequencies of a 'pro-environmental agreement' (PEA) obtained from an ensemble of 1000 simulations initialised with a distribution of opinions reflecting the 'six Americas' (Maibach et al., 2011) (subpanel $t = 0$ in *b*) for different levels of noise and confirmation bias. PEA is a state of the model defined by low dispersion, $\sigma < 0.1$, and high average concern about climate change, $\bar{x} \geq 4/6$, at time $t = 10^5$. Panel *(a)* shows the results for ambiguity noise (as in figure 2.2) and, correspondingly, subpanels a_1–a_9 in *(b)* show example distributions of agent opinions at $t = 10^5$ for locations in *(a)*. PEA is rarely reached when bias is strong (locations a_1–a_3), noise is low (a_4) or noise is very high (a_3, a_6 and a_9). PEA is reached in less than $36\,\%$ of the simulations when bias is weak (a_7–a_9). However, PEA is frequently reached (in $77\,\%$ of the simulations) when both the bias and the ambiguity noise are moderate, $\epsilon \approx 0.175$ and $\nu \approx 0.13$ (location a_5). Under selectivity noise (*c*), adaptation noise (*d*) or exogenous noise (*e*), the maximum frequency of reaching PEA does not exceed $36\,\%$, regardless of the noise-to-bias ratio.

climate change concerns and, consequently, reduced frequency of PEA. There is an optimal noise-to-bias ratio, $\epsilon \approx 0.175$ and $\nu \approx 0.13$ (location a_5), for which the society reaches PEA with a much higher frequency ($77\,\%$ of the simulations). For a better understanding of the model behaviour, figure 2.4 shows how the opinions evolve in three example simulations of societies with moderate bias, $\epsilon = 0.13$, and ambiguity noise, $\nu = 0.08$, which represents a critical point where PEA is reached in some cases (figure 2.4*a*) but not in others owing to a lack of average concern (2.4*b*) or a lack of agreement (2.4*c*). Even in cases with general agreement among the agents, opinions still fluctuate (figure 2.4*a* and 2.4*b*). The average opinion, $\bar{x}$, however, remains quite extreme and stable, especially when ambiguity noise and bias are similar to the configuration a_5 in figure 2.3*a* (see Supplementary Figure A.12).

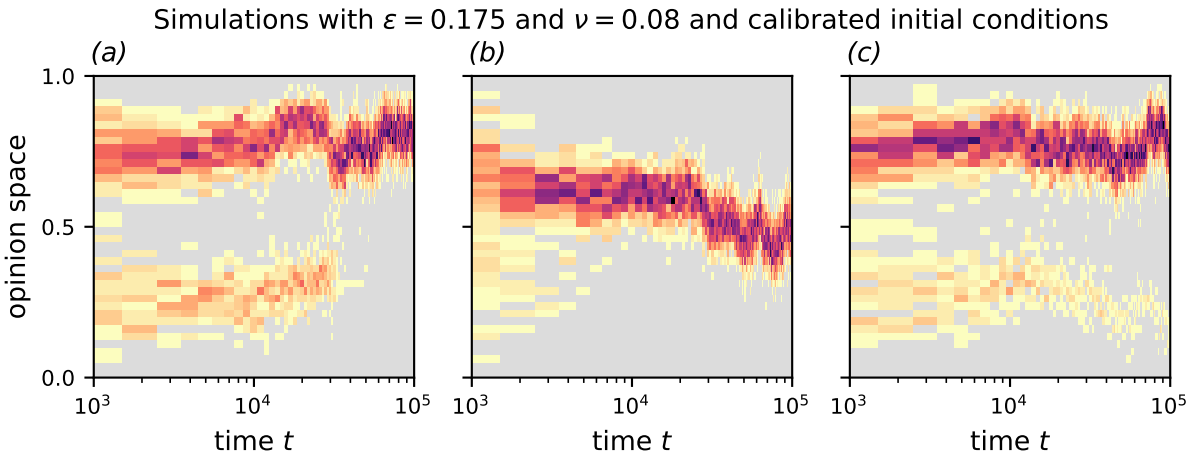


Figure 2.4: Typical opinion dynamics in simulations with calibrated initial conditions (figure 2.3) under moderate bias, $\epsilon = 0.175$, and moderate ambiguity noise, $\nu = 0.08$. This configuration lies in the critical region between a_4, where PEA is barely reached, and a_5, where PEA is very frequently reached. The colour denotes the agent density in each bin over time (note the logarithmic scale on the time axis). Dark red colours represent high density (i.e. many agents have opinions x_i falling in the corresponding bin at time t), bright yellow colours represent low density, and grey represents zero density. A consensus with very low dispersion can be reached relatively early (panel *a*), late (panel *b*) or never (within the simulated time horizon, panel *c*). In the latter case, some agents cluster in the lower half of the opinion space. The mean opinion at $t = 10^5$ in the simulations in panels *(a)* and *(c)* are above the threshold for a PEA, $\overline{x} \geq 0.66$ for at least the second half of the simulation, but the dispersion for the simulation in panel *(c)* is above the threshold for PEA, $\sigma = 0.1$, throughout the simulation and, therefore, only the simulation in panel *(a)* represents a PEA.

Interestingly, we observe this pattern only with ambiguity noise. In contrast, under selectivity noise (figure 2.3*c*), adaptation noise (figure 2.3*d*) or exogenous noise (figure 2.3*e*), the frequency of PEA remains below $36\,\%$, regardless of the levels of noise or bias. For instance, in the case of high selectivity noise, agents nearly always reach an agreement but only by compromising with the sceptical minority, and, thus, the consensus opinion is typically not concerned or alarmed, $\overline{x} < 0.6$. This implies that ambiguity noise in social influence likely promotes a scenario in which agents come to agree on a high level of climate change concern, whereas other types of noise do not have such an effect.

2.4.5 Robustness

In this section, we highlight several critical assumptions that are inherent to the results presented so far and briefly examine how altering these assumptions affects our results. First, in line with most other social-influence models, we assume one-to-one interactions in which a single focal agent assimilates to the view of a single sender. In reality, there are situations in which an individual considers the opinions of many others (*many-to-one* interaction, Flache & Macy, 2011), for example when the individual hears about election polls, or situations in which a single individual expresses an opinion publicly to many others simultaneously (*one-to-many* interaction, Keijzer et al., 2018), for example, when the individual posts statements on a social media platform. We find that the assumed communication regime has a pronounced effect on the opinion formation pattern in the presence of ambiguity noise. Under a many-to-one communication regime, the thresholds for drift and agreement resulting from ambiguity noise are much higher than in the one-to-one case. That is, high ambiguity noise is required to induce agreement, and a combination of strong bias and high ambiguity noise is required to obtain drifting. One-to-many communication, in contrast, lowers the threshold for reaching an agreement

and generates high-frequency drift. That is, even low levels of noise induce agreement regardless of bias, but the consensus opinion fluctuates across the opinion space. This represents a society in which all individuals share an opinion but, as a collective, they are highly volatile about this opinion. Exemplary simulation runs with these alternative communication regimes can be found in the Supplementary Figures A.13–A.16. In sum, while a many-to-one interaction may sometimes make disagreement more robust even under relatively high levels of noise, a one-to-many type of interaction might sometimes amplify the interference of noise.

Second, we assume that all agents are connected and thus able to directly influence each other. This assumption may hold for small groups, but social influence patterns are typically more characteristic of networks in which nodes represent agents and links represent influence channels between them (based on social relations), and several studies have shown the influence of network structure on opinion dynamics (Schawe et al., 2021; Squazzoni et al., 2014; Steiglechner et al., 2023). For example, it is well-known that the degree of transitivity (a defining characteristic of human social networks) can critically alter the dynamics of opinions. A thorough analysis of the effects of different network structures on the impact of the different types of noise would go beyond the scope of this study. Our exploratory analysis, however, confirms that the qualitative patterns obtained with a fully connected network are similar to those obtained with a transitive network, such as the standard small-world network (Watts & Strogatz, 1998), thus attesting to the robustness of the model dynamics to substantively different network topologies (Supplementary Figures A.17 and A.18).

Third, we assume a zero-mean Gaussian noise distribution—arguably the most plausible assumption one can make about stochastic variation in the real world. However, the shape of the noise distribution can have important implications. For example, Gaussian noise is unbounded and, as such, there is a small but non-zero chance that an agent with opinion $x_i = 1$ communicates a message $m_i = 0$ if $\nu > 0$. In Supplementary Figure A.19, we compare the simulations presented in figure 2.3 (in which we assume Gaussian noise and calibrated initial conditions) with simulations in which we assume that noise is instead drawn from a zero-mean, bounded uniform distribution, $\xi \in [-\nu, \nu]$, for the four types of noise (refer to Pineda et al., 2011, for a study using the same noise distribution). The results are nearly identical in most cases, although the optimal combination of ambiguity noise and confirmation bias produces much less PEA when noise is drawn from a bounded uniform distribution. Bounded uniform exogenous noise is especially effective in fostering PEA.

2.5 Discussion

We have presented a taxonomy of noise and provided insights into the surprisingly large differences obtained from implementations of these different types of noise in the BC model. In general, noise can induce agreement among moderately biased agents by smoothing out divisions between agents in different clusters. However, the conditions for this effect to occur depend on the type of noise considered. For example, selectivity noise—acting on the connection between agents—always promotes agreement, whereas adaptation and exogenous noise—acting on the agent after or independent of social influence—promote agreement only within a narrow range of bias and noise levels. These results are consistent with the previous modelling studies considering these types of noise (or similar conceptualisations of them) in different opinion dynamics models (see corresponding references in section 2.3). We have focused our analysis predominantly on ambiguity noise because its effects on consensus formation have been the least extensively addressed in prior models of social influence. By affecting communicated messages rather than their recipients, ambiguity noise has indirect consequences for the two relevant opinion formation processes in our model: selection and adaptation to social influence. It

seems intuitive to think that clarity and unbiased reasoning in public debates are key to enabling a society with initially diverse or polarised opinions to reach an agreement (McMahan & Evans, 2018; Smaldino & Turner, 2022). However, clear communication might be at odds with successful consensus formation, as indicated by conceptual work on ambiguity as a strategic tool (Eisenberg, 1984; Pinker et al., 2008). Our results show that moderate ambiguity in expressed opinions facilitates agreement under a wide range of bias levels—a pattern that differs from those of selectivity and adaptation noise.

To go beyond hypothetical scenarios of uniform initial opinions (as assumed in most formal models; Carro et al., 2013; Sobkowicz, 2020), we initialised our model with a data-driven distribution of opinions about climate change (Maibach et al., 2011). According to this distribution (and in line with other surveys on climate change opinions; European Social Survey European Research Infrastructure (ESS ERIC), 2020; Pew Research Center, 2022), the majority of people are somewhat concerned about climate change (with $18\,\%$ in the 'alarmed' and $33\,\%$ in the 'concerned' categories), but people appear far from having reached a consensus (with significant fractions of $19\,\%$ in the 'cautious', $12\,\%$ in the 'disengaged', $11\,\%$ in the 'doubtful' and $7\,\%$ in the 'dismissive' categories in 2008) (Maibach et al., 2011). In some cases, opinion diversity can be beneficial to collective decision-making, for example, when diversity prevents conformity, groupthink or overconfidence (Coglianese, 2001; Smaldino et al., 2023). However, as a large body of climate experts judge climate change as a 'threat to human well-being and planetary health' (IPCC, 2023) and call for an 'emergency response' (Lenton et al., 2019), continued disagreement among large sectors of the population may undermine the support for policies aiming to produce an adequate response.

Surprisingly, we found that a group of agents affected by a moderate bias and a medium ambiguity noise provided the best conditions for aligning towards a high level of concern about climate change. The results of the climate change scenario are qualitatively similar to the hypothetical one, but they highlight even more prominently the separate effects of different noise types. With the help of normally distributed ambiguity noise, agents can reach an agreement under a moderate bias and, crucially, this agreement is reached without having to compromise the majority's pro-environmental attitudes. In this scenario, ambiguity noise clearly outperforms other types of noise and produces the unique pattern of a society driving itself towards a shared, extreme opinion. Of course, we should be cautious about drawing strong conclusions about predictions or prescriptions concerning climate change opinions, as the way those opinions are formed and updated surely involves many factors not included in our model. Nevertheless, our model provides a good jumping-off point for such considerations.

Mathematical models, even the most sophisticated ones, are obviously only approximations of real-world social processes. For example, our model leaves aside media influence, social structure, value systems, ideology or trust in experts. While we are confident in the robustness of the effects of ambiguity noise, there are several avenues to increase the realism of the model. We focus here on three that we deem particularly promising. First, people have only few salient social influences (Bond et al., 2012) with homophily constraining the sources of those influences (McPherson et al., 2001). Our analysis indicates that embedding agents in a small-world network does not critically alter the qualitative results regardless of the network parameters (see section 2.4.5). Still, future research could investigate how homophily, represented as a bias in the selection of partners in such networks (as in Dandekar et al., 2013; Smaldino & Jones, 2021), alters the opinion dynamics. Second, ambiguity is not entirely unintended. People strategically adapt the degree of ambiguity when expressing opinions to their peers (Eisenberg, 1984; Frankenhuis et al., 2023; Hankins et al., 2023; Smaldino & Turner, 2022). By assuming that the noise in the model is heterogeneous and adaptive, future research could investigate how agents develop communication strategies that use ambiguity strategically to

spread opinions to their peers more effectively. Similarly, future research could investigate how the assumption that confidence bounds are heterogeneous and adaptive affects our results. Third, we focused on studying each type of noise separately, favouring a comparison of their contrasting effects and neglecting their interplay. Future research could investigate how interactions between the different types of noise affect opinion patterns (as in the study by Grauwin and Jensen (2012), which explored the effects of interacting selectivity and exogenous noise).

Social influence is only one of the many factors shaping opinions, but it plays a central role in many political debates (Nowak et al., 1990), and particularly in the debate on climate change (Kjeldahl & Hendricks, 2018; Pearson et al., 2016; Toomey, 2023). We have argued here that ambiguity in communicated messages is an inherent feature of such social influence and is particularly pronounced when dealing with a complex topic like climate change. There are ways to reduce ambiguity noise in a social debate, for example, by enforcing more transparent and rigorous communication (Kahneman et al., 2021) or by fostering the use of clearly identifiable markers (Smaldino & Turner, 2022), such that socially transmitted opinions are more representative of the actual opinions (e.g. by wearing clothes items or using hashtags associated with the support for a particular opinion (Powell et al., 2023)). Our study, however, supports literature promoting the benefits of noise or ambiguity in communication to foster agreement across society (e.g. Bak-Coleman et al., 2021; Couzin et al., 2011; Eisenberg, 1984; McMahan & Evans, 2018; Shirado & Christakis, 2017; Smaldino et al., 2023). Our findings imply that in the presence of confirmation bias, which is a crucial cognitive factor in the climate change debate (Johnson & Levin, 2009; Moser et al., 2022), communication that leaves some space for ambiguity might prove beneficial not only for reaching an agreement but also for strengthening a shared concern on a topic like climate change. This is an important preliminary step for effective policy-making to address such political challenges, but one should be aware of the different kinds of noise and their different impacts on collective opinion formation.

Data accessibility. The model is coded in python[6]. To foster reproducibility, transparency, and flow of ideas, we make the code publicly available at github.com/PeterSteiglechner/noise-in-OD and have archived it within the Zenodo repository https://doi.org/10.5281/zenodo.10644179. We have also implemented the model in the *defSim* package (Laukemper et al., 2020) to facilitate investigations on how these types of noise affect opinion patterns under different model assumptions. Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.7095862.

Authors' contributions. P.S.: conceptualisation, formal analysis, visualisation, writing—original draft, writing—review and editing; M.A.K.: conceptualisation, formal analysis, visualisation, writing—original draft, writing—review and editing; P.E.S.: conceptualisation, writing—review and editing; D.M.: writing—review and editing; A.M.: conceptualisation, project administration, writing—review and editing. All authors gave final approval for publication and agree to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. P.S., D.M. and A.M. acknowledge funding from the German Research Foundation (DFG) through the project SEATRAC (project number: 423711127) as part of the Priority Programme 1889: Regional Sea Level Change and Society (SeaLevel). M.A.K. acknowledges IAST funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010.

[6]Python Software Foundation. Python Language Reference, version 3.9.5. Available at http://www.python.org.

Chapter 3

Social identity bias and communication network clustering interact to shape patterns of opinion dynamics

Peter Steiglechner[1,2,*], Paul E. Smaldino[3,4], Deyshawn Moser[5,6], Agostino Merico[1,2]

[1]Systems Ecology Group, Leibniz Centre for Tropical Marine Research (ZMT), Bremen, Germany

[2]School of Science, Constructor University, Bremen, Germany

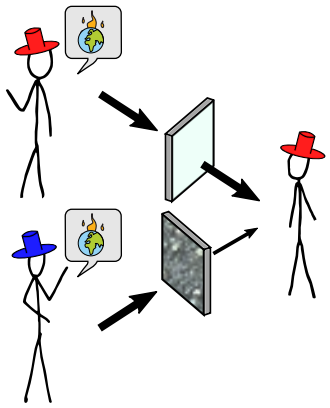
[3]Department of Cognitive and Information Sciences, University of California Merced, Merced, USA

[4]Santa Fe Institute, Santa Fe, USA

[5]Institutional and Behavioural Economics Group, Leibniz Centre for Tropical Marine Research (ZMT), Bremen, Germany

[6]School of Business, Social & Decision Sciences, Constructor University, Bremen, Germany

[*] *Corresponding author:* `peter.steiglechner@leibniz-zmt.de`



This chapter contains a manuscript published as

Steiglechner, P., Smaldino, P. E., Moser, D., & Merico, A. (2023). Social identity bias and communication network clustering interact to shape patterns of opinion dynamics. *Journal of The Royal Society Interface*, *20*(209), 20230372. https://doi.org/10.1098/rsif.2023.0372.

Abstract

Social influence aligns peoples' opinions, but social identities and related in-group biases interfere with this alignment. For instance, the recent rise of young climate activists (e.g. 'Fridays for Future' or 'Last Generation') has highlighted the importance of generational identities in the climate change debate. It is unclear how social identities affect the emergence of opinion patterns, such as consensus or disagreement, in a society. Here, we present an agent-based model to explore this question. Agents communicate in a network and form opinions through social influence. The agents have fixed social identities which involve homophily in their interaction preferences and in-group bias in their perception of others. We find that the in-group bias has opposing effects depending on the network topology. The bias impedes consensus in highly random networks by promoting the formation of echo chambers within social identity groups. In contrast, the bias facilitates consensus in highly clustered networks by aligning dispersed in-group agents across the network and, thereby, preventing the formation of isolated echo chambers. Our model uncovers the mechanisms underpinning these opposing effects of the in-group bias and highlights the importance of the communication network topology for shaping opinion dynamics.

Keywords: *Agent-based model, In-group bias, Computational social science, Opinion dynamics, Generational conflict, Climate change*

3.1 Introduction

Social influence is a key driver of opinion formation. Opinions should presumably align over time as people discuss an issue and observe what others think about that issue (Barnes et al., 2020; Moussaïd et al., 2013; Nowak et al., 1990). Yet, people do not seem able to find a consensus on many pressing political, environmental, or economic issues, including climate change (Dunlap et al., 2016), public health measures (like wearing face masks to prevent the spread of Covid-19; Milosh et al., 2021), or government-funded social measures (Balietti et al., 2021). Among these examples, climate change is particularly striking because public opinions remain divided, even though an overwhelming scientific consensus has since long been established. Social influence plays a critical role in this debate (Bak-Coleman et al., 2021; Kjeldahl & Hendricks, 2018; Pearson et al., 2016; Wallis & Loy, 2021) as pro-environmental attitudes can spread among people, but so can climate scepticism. What makes collective opinion formation a particularly complex problem is that social learning is not a straightforward copying of other's opinions but follows context-dependent strategies (Kendal et al., 2018) which are typically heuristic, affected by cognitive biases (Balietti et al., 2021; Pearson & Schuldt, 2018), and based on a limited number of social connections among individuals (Bond et al., 2012).

A major factor that impacts social influence is social identity (Price, 1989; Smaldino, 2022), especially in the debate about climate change (Estrada et al., 2017; Fielding & Hornsey, 2016; Kjeldahl & Hendricks, 2018). Individuals affiliate with a specific group, the 'in-group', based on similar personal or cultural characteristics such as political orientation, age, sex, or occupation. The way they interact with each other depends on whether or not they share such a social identity (Hornsey, 2008; Tajfel, 1974). For example, identity-related

factors influenced how people perceived specific policies designed to mitigate the spread of Covid-19 (Flores et al., 2022). Similarly, identification with different generations, like 'youngsters' or 'elders', can add an independent inter-group dimension to the debate on climate change. Age has indeed been shown to correlate with opinions on climate change (Gonyea & Hudson, 2020; Hall et al., 2018; Hornsey et al., 2016). More importantly, age and related social identity may shape communication and interaction behaviours in the climate change debate by establishing generational in- and out-groups.

Social identity can affect different aspects of communication. First, social identity influences who interacts with whom. People are homophilic, i.e. they preferentially interact with members of their in-group (Lazarsfeld & Merton, 1954; McPherson et al., 2001). For example, young people observe or discuss more likely the opinions of their young peers and vice versa. Second, social identity influences how individuals perceive socially-transmitted information. In particular, people tend to view information coming from in-group sources as more trustworthy and relevant than information originating outside the group. Group membership can thus become an important factor when evaluating the subjective opinions of others. Such differential evaluation of information, known as in-group bias, is a well-studied and persistent feature of human behaviour (Bartels, 2002; Brewer, 1979; Fielding & Hornsey, 2016; Hewstone et al., 2002; Mackie et al., 1992; Powell et al., 2023; Richerson et al., 2016; Turner et al., 1989), although the impacts of this bias vary depending on the culture of the people involved (Moser et al., 2022) and the nature of the relevant social identities (Brewer, 1999). In-group bias usually reinforces similarities between in-group members (Brewer, 1979) and, at the same time, draws like-minded people into the group (Fielding & Hornsey, 2016). Apart from such in-group favouritism, social identity can also involve out-group derogation or aversion, causing divides between groups to become more pronounced (Brewer, 1999; Hewstone et al., 2002; Smaldino et al., 2017). In summary, social identity does not necessarily determine *what* a particular individual believes about a debated issue like climate change, but social identity can affect (1) *who* that person interacts with and (2) *how* that person perceives opinions of others.

Mathematical models have become a powerful tool to study opinion formation and constitute a valuable complement to traditional social science approaches, such as laboratory experiments or surveys (Bak-Coleman et al., 2021; Galesic et al., 2023). Models can be used to explore a variety of scenarios, theories, or assumptions. In particular, they can provide a link between how cognitive processes, such as in-group bias, manifest at the individual level and how this plays out at the collective level. Most opinion dynamics models are agent-based and simulate how human agents embedded in a social structure update their opinions as they incorporate new information – either through social influence or external stimuli. Such models typically define (1) how people are connected to each other (societal structure), (2) how their opinions are represented, (3) how they acquire new information, and (4) how they process new information to update their opinions (update rule). Homophily or biases are typically integrated by modulating certain components of the model at the individual level. In particular, homophily is often encoded in the societal structure (e.g. Dandekar et al., 2013; Mäs et al., 2013; Smaldino & Jones, 2021) and biases are typically encoded in the agents' update rules (e.g. Flache et al., 2017; Sobkowicz, 2018).

Different opinion dynamics models have studied various ways to conceptualise biases (e.g. Chen et al., 2017; Dandekar et al., 2013; Sobkowicz, 2018). Most of these models focus on biases related exclusively to the exchange of opinions on the debated topic(s). For example, bounded confidence models assume that the opinion distance of two agents fully determines whether they perceive each other as similar and, thus, whether they are influenced by each other. Social identity theory, however, suggests that such biased influence does not depend on peoples' opinions alone. Perceived similarity and, consequently, the degree of influence between

people also depends on their social identities and on group perceptions. We, therefore, argue—in line with previous works, such asAlizadeh et al. (2015), Feliciani et al. (2021), Pearson and Schuldt (2018), Smaldino (2022), and Squazzoni et al. (2014)—that it is ultimately the interplay of the exchange of opinions *and* of social information that shapes opinion formation. While there is much research in social psychology and related sciences on the influence of social identity on opinion formation – especially at the individual level – such aspects are still understudied with mathematical models of collective opinion formation (Galesic et al., 2021; Sobkowicz, 2020; Squazzoni et al., 2014).

In this study, we investigate the effects of social identity and the related in-group bias on patterns of opinion formation using an agent-based model. Our main assumptions are that individuals align their opinions to those of their social contacts, in line with the social influence literature (e.g. Festinger, 1954; Nowak et al., 1990), and that social identity moderates such alignment (in line with social identity theory; Hewstone et al., 2002), leading to greater shifts in opinions due to interactions from in-group members vs. out-group members. Moreover, we assume that individuals interact in non-random ways but that they have stable in- and out-group contacts, which can be represented as a fixed network. While the network topology is in principle uncertain, we assume that agents tend to have more in-group than out-group contacts, in line with literature on homophily (e.g. McPherson et al., 2001). In particular, the model is designed along the following principles. Agents hold opinions on a specific topic, for example, climate change, which evolve when they communicate with their neighbours in a fixed network. Agents then use a heuristic approach, formulated following Bayesian calculus, to adapt to perceived opinions. The basic opinion formation process follows previous studies by Martins (Martins, 2009) and Sobkowicz (Sobkowicz, 2018). We extend their framework by including social identity. Agents identify with one of two groups, for example, youngsters or elders in the context of climate change, and we assume that this social identity is visible to others. The effects of social identity are twofold: agents are homophilic in their interaction preferences with respect to social identity and agents may be biased in the way they perceive others.

This model design allows us to investigate our main research question: does in-group bias impede or foster consensus among homophilic agents? It may seem obvious that in-group bias should create divisions between groups, inhibiting consensus formation. However, in-group bias may also speed up the convergence of opinions within in-groups, thus allowing opinions to spread further across the network before polarisation can take hold (Gabriel & O'Connor, 2024; Smaldino et al., 2023). This suggests that the structure of social interactions may moderate the effects of in-group bias. We thus ask: is the effect of in-group bias consistent over different network topologies? If not, what is the mechanism by which in-group bias can foster consensus? By setting up the model with different homophilic network topologies, we show that the in-group bias impedes consensus in societies if the topology is highly random. In contrast, the in-group bias fosters consensus in societies if the topology is highly clustered. These contrasting effects are robust outcomes unless homophily is very high or agents are strongly predisposed in their initial opinions such that an enhanced disagreement between the identity groups at the outset of the simulations inhibits the formation of consensus. The results of our model suggest that the impact of social identity and in-group bias can only be evaluated in relation to the underlying topology of the communication network.

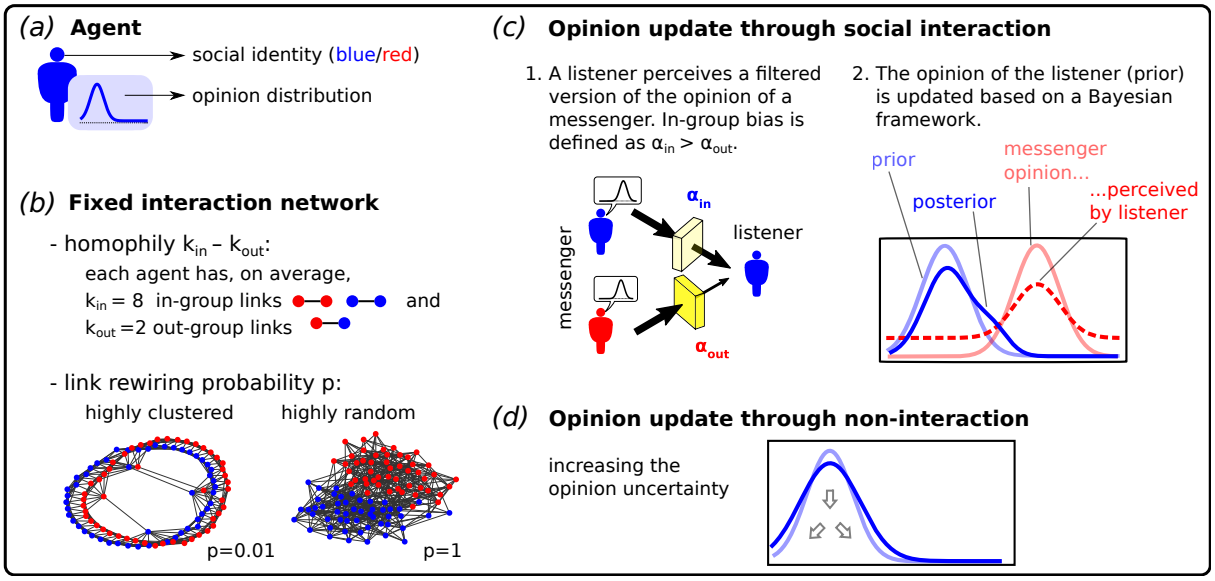


Figure 3.1: Schematic representation of (a) an agent with associated properties, (b) the interaction network, which remains unchanged throughout the simulation and is characterised by homophily and link rewiring (creating network topologies varying from highly clustered to highly random), and the opinion update process, which includes (c) social interaction or (d) non-interaction.

3.2 Model description

3.2.1 Summary

Our model simulates opinion formation in a small, virtual society consisting of n agents. Figure 3.1 shows a conceptual representation of the model and the effects of social identity on the model processes. Each agent represents an individual human being and is characterised by (1) an opinion on a disputed issue like climate change, expressed as a distribution, and (2) a social identity, expressed as an exclusive affiliation to one of two distinct groups (figure 3.1*a*). Agents are embedded in a fixed interaction network, which is characterised by (1) a degree of identity-driven homophily and (2) a probability of link rewiring (figure 3.1*b*). The network topologies range from highly clustered to highly random. Over time, the opinions of all agents evolve via one of two processes. With a certain probability, a focal agent interacts with a linked neighbour and its opinion distribution is modified by that interaction. In-group bias involves that agents are more influenced by interactions with in-group members than by interactions with out-group agents (figure 3.1*c*). If the agent does not interact, its opinion distribution broadens, thus, increasing the susceptibility of that agent's opinion to future social influences (figure 3.1*d*). In the following sections, we provide detailed explanations of these model features. With the objective of fostering reproducibility, transparency, and flow of ideas, we make the model available as open-source software (Steiglechner, 2023) so that it can be used, modified, and redistributed freely.

3.2.2 Agent characteristics

Each agent is characterised by an opinion about the debated issue. With 'opinion', we mean the agent's subjective point of view on the topic, such as its degree of concern about climate change. Opinions and

beliefs are complex cognitive constructs, involving mental representation and rational processing of both direct evidence and social influence. For simplicity, we use the word 'opinion' in a way consistent with much of the opinion dynamics literature, in which opinions are updated through social influence and not direct experience. As such, our results will apply most readily to opinions shaped largely through social influence. Adopting a framework used in previous models (Galesic et al., 2021; Martins, 2009; Sobkowicz, 2018), we represent the opinion of agent i at time t as a distribution $x_i(b,t)$ (x_i in the following) over the belief space $\mathcal{B} = [-1, 1]$. The opinions are initially assumed as Gaussian distributions characterised by a mean and a variance, the latter reflecting the agent's uncertainty around the mean opinion. Over time, the shape of these distributions can change and may even become multi-modal. Each value b in the belief space represents a statement, for example about climate change, ranging from $b = -1$ ('I am not at all concerned about climate change') to $b = +1$ ('I am extremely concerned about climate change'). The opinion $x_i(b,t)$ represents the level of support of agent i for these statements b at time t. For computational reasons, we approximate the distribution and the belief space with 200 discrete, equally spaced values $\mathcal{B} = \{-0.995, -0.985, \ldots, 0.995\}$.

Each agent is also characterised by a social identity. This is expressed as an affiliation to one of two groups, red or blue, which can represent 'youngsters' and 'elders', for example. The social identity groups contain an equal number of agents. Social identities are visible to others and remain fixed throughout a simulation. Although in reality opinions often correlate with identities (Druckman et al., 2021), we assume that the agents' opinions are independent of their social identities (but relax this assumption later in Section 3.3.6 when we analyse a scenario in which social identities predispose the agents to specific opinions). The social identity of an agent defines its in-group—those agents sharing the same social identity—and the out-group—those with different social identities.

3.2.3 Interaction network

Agents are situated in an interaction network consisting of nodes and undirected, unweighted links. Nodes represent agents, and only agents connected through direct links can communicate. This reflects the fact that, in reality, most people have relatively few salient social influences that shape their opinions (Bond et al., 2012). The network consists of two types of links: in-group links, connecting agents who share the same social identity, or out-group links, connecting agents with different social identities. In- and out-group links represent the same influence channels, although, in reality, these channels may have different characteristics. Assuming that social relationships evolve at a much slower pace than opinions, the network is created before the simulation begins and remains unchanged throughout the simulation. The network is defined by three parameters: (1) the average number of in-group links per agent, denoted as k_{in}, (2) the average number of out-group links per agent, denoted as k_{out}, and (3) the link rewiring probability, denoted as p (more on this parameter later). We define a network as homophilic if $k_{\text{in}} > k_{\text{out}}$ (see Dandekar et al., 2013; Karimi et al., 2018, for similar conceptualisations of homophily). For instance, in a fully homophilic network with an average node degree of 10, the agents would have $k_{\text{in}} = 10$ in-group links and $k_{\text{out}} = 0$ out-group links, indicating a complete separation of the social identity groups. In a non-homophilic network, the agents would have, on average, $k_{\text{in}} = 5$ in-group links and $k_{\text{out}} = 5$ out-group links.

The network is constructed by creating two separate ring lattices, one for each social identity group. To establish in-group links, we connect each agent to its k_{in} nearest neighbours within its own group in the

corresponding ring lattice[1]. To establish out-group links, we connect each agent in the red group lattice to its k_{out} closest agents in the blue group lattice. For example, in a network with $n = 100$ agents, where agents 1 to 50 are red and agents 51 to 100 are blue, for $k_{\text{out}} = 2$, agent 1 is connected to its two closest out-group agents, 51 and 52, agent 2 is connected to agents 52 and 53, and so on (see Supplementary Figure B.1). After establishing this deterministic, homophilic network structure, all links are rewired with probability p. For the rewiring of in-group links, we follow the procedure described in the Watts-Strogatz model (Watts & Strogatz, 1998). For the rewiring of out-group links, an existing link between agents i and j is substituted with a link between agent i and a randomly selected agent k from the out-group of agent i. This rewiring process preserves the total number of in-group and out-group links, thereby maintaining the degree of homophily in the network.

Homophily (k_{in} and k_{out}) and rewiring (p) both play an important role in the topology of the network. In our main analysis, we focus primarily on networks with moderate homophily, specifically $k_{\text{in}} = 8$ and $k_{\text{out}} = 2$ (although we vary the degree of homophily in the sensitivity analysis). With this degree of homophily, we define a network as *highly clustered* when it is created with minimal or no rewiring ($p \to 0$) and as *highly random* when it is created with maximum rewiring ($p \to 1$). Highly clustered networks are characterised by a high clustering coefficient and a long average path length (see Supplementary Figure B.2). Moreover, they possess a regular structure (i.e. all agents have the same number of in- and out-group links) and a local topology (i.e. most of the neighbours of an agent are connected among themselves). Highly random networks are characterised by a low clustering coefficient and a short average path length, and the node degrees of the agents are heterogeneous (i.e. while some agents may be connected exclusively to in-group members, others may have many out-group links). Highly random networks resemble those generated by the stochastic block model (Holland et al., 1983) with a prescribed number of in- and out-group links.

3.2.4 Opinion update

At every time step, all n agents are selected asynchronously and in random order. With probability q, a selected agent (the listener) interacts with one of its neighbours (the messenger). With probability $1 - q$, the listener does not interact, and its opinion distribution becomes more uncertain.

Opinion update through social interaction

Interaction implies that the opinion of listener i changes based on observing the opinion of messenger j. We assume that the listener i only sees a distorted version of the messenger's opinion distribution x_j. Specifically, following a previous opinion formation model (Martins, 2009), the listener sees the distribution x_j with an uncertainty that is higher than the actual value. This reflects the fact that humans tend to evaluate others in a conservative way (Phillips & Edwards, 1966), perceiving their opinions as less conclusive than what they are. As described in the introduction, social identity can undermine opinion change. For example, individuals may trust in-group members more, feel a stronger need to conform with them, or simply relate more to their in-group peers due to a shared language. To implement this bias, we assume that the degree of distortion depends on the social identity of the messenger. That is, for interactions with an in-group messenger j, listener i sees x_j as:

[1]Note that if k_{in} is uneven, for example, $k_{\text{in}} = 5$, then we connect each agent to its two closest agents in both clock- and counter-clockwise directions in the ring lattice and, with probability $1/2$, additionally to the third closest agent in the clockwise direction, such that the average in-group link degree is $k_{\text{in}} = 5$.

$$p_i(x_j) = \alpha_{\text{in}} \cdot x_j + (1 - \alpha_{\text{in}}) \cdot \mathcal{U} \ , \tag{3.1}$$

with in-group perception α_{in} and, equivalently, with out-group perception α_{out} for interactions with out-group messengers. $\mathcal{U}$ denotes the uniform distribution on the belief space $\mathcal{B}$. This perception step acts as a filter for x_j (see figure 3.1*c*) and $\alpha_{\text{in/out}}$ represent the transparencies of the filters for in-/out-group messengers, respectively. A value of $\alpha_{\text{in/out}} = 1$ (fully transparent filter) implies that the listener perceives x_j accurately, whereas $\alpha_{\text{in/out}} = 0$ (fully opaque filter) implies that the listener perceives a uniform distribution that is unrelated to the messenger's opinion. Negative values of $\alpha_{\text{in/out}}$ represent repulsive social influence, i.e. the listener would perceive the support of the messenger for a certain belief as counter-evidence for that belief. In this study, we restrict ourselves to positive values of $\alpha_{\text{in/out}}$.

We define agents as affected by in-group bias when their in-group perception is higher than their out-group perception (i.e. $\alpha_{\text{in}} > \alpha_{\text{out}}$), meaning that a biased listener perceives the opinion of an in-group messenger as more certain than the same opinion of an out-group messenger. The in-group bias reflects a preference for accepting the opinions of in-group individuals over those of out-group individuals, with stronger bias indicating a larger disparity between the two. For simplicity, we assume that α_{in} and α_{out} are the same for all agents.

The opinion of the listener is updated following Bayesian calculus. In line with previous models (Acemoglu & Ozdaglar, 2011; Bartels, 2002; Martins, 2009; Sobkowicz, 2018), we consider the opinion distribution of the listener as the prior and the distorted version of the messenger's opinion, as seen by the listener, as the likelihood. After an interaction with an in-group messenger j, the updated opinion of listener i becomes the Bayesian posterior (before normalisation):

$$x_i \leftarrow x_i \cdot p_i(x_j) \ = \ \alpha_{\text{in}} \cdot x_i \cdot x_j + (1 - \alpha_{\text{in}}) \cdot x_i \cdot \mathcal{U} \tag{3.2}$$

and, equivalently, with α_{out} after an interaction with an out-group messenger. Equation (3.2) encompasses two competing forces: (1) an assimilative force that pulls the opinion of the listener towards the opinion of the messenger and (2) a conservative force that keeps the opinion of the listener unchanged, especially if the opinion of the messenger is very distant. The relative strengths of these forces depend on α_{in} and α_{out} and on the overlap between x_i and x_j. Supplementary Figure B.3 provides an illustrative example of how these forces shape the posterior opinions of agents depending on $\alpha_{\text{in/out}}$ and different messenger opinions.

Opinion update through non-interaction

At each time step, only a fraction of the agents (on average $q \cdot n$) change their opinions following social interactions. For the remaining agents, non-interaction slightly increases the uncertainty characterising their opinion distributions. This reflects the fading of strong emotions or imperfect memory of arguments (see Geschke et al., 2019; Mäs et al., 2013; Sobkowicz, 2018, for models with similar processes). The posterior opinion of a non-interacting agent is obtained by solving the diffusion heat equation for one time step $t \to t+1$ (using implicit differentiation via the backward time-centred space method):

$$\frac{\mathrm{d}}{\mathrm{d}t}\, x_i(b,t) = \kappa \cdot \frac{\mathrm{d}^2}{\mathrm{d}b^2}\, x_i(b,t) \quad \overset{\substack{\text{discrete} \\ \text{solution}}}{\longrightarrow} \quad \frac{{x_i}_b^{t+1} - {x_i}_b^{t}}{\Delta t} = \kappa \cdot \frac{{x_i}_{b+\Delta b}^{t+1} - 2 \cdot {x_i}_b^{t+1} + {x_i}_{b-\Delta b}^{t+1}}{\Delta b^2} \tag{3.3}$$

with zero Dirichlet boundary conditions at the edges of the belief space. This process depends on the parameter κ, which determines the speed with which the opinion distribution decays during non-interaction. While social interaction tends to narrow down opinion distributions, non-interaction broadens opinion distributions, making them more susceptible to change during future interactions. Consequently, if an agent does not interact for a sufficiently long time, it eventually adopts a uniform opinion distribution, indicating complete neutrality or indifference towards the issue. The impact of a more recent social interaction on an opinion, thus, tends to outweigh that of previous interactions.

3.2.5 Initialisation and analysis of results

At the beginning of a simulation, $t = 0$, we create a society of $n = 100$ agents as follows. We divide the agents into two evenly sized social identity groups, red and blue, and situate them in a fixed network with a moderate level of homophily such that agents are on average linked to $k_{\text{in}} = 8$ in-group and $k_{\text{out}} = 2$ out-group members and with a specific degree of clustering determined by the rewiring probability p. Then, we initialise the agents with opinions represented as Gaussian distributions, $x_i \sim \mathcal{N}(\mu_{i,0}, \sigma_{i,0})$, with fixed variance $\sigma_{i,0} = \sigma_0 = 0.2$ and randomly sampled mean $\mu_{i,0} \in [-1, 1]$. At the bounds of the belief space, the distributions are truncated and, therefore, mean and variance are not equivalent to those of the initial Gaussian distribution. This choice of opinion initialisation allows us to distribute agent opinions uniformly over the belief space (except at the bounds), such that the initial state represents a very diverse and heterogeneous society. In Section 3.3.6, we explore an alternative scenario in which agents have predisposed initial opinions depending on their social identities, i.e. an agent i is initialised with a negative $\mu_{i,0}$ with a higher probability of $0.5 + \delta/2$ (with predisposition δ) if it has the red identity, and a lower probability of $0.5 - \delta/2$ if the agent has the blue identity. Finally, we fix the interaction probability $q = 0.2$ and the opinion decay speed during non-interaction $\kappa = 0.0002$. We analyse the robustness of our assumptions by varying the values of these parameters in Sections 3.3.4 to 3.3.6.

Our main analysis focuses on comparing the opinion patterns obtained with different in-group and out-group perceptions and, thus, also different strengths of in-group bias. To avoid potential divisions by 0 when normalising the posterior in equation (3.2), we choose $\alpha_{\text{in/out}} \in [0, 0.99]$. For illustrative purposes, we later focus on a set of distinct values of α_{in} and α_{out}. We consider three particular cases, U_1, U_2, and U_3, representing societies of unbiased agents, and one case, B, representing a society of biased agents. There are three unbiased cases because agents can be characterised by different levels of perception. Regardless of social identity, agents in U_1 are 'sceptical' towards others ($\alpha_{\text{in}} = \alpha_{\text{out}} = 0.25$), agents in U_2 are 'neutral' towards others ($\alpha_{\text{in}} = \alpha_{\text{out}} = 0.5$), and agents in U_3 are 'credulous' towards others ($\alpha_{\text{in}} = \alpha_{\text{out}} = 0.75$). In contrast, agents in societies B are 'credulous' towards in-group agents ($\alpha_{\text{in}} = 0.75$) and 'sceptical' towards out-group agents ($\alpha_{\text{out}} = 0.25$). Comparing the results of U_1, U_2, and U_3 with those of B allows us to isolate the effects of the bias.

We start our analysis by describing the specific opinion patterns produced by the model in general terms (Section 3.3.1). Then, we systematically analyse how often societies reach a consensus after a fixed number of t time steps, i.e. the consensus frequency, C_{t}, for societies characterised by highly random (Section 3.3.2) or highly clustered networks (Section 3.3.3). There are many ways to quantify disagreement or consensus (Bramson et al., 2017; Flache & Mäs, 2008; Gestefeld et al., 2022; Turner & Smaldino, 2018). Here, we measure the level of disagreement with the standard deviation, σ, of the opinion means of all agents and we define consensus when σ is below a threshold $\sigma_{\text{cons}} = 0.01$. Simulations are terminated after $t = 5000$

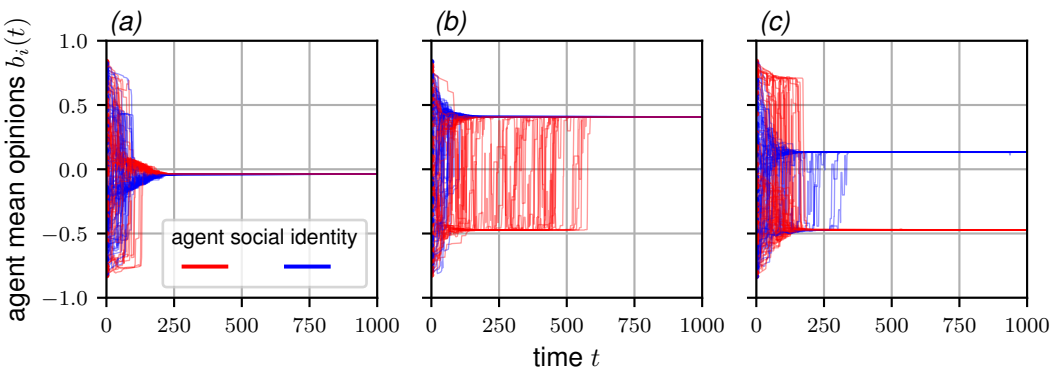


Figure 3.2: Time evolution of the agents' mean opinions in three example simulations of a society B composed of biased agents with in-group perception $\alpha_{\text{in}} = 0.75$ and out-group perception $\alpha_{\text{out}} = 0.25$, and an interaction network characterised by moderate homophily, $k_{\text{in}} = 8$ and $k_{\text{out}} = 2$, and maximum link rewiring, $p = 1$. Panels a to c show realisations of the same society but with different random seeds, which affects (1) the initial conditions, (2) the network topology, and (3) the update order. In these examples, opinions (a) converge quickly, at $t = 198$, to a consensus, (b) separate into two distinct clusters, corresponding to the two social identity groups, that converge to a consensus at $t = 585$, when one opinion cluster absorbs the other, or (c) separate into two distinct clusters that remain separate over the entire simulation.

steps, when an agent has, on average, updated its opinion 1000 times during one-on-one interactions (with interaction probability $q = 0.2$). The results do not change qualitatively for shorter and longer simulation times (see Section 3.3.4). Because the model is stochastic with respect to network topology (if $p > 0$), update order, and initial opinions, we present consensus frequencies, C_{t}, as averages over ensemble runs of 1000 realisations of a society with the same parameter configuration but different random seeds.

3.3 Results

3.3.1 Consensus can emerge through an abrupt, stochastic transition

Figure 3.2 shows the temporal evolution of opinions in three example simulations of a society B composed of biased agents with $\alpha_{\text{in}} = 0.75$ and $\alpha_{\text{out}} = 0.25$. The societies in these examples are characterised by homophilic and highly random networks, i.e. $k_{\text{in}} = 8$, $k_{\text{out}} = 2$, and $p = 1$. In the early phase of a simulation, opinions converge within small local neighbourhoods. This leads to the formation of opinion clusters. Neighbourhoods in the highly random network are densely interconnected such that these clusters typically dissolve quickly, causing fast overall convergence towards a relatively moderate consensus opinion (figure 3.2*a*). If, however, distinct clusters become sufficiently large and the convergence within these clusters dominates over the alignment among different clusters, then disagreement stabilises and fast consensus is prevented (figure 3.2*b* and 3.2*c*). Disagreement, however, is a transient, meta-stable state of the system. In some cases (e.g. figure 3.2*b*), alignment pressures among the opinion clusters dominates over their internal cohesion. This leads to an abrupt, stochastic transition towards a consensus where one cluster absorbs the other and the resulting consensus opinion tends to be more extreme than when consensus is reached very early in the simulation (as in figure 3.2*a*; see Supplementary Figure B.4). In other cases (e.g. figure 3.2*c*), disagreement persists because opinion clusters turn into echo chambers, in which the agents are confronted almost exclusively with similar opinions. Over time the agents become increasingly certain about their opinions (i.e. the variances of their opinion distributions narrows down) and, thus, unresponsive to agents with different opinions. In this scenario, we find that the

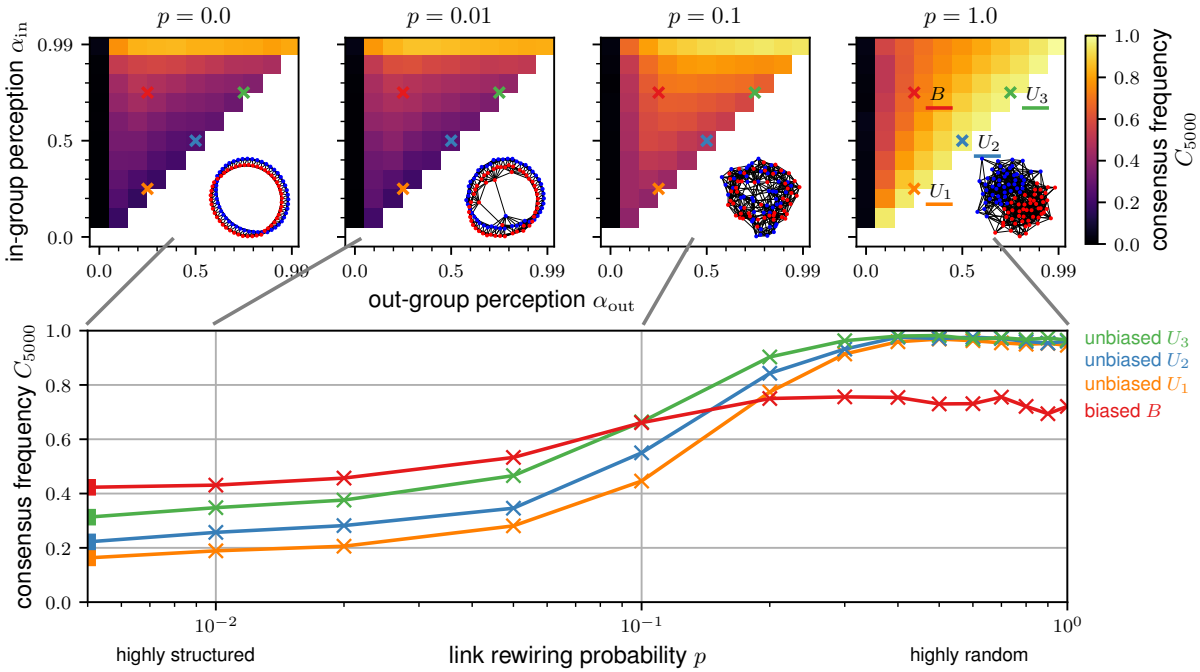


Figure 3.3: Consensus frequency C_{5000} before $t = 5000$ time steps for societies characterised by networks with homophily, $k_{\text{in}} = 8$ and $k_{\text{out}} = 2$, and different link rewiring probabilities $p \in [0, 1]$. In highly random networks ($p \gg 0.1$), consensus occurs less often in a society B of biased agents (red) than in societies U_1 (orange), U_2 (blue) or U_3 (green) of unbiased agents. However, in highly clustered networks ($p < 0.1$), a society B reaches consensus more frequently than societies U_1, U_2, or U_3. The inset panels $p = 0$, $p = 0.01$, $p = 0.1$, and $p = 1$, show the consensus frequencies over the full parameter space of α_{in} and α_{out}, where societies of unbiased agents, such as U_1, U_2, and U_3, are located on the diagonal and societies of biased agents, such as B, are located in the upper left corner. With increasing bias (i.e. moving further away from the diagonal), consensus frequencies consistently decrease or increase for, respectively, highly random or highly clustered networks, unless $\alpha_{\text{out}} \to 0$. Each inset panel also shows an example of a corresponding network topology, with the node colour (blue and red) representing the two social identities.

standard deviation of all mean opinions, σ, is typically in the range from 0.2 to 0.6, similar to its initial value at $t = 0$ (and thus disagreement and consensus are qualitatively distinct patterns regardless of the exact value of the threshold, $\sigma_{\text{cons}} = 0.01$).

3.3.2 In-group bias impedes consensus in homophilic, highly random networks

Figure 3.3 shows the frequency, C_{5000}, with which societies in our ensemble simulations reach a consensus within $t = 5000$ steps, depending on the link rewiring probability, p, in the network and the agents' in- and out-group perception, α_{in} and α_{out} (in particular, for societies U_1, U_2, U_3, and B). We first focus on societies characterised by highly random networks, $p \gg 0.1$. In general, consensus is frequently reached within $t = 5000$ time steps unless the out-group perception of the agents is very low, $\alpha_{\text{out}} \to 0$. However, consensus occurs somewhat less frequently if the agents are biased (upper left part of panel $p = 1$ in figure 3.3) than if they are unbiased (diagonal part of panel $p = 1$), regardless of the exact values for α_{in} and α_{out}. For example, for $p = 1$, societies U_1, U_2, and U_3 reach consensus within 5000 time steps in, respectively, $95\,\%$, $96\,\%$ and $97\,\%$ of the simulations, but society B reaches consensus 'only' in $72\,\%$ of the simulations. These numbers increase only marginally over much longer simulation times (see Section 3.3.4). Biased agents see opinions of in-group messengers as more certain and, in homophilic networks, they also tend to interact more with

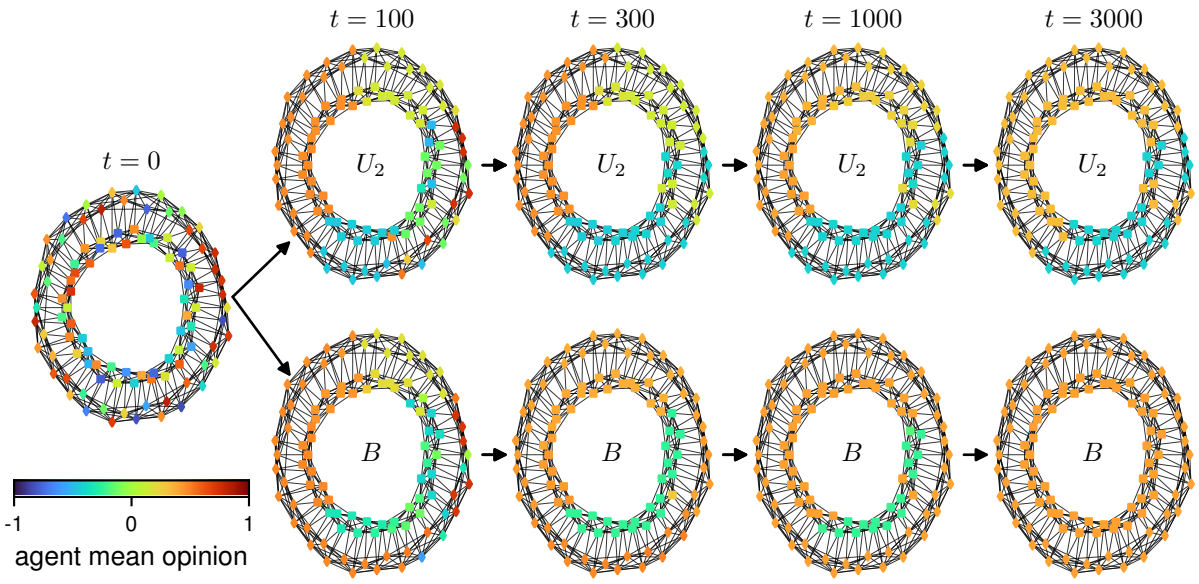


Figure 3.4: Example simulation for a society U_2 of unbiased, 'neutral' agents ($\alpha_{\text{in}} = \alpha_{\text{out}} = 0.5$) and a society B of biased agents ($\alpha_{\text{in}} = 0.75$ and $\alpha_{\text{out}} = 0.25$) where the society is characterised by a homophilic and highly clustered interaction network, $k_{\text{in}} = 8$ and $k_{\text{out}} = 2$, and $p = 0$. The shapes and positions of the nodes (outer and inner circle) represent the agents' social identities. The colours of the symbols represent the mean opinions of the agents. In U_2 (upper panels), a small, isolated opinion cluster emerges (light blue) and turns into a minority echo chamber that remains stable beyond $t = 3000$. In B (lower panels), the relatively higher alignment pressure among in-group members prevents the formation of isolated opinion clusters. One social identity group (the outer circle) reaches a consensus, which the out-group agents align to before $t = 3000$. Consensus is thus reached in society B, but not in society U_2. In this particular example, consensus is also not reached in the other societies of unbiased agents, U_1 and U_3 (see Supplementary Figure B.5).

in-group members. This combination facilitates and stabilises the emergence of echo chambers (see figure, which mostly coincide with the identity groups 3.2*c*). In sum, in-group bias impedes consensus in homophilic and highly random networks.

3.3.3 In-group bias fosters consensus in homophilic, highly clustered networks

In contrast to impeding consensus in societies with highly random networks ($p \gg 0.1$), in-group bias promotes consensus in societies with highly clustered networks ($p < 0.1$, figure 3.3). In general, higher clustering in the network, i.e. fewer rewired links, reduces the consensus frequency regardless of α_{in} and α_{out}. The reason for this is that paths between agents in a clustered network are typically longer than in a random network, which leads to weaker influence between them and to longer convergence times. However, unbiased agents are much more affected by clustering than biased agents. For example, in highly clustered networks with $p = 0$, a consensus is reached in $42\,\%$ of a society B ($72\,\%$ for $p = 1$), but only in $22\,\%$ of a society U_2 ($96\,\%$ for $p = 1$). This qualitative pattern remains robust over all time scales, even if societies characterised by such clustered networks reach consensus at time scales beyond $t = 5000$ steps (see Section 3.3.4 for more details). Note that this positive effect of the in-group bias on consensus formation holds only for societies in which the agents are at least somewhat susceptible to out-group influences, $\alpha_{\text{out}} > 0$ (figure 3.3).

This result, that in-group bias can promote consensus in highly clustered networks ($p < 0.1$), may appear counter-intuitive. For a better understanding, we provide a video in the electronic supplementary material[2] that shows a typical simulation for the societies U_1, U_2, U_3, and B, all characterised by a homophilic and highly clustered network ($p = 0$) and an identical configuration of initial opinions. Figure 3.4 shows snapshots of the society B, which reaches consensus, and of the society U_2, which does not reach consensus. In U_2, isolated opinion clusters can form across social identity divisions, with the effect that the agents are fully detached from contrasting opinions elsewhere in the network. Therefore, opinions are clustered by space rather than social identity in this scenario. In B, the alignment pressure exerted by the in-group outweighs the alignment pressure exerted by neighbouring out-group agents, even when the path length between the in-group agents tends to be large. Agents are, thus, more likely to reach a consensus within their in-group and the possibility for the formation of isolated echo chambers is reduced. Once one social identity group has reached a consensus, the agents from the other group are collectively pulled to that consensus opinion and, eventually, a society-wide consensus is established. This pattern of opinion dynamics is fostered by the in-group bias affecting agents in societies B but not agents in societies $U_{1,2,3}$.

3.3.4 The interplay between in-group bias and network clustering is a robust pattern

Our main results—in-group bias impedes consensus in highly random networks (Section 3.3.2) and fosters consensus in highly clustered networks (Section 3.3.3)—are robust under a wide range of parameter values. Figure 3.5 shows the consensus frequencies for highly random networks ($p = 1$, dots), and for highly clustered networks ($p = 0$, crosses) at different values of all model parameters. For $p = 1$, the consensus frequencies obtained in societies B of biased agents (red solid lines in figure 3.5), are reliably smaller than those in societies U_2 of unbiased, 'neutral' agents (blue solid lines). Similarly, for $p = 0$, the consensus frequencies in societies B (red dashed line in figure 3.5), are reliably larger than those in societies U_2 (blue dashed lines) with only two exceptions for extreme homophily or for strong predisposition (more details in Sections 3.3.5 and 3.3.6), or for both combined (see Supplementary Figure B.6).

While different values of n, σ_0, κ, q, and t do not change the general results of the study, some interesting effects are discernible at a higher level of detail. First, a larger number of agents, n, increases or decreases the consensus frequency depending on the network topology (figure 3.5*a*). Very large societies virtually always reach consensus when the interaction network is highly random ($p = 1$), but they barely reach consensus within $t = 5000$ time steps when the network is highly clustered ($p = 0$). Second, the parameters σ_0, κ, and q (figure 3.5*b–d*) determine how the uncertainties of the opinion distributions evolve and are, thus, crucial for consensus formation. The formation of consensus is facilitated by high uncertainties in the Gaussian initial opinions, fast decay of opinion distributions during non-interaction, and reduced frequency of interactions. Finally, the consensus frequencies depend on the simulation time scale (figure 3.5*e*). Consensus may emerge already after $t = 100$ time steps. In highly random networks, consensus frequencies reach high values relatively quickly, but in highly clustered networks consensus forms much slower and may still emerge after a much longer simulation time. Eventually, for the agents in society B, consensus becomes virtually certain in highly clustered networks ($p = 0$, dashed) but the consensus frequency saturates at $C_{10^6} = 84\,\%$ in highly random networks (solid red line).

[2]Available at https://doi.org/10.6084/m9.figshare.c.7095862

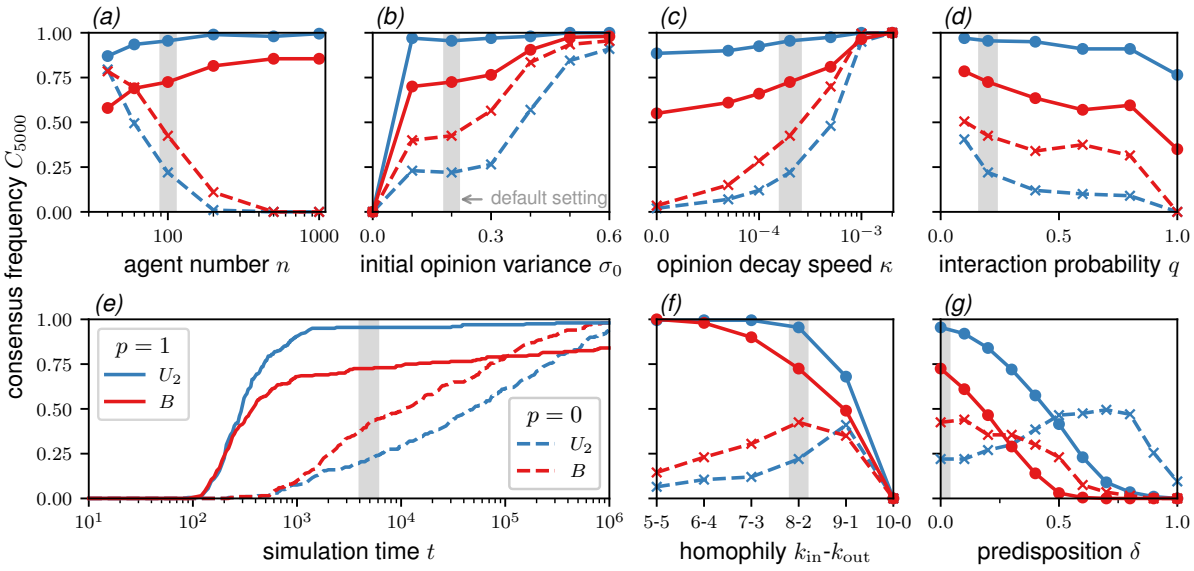


Figure 3.5: Sensitivity analysis for the consensus frequency C_{5000} of societies U_2 (blue lines) and B (red lines) characterised by highly random networks ($p = 1$, solid lines with dots) or highly clustered networks ($p = 0$, dashed lines with crosses). The parameters investigated are (a) the number of agents, n, (b) the variance of the initial Gaussian opinion distributions, σ_0, (c) the opinion decay speed, κ, during non-interaction (d) the probability of interaction, q, (e) the simulation time, t, (f) the degree of homophily, k_{in} and k_{out}, and (g) the predisposition of the initial opinions, δ. The vertical grey areas mark standard parameter values used for obtaining the results presented in the previous figures ($n = 100$, $\sigma_0 = 0.2$, $\kappa = 0.0002$, $q = 0.2$, $t = 5000$, $k_{\text{in}} = 8$, $k_{\text{out}} = 2$, and $\delta = 0$). While the parameter choices affect the exact values of the consensus frequencies, the general results are robust. Specifically, the in-group bias affecting the agents in societies B prevents consensus in highly random networks (the red solid lines are always below the blue solid lines) and the bias fosters consensus in highly clustered networks (the red dashed lines are always above the blue dashed lines). The latter holds true in all but two extreme cases, where societies are characterised by very strong homophily, $k_{\text{in}} \geq 9$ and $k_{\text{out}} \leq 1$, or high predisposition $\delta \geq 0.4$.

3.3.5 Moderate homophily promotes consensus in highly clustered networks

The presented results are based on interaction networks with moderate homophily, specifically, where agents have, on average, $k_{\text{in}} = 8$ in-group links and $k_{\text{out}} = 2$ out-group links. The main results (that the in-group bias has opposite effects on consensus formation depending on the random or clustered network topology, see Sections 3.3.2 and 3.3.3) also hold true for less homophilic or even non-homophilic networks (figure 3.5*f*). Only in networks characterised by very strong homophily, $k_{\text{in}} > 9$ and $k_{\text{out}} < 1$ (with an average of 10 links per agent), does in-group bias impede consensus in both random and clustered network topologies. Homophily by itself has diverse effects on consensus formation, regardless of the in-group bias. Increasing homophily impedes consensus in highly random networks ($p = 1$, solid lines in figure 3.5*f*), but it promotes consensus in highly clustered networks ($p = 0$, dashed lines) unless the two groups are (nearly) fully separated.

3.3.6 Predisposition of unbiased agents promotes consensus in homophilic, highly clustered networks

The presented results are based on the assumption that opinions are uniformly distributed at the start of a simulation and independent of the agents' social identities. We, therefore, tested the case of predisposed agents (figure 3.5*g*), in which the initial opinions of agents with different social identities tend to be positioned

on opposite sides of the belief space (see Supplementary Figure B.7, for a more detailed description). Our main results hold true for weak predisposition, but not for strong predisposition. In the latter case, in-group bias impedes consensus in both random and clustered networks. In general, predisposition impedes consensus. Surprisingly, however, predisposition (up to $\delta = 0.7$) can promote consensus in highly clustered networks, but only when the agents are unbiased (society U_2, dashed blue line).

3.4 Discussion

Mathematical modelling serves various purposes in the social sciences (Edmonds et al., 2019; Epstein, 2008; Smaldino, 2023). While theories of human behaviour remain largely verbal and ambiguous, idealised modelling can be an important step in the development of more detailed theories and towards a better understanding of the social phenomena (see Craver, 2006; Smaldino, 2017; Wimsatt, 1987, for further discussions). For example, a large and influential literature on opinion dynamics uses abstract, idealised models of social influence to explore factors that lead to phenomena such as consensus, polarisation, factionalisation, and extremism. Our modelling study falls into this tradition, yet our contribution is to point out how in-group bias—a common and well-documented facet of human behaviour—influences opinion dynamics in important and non-obvious ways. We have thus presented an agent-based model of opinion formation based on social influence with the social identities of the agents driving their interactions and perceptions. In line with theoretical arguments about the importance of social identity (Hewstone et al., 2002; Kahan et al., 2011; Pearson et al., 2016; Smaldino, 2022; Toomey, 2023) and with empirical studies indicating that identity may be a crucial driver of polarisation (Cook & Lewandowsky, 2016; Flores et al., 2022; Landrum et al., 2017; Mason, 2015), opinion patterns in our model are crucially shaped by social identity. The model shows how in-group bias can have different societal-level effects depending on the network structure. Specifically, in-group bias prevents consensus in highly random networks, but fosters consensus in highly clustered networks—as long as homophily and predisposition are not extreme.

The outcome of consensus or disagreement in the model depends on the interplay between two forces during social interactions: an assimilative force, acting to align opinions and narrow them down, and a conservative force, acting to preserve the original opinions. These two forces and their interplay are common in many opinion dynamics models (e.g. Axelrod, 1997; Edmonds & ní Aodha, 2019; Flache et al., 2017; Turner & Smaldino, 2018), but in contrast to most of these studies, we do not assume that agents fully ignore others (as e.g. in bounded confidence models Deffuant et al., 2000; Hegselmann & Krause, 2002) or that opinions diverge when their disagreement exceeds some threshold (as e.g. in Flache & Mäs, 2008; Mäs et al., 2010). In our model, agents biased with respect to social identity adapt more promptly to in-group members than to out-group members because they perceive in-group opinions as more conclusive than out-group opinions. This purely positive social influence implies that consensus is an inevitable outcome (as in other studies, such as Abelson, 1967; Bartels, 2002; Flache et al., 2017, with purely assimilative influences). However, to reach a consensus in a reasonable time, the agents need to remain sufficiently susceptible to opinion change. Disagreement can solidify and persist in our model only when agents are trapped in isolated opinion clusters and when those clusters turn into echo chambers, inside which the agents' opinions become increasingly narrow. This outcome is in line with empirical evidence that echo chambers play an important role in political debates (Boutyline & Willer, 2017; Cinelli et al., 2021) including the debate on climate change (Jasny et al., 2015; Williams et al., 2015). Moreover, disagreement can be a robust outcome of our model without opinion polarisation. This

pattern is consistent with survey-based research (Flores et al., 2022; Mason, 2015) showing that, for example, US-citizens are divided more along social identity lines than actual opinion differences.

Our model shows that, in the long run, in-group bias prevents consensus in societies characterised by networks with low clustering, i.e. with highly random link structures. In such networks, the average path length, especially between in-group members, is short, thus fostering in-group alignment. The bias exacerbates this effect by turning opinion clusters more easily into echo chambers and, thereby, consolidating disagreement between identity groups. This view is consistent with the notion that in-group bias is negatively associated with consensus building (e.g. Johnson & Levin, 2009; Lau & Murnighan, 1998; Mason, 2015). A similar effect of an identity-related bias was described in a previous modelling study (Flache & Mäs, 2008) with agents characterised by different demographic attributes that influenced whether they aligned to or repelled from the opinions of others. Similar to Flache and Mäs (2008), we find that a strong in-group bias generates disagreement but, in contrast to them, we obtain this result without including repulsive social influences between dissimilar agents.

The most important result of our study is that in-group bias can have opposite effects depending on the topology of the communication network. High clustering in the network inhibits consensus formation in general. However, in contrast to the result discussed above, the presence of in-group bias facilitates consensus in this case. In such highly clustered networks, the average path length between agents tends to be long, such that agents at opposite ends of the network exert little influence on each other. Consequently, local opinion clusters emerge, irrespective of social identities, which inhibit a society-wide consensus. In this case, opinions are separated by space and in-group bias helps to dissolve such clusters, thus, promoting consensus. This view is consistent with the notion that, under certain conditions, in-group bias is positively associated with consensus building (e.g. O'Connor & Weatherall, 2018) and with empirical evidence showing that concerns about climate change have increased faster among younger generations than older generations (Swim et al., 2022). In our model, the mechanisms by which the in-group bias leads to consensus are analogous to those considered by Mäs et al. (2013). Mäs et al. used the model of Flache and Mäs (2008) with the addition of homophily in relation to the way agents choose their interacting partners. The presence of in-group bias and homophily caused the agents to moderate extremists within their respective in-groups, which then promoted consensus among these moderate groups in the long run. Our results exhibit the same two-step process in highly clustered networks, but we also find that besides homophily and bias, this process can be driven solely by the topology of the communication network. The importance of network topology in relation to identity was recently observed also in a two-armed bandit game model (Fazelpour & Steel, 2022) in which agents learned from their peers but distrusted information from out-group members. Similar to our results, the (very simple) social networks considered by Fazelpour and Steel (2022) determined whether the bias prevented or promoted collective performance.

Although the positive effect of the bias on consensus formation in highly clustered networks is robust over a wide range of parameter configurations, there are some limitations in two extreme cases: strong homophily and marked predisposition. In highly clustered networks, homophily and predisposition affect opinion formation similar to in-group bias. They promote the convergence of opinions among in-group members that are dispersed over the network, thereby, homophily and predisposition can foster consensus. Although counter-intuitive at first sight, this result reflects the well-established concept that any process that prevents social learners from converging too quickly towards a local optima can improve the collective performance (Barkoczi & Galesic, 2016; Galesic et al., 2023; Smaldino et al., 2023). However, the combination of in-group bias with either

strong homophily or marked predisposition (or a combination of the two) reverses this positive effect and prevents consensus formation.

Social identities and related biases are particularly relevant in the debate on climate change (Clayton & Opotow, 2003; Pearson & Schuldt, 2018). As discussed before, climate change entails an intergenerational conflict (Meleady & Crisp, 2017; Ross et al., 2019; Swim et al., 2022), which is commonly addressed in the debate ("We young people [...] must hold the older generations accountable for the mess they have created [...]."—Greta Thunberg[3]). There is ample empirical evidence on the fact that social identity affects the way people perceive the opinions of others on topics like climate change (Esposo et al., 2013; Landrum et al., 2017; Mackie et al., 1992). This evidence suggests that it may be easier to dismiss alarmed statements from an out-group contact as 'merely their opinion', given that such opinions are subjective rather than rooted in logic- or evidence-based reasoning. Recent studies have proposed strategies to reduce the relevance of social identities and related biases in political debates, for example, by exposing people to trusted expert opinions, like those presented in the IPCC report (Flores et al., 2022), by emphasising non-political similarities between different groups, like connecting people with similar musical taste (Balietti et al., 2021), or by fostering contact between members of different groups to reduce prejudices (Hewstone et al., 2014). Analogously, other studies (e.g. Brown et al., 1999) have proposed strategies that instead aim to highlight social identity and exploit related biases in order to overcome disagreement in political debates. An example of such a strategy consists of channelling policy communication through in-group specific messengers, such as Greta Thunberg (Bavel et al., 2020; Fielding et al., 2020; Sabherwal et al., 2021). Another example is the use of large social identity groups to reinforce people's perceived efficacy (by the sheer size of the group) during global crises (Bavel et al., 2020; Masson & Fritsche, 2021).

Over the past years, the generational identity gap and the associated conflicts have arguably increased as youth movements surged in popularity across the globe (Gonyea & Hudson, 2020; Marris, 2019). Has this social identity contributed to a shared view of climate change among youngsters and beyond? Or have social identities and related in-group biases further polarised society? Answers to these questions remain elusive. Models, like the one we have presented here, are, by definition, approximations of reality. And although they cannot include the broad spectrum of processes and cognitive biases characterising our society, their simplicity enables us to test mechanisms that can generate different macroscopic patterns such as consensus or disagreement (Smaldino, 2020). Our model shows that in-group bias can have such opposing effects depending on the communication network characterising a society. While it is relatively well established that networks of social influence are somewhat affected by homophily (Colleoni et al., 2014), more fine-grained aspects of the network topology are arguably less clear. For example, connections between users of social media platforms such as 'Twitter' (now 'X'), are driven by identity, but the network comprises a mix of short- and long-range connections and local associations remain important (Herdağdelen et al., 2013). We show here that these aspects play a crucial role in the emergence of opinion patterns. Sparse long-range connections and strong local clustering generally impede consensus. Under such conditions, in-group bias can stimulate consensus formation. Contrastingly, the bias impedes consensus formation in networks with many long-range and less locally confined connections (similar, for example, to internet forums). The effects of in-group bias are thus moderated by network structure. This is important because real-world network structures are quite diverse, as the analysis of networks in social media platforms like 'Twitter' suggests (Bodrunova et al., 2019; Colleoni et al., 2014), which can lead to unanticipated communication and opinion patterns.

[3]CNN (@CNN) 'We must hold the older generations accountable for the mess they have created ... and say to them you cannot continue risking our future like this.' Teen climate activist Greta Thunberg calls on young people to use their anger as activism. *Twitter* (now *X*). 25 Dec 2018. https://twitter.com/CNN/status/1077444076176359426

Future research could focus on investigating more closely the network structures that are characteristic of certain real-world debates, and in particular their clustering and path length properties. This research may indicate which of the effects of biases related to social identity dominate in such debates. How heterogeneity in the degrees of in-group bias or differences in the agents' social power in the network can alter the effects of in-group bias could be another interesting avenue for future research. Although intergroup relations are a key factor in driving opinion formation in society and, thus, for example, in facilitating or hampering the design of effective climate change mitigation policies (Johnson & Levin, 2009; Pearson & Schuldt, 2018; Ross et al., 2019), a surprisingly low number of modelling studies have tackled the problem. Ultimately, by shedding new light on the contrasting effects that the interaction among social identity, in-group bias and network topology can have on opinion dynamics, we hope that our study can inspire additional modelling work in this field.

Data availability. We make the model available as open-source software (Steiglechner, 2023) so that it can be used, modified, and redistributed freely. Further analysis are provided in the electronic supplementary material, which is available online at https://doi.org/10.6084/m9.figshare.c.6960070.

Authors' contributions. P.S., A.M., and P.E.S. conceived and designed the research project. P.S. developed the model, wrote the code and analysed the results. P.S. wrote the initial draft of the manuscript and all authors contributed to review and editing. All authors gave their final approval for publication.

Conflict of interest declaration. We declare we have no competing interests.

Funding and acknowledgements. We are grateful to Achim Schlüter for insightful discussions about cognitive biases and their impacts on real-world social interactions. This work was supported by the German Research Foundation (DFG) though the project SEATRAC (project number 423711127) as part of the Priority Programme 1889: Regional Sea Level Change and Society (Sea Level).

Chapter 4

Perceived and actual opinion polarisation in the German climate change debate

Peter Steiglechner[1,2,*], Paul E. Smaldino[3,4], Agostino Merico[1,2]

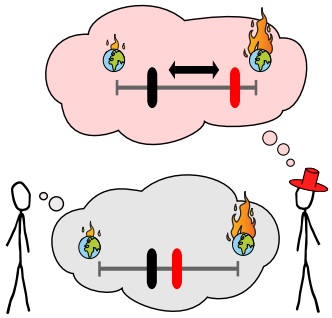
[1]Leibniz Centre for Tropical Marine Research (ZMT), Bremen, Germany

[2]Constructor University, Bremen, Germany

[3]Department of Cognitive and Information Sciences, University of California Merced, Merced, USA

[4]Santa Fe Institute, Santa Fe, USA

[*]*Corresponding author:* psteiglechner@constructor.university



This chapter contains a manuscript intended for submission to a peer-reviewed journal.

Abstract

The debate about climate change seems to have become increasingly polarised in the last decade. By driving public uncertainty and policy stalemate, polarisation constitutes a significant barrier to our efforts against the climate crisis. Many approaches to measure polarisation do not differentiate, however, between perceived and actual polarisation, and when they do, they often conflate ideological and affective polarisation. Here, we present a new method to measure perceived ideological polarisation by formalising mathematically how individuals perceive the opinions of others. We account for the fact that these perceptions co-evolve with opinions in society and depend on political identities. Applying this method to data on climate opinions among Germans, we find that, even in the absence of any affective or structural drivers, people perceive much more polarisation than their actual opinions suggest when they are sufficiently biased. Notably, perceived polarisation varies greatly among different partisan groups. Our method and analysis offer a fresh perspective on (mis-)perceived ideological polarisation and can be applied to simulated or empirical data covering various topics. Our conclusion—that Germans may overestimate the degree of ideological polarisation within the climate debate—is a relevant consideration for social cohesion in climate politics.

Keywords: *False polarisation, Perceived polarisation, Computational social science, Opinion dynamics, Climate change*

4.1 Introduction

Societal polarisation has received great attention in social, political, and scientific debates (Abramowitz & Saunders, 2008; Baldassarri & Page, 2021; Fiorina & Abrams, 2008; Levin et al., 2021; Waldrop, 2021) as it may be a key contributing factor to the escalation of global crises (World Economic Forum, 2024). In this study, we focus on societal polarisation concerning one particular topic: climate change. To what degree are peoples' attitudes towards climate change diverging? This question has been discussed broadly in the scientific literature (Dunlap et al., 2016; Falkenberg et al., 2022; Leiserowitz et al., 2021; McCright & Dunlap, 2011a; Mewes et al., 2024; Smith et al., 2024), in particular since climate grassroots movements, such as the 'Fridays for Future'-movement which was initiated by Greta Thunberg in 2018, surged in popularity and pushed the issue of climate change on top of the political agenda (Marris, 2019). Societal polarisation in the climate debate is especially critical because the effectiveness of measures to tackle it, that is to mitigate and adapt to the climate crisis, depends on social cohesion, stable political consensus, and a concerted effort by the public (Dunlap et al., 2016). However, polarisation is a subjective experience and we suggest in this study that people may perceive more polarisation on climate change than actually exists.

The term 'polarisation' is used quite ambiguously in the scientific literature and can have different meanings (Armaly & Enders, 2021; Judge et al., 2023; Mason, 2015). Ideological (or opinion) polarisation refers to a divergence of opinions. For example, some people become more concerned about climate change while others care increasingly less. But polarisation can also capture structural aspects, such as the increased clustering of opinions in social networks ('echo chambers') and the reduced communication between individuals of different

clusters (Cinelli et al., 2021; Hohmann et al., 2023). Moreover, affective polarisation describes the increase in negative feelings between certain groups, for example, the amplified resentment between respective supporters of Democrats and Republicans in the US, independent of their actual opinion differences (Druckman et al., 2021; Iyengar et al., 2012; Mason, 2015). In the following, we focus purely on the ideological aspect, i.e. opinion polarisation, and neglect structural and affective aspects (although we will comment on their interrelations later).

Quantitative measures of opinion polarisation are typically based on the distribution of opinions in society or changes of that distribution over time (Bramson et al., 2017; DiMaggio et al., 1996; Dunlap et al., 2016). However, both in the public and academic discourses, there is often no distinction between actual and perceived opinion polarisation (Judge et al., 2023; Westfall et al., 2015). Actual opinion polarisation represents a divergence of opinions in society, i.e. an increase of the differences between people's opinions, and only this can be measured directly from the distribution of opinions. Perceived opinion polarisation represents the divergence of opinions in society *as seen by the people themselves* (Enders & Armaly, 2019; Lees & Cikara, 2021; Levendusky & Malhotra, 2016; Westfall et al., 2015). Although seemingly subtle, this difference is important because (mis-)perceived polarisation does not necessarily require opinions to actually diverge. Indeed, much evidence suggests that perceived polarisation overshadows actual opinion divergence (Enders & Armaly, 2019; Lees & Cikara, 2021; Levendusky & Malhotra, 2016; Sherman et al., 2009; Westfall et al., 2015). For example, survey data analysed by Enders and Armaly (2019) showed only a small increase in the difference of actual opinions between Republican and Democrat supporters after 2000, but displayed a large increase in the perceived differences between the groups. When the respondents were asked what they believed the out-party's opinion to be, they overestimated the actual distance between in- and out-party. This mismatch extends to the issue of climate change in which people also systematically and substantially overestimate the degree of public disagreement on climate action in society (Andre et al., 2024). It is not obvious where this mismatch comes from. Ultimately, however, it is the perceived rather than the actual polarisation that influences political debates and the design of effective policies (Andre et al., 2024; Enders & Armaly, 2019).

The empirical studies that measured perceived polarisation (Armaly & Enders, 2021; Enders & Armaly, 2019; Lees & Cikara, 2021; Levendusky & Malhotra, 2016; Westfall et al., 2015) typically asked the participants in the surveys to report (i) their own opinion, (ii) the party or the ideology that they most affiliate with, and (iii) a respective rival party/ideology on the same one-dimensional scale. For example, these studies asked participants "What do you think the typical Democrat/Republican voter would want to happen to capital gains tax rates?" (Levendusky & Malhotra, 2016). This question, by design, comprises a variety of factors, such as misrepresentation of the actual opinions of others, negative feelings towards out-groups or the 'other party' (affective polarisation Armaly & Enders, 2021), politically motivated reasoning (Bayes & Druckman, 2021; Kahan, 2016), and anchoring biases by placing first themselves and then a rivalry group on the same scale (Levendusky & Malhotra, 2016). As a consequence, these surveys capture the perception of overall polarisation which confounds all of these factors. And, consequently, the studies are not suitable to distinguish between actual and perceived *ideological* polarisation, i.e. the polarisation that purely considers opinion differences (Druckman et al., 2021). We argue that a mismatch between perceived and actual opinion polarisation does not require drivers such as affective or motivated reasoning with respect to a rivalry political group, although such drivers may amplify the mismatch.

One alternative approach to understand (perceived) opinion polarisation and its drivers is mathematical modelling. By constituting a virtual laboratory, mathematical models can be used to investigate the social con-

ditions that foster the persistence of disagreement or drive opinion polarisation (Baldassarri & Page, 2021; Flache et al., 2017; Galesic et al., 2021; Levin et al., 2021; Sobkowicz, 2020; Waldrop, 2021). Some modelling studies suggest that the emergence of opinion polarisation depends critically on the structural properties of the underlying social network (Smaldino et al., 2023; Steiglechner et al., 2023), the initial asymmetry in opinions (Carpentras et al., 2022; Turner & Smaldino, 2018), the biases in the information processing of individuals, such as the confirmation bias (Dandekar et al., 2013) or biases related to social identity (Schweighofer et al., 2020; Steiglechner et al., 2023), and the noise in the communication process (Steiglechner et al., 2024; Turner & Smaldino, 2018). Although modelling is generally considered a vital approach to better understand the drivers of opinion polarisation (Bak-Coleman et al., 2021; Waldrop, 2021), opinion dynamics models usually only study the emergence of actual polarisation, i.e. the objective divergence of opinions, and disregard how people perceive opinions. They also tend to overestimate the frequency with which individuals change their opinions, thus making them questionable candidates to provide a realistic explanation of polarisation. If the perceived opinion polarisation is indeed far greater than the actual divergence of opinions—as the empirical evidence suggests—then models that neglect this distinction are ultimately investigating the wrong phenomenon (Sobkowicz, 2020).

Perception is inherently subjective. Social identity theory (Hewstone et al., 2002; Tajfel, 1974) suggests that people see themselves as part of groups, for example based on demographic attributes or affiliation to a political stream or party, and that they predominantly compare themselves in relation to their in-group members rather than the whole society. Social identities—and partisan identities in particular—thus provide a lens to perceive the political world (Bartels, 2002; Flores et al., 2022; Lüders et al., 2023; Macy et al., 2019; Powell et al., 2023). As partisan groups vary in their opinion distributions, its members develop different ways to perceive opinions (Abrams et al., 1990; Homer-Dixon et al., 2013)—their belief systems differ (Brandt & Sleegers, 2021). Consider, for example, the US debate on climate change: In the early 2000s, disagreement on climate change was common between Democrat and Republican voters. This changed around 2008, when the issue became a relevant political dimensions after the election of Barack Obama (Dunlap et al., 2016). Democrats (with a more homogeneous stance on climate change (Dunlap et al., 2016)) might perceive the political gap between a climate-change denier and a climate-change advocate larger than they would have perceived it in the early 2000s, because an individual's opinion on the subject had gained salience as an identity signal. Changes in the subjective perceptions may thus create a sense of polarisation even if the actual opinions do not diverge and in the absence of an external driver such as an increasing resentment of other parties or groups.

We pursue two goals in this study. First, we present a new mathematical description of how perceptions of opinions co-evolve based on partisan identities and use this method to understand how accounting for such dynamics may drive perceived opinion polarisation regardless of actual opinion divergence, affective drivers, or motivated reasoning (Section 4.2). Specifically, we assume that the opinion of an individual can be represented as a numerical value over a certain range—the opinion space—in line with most of the opinion dynamics literature (Flache et al., 2017; Noorazar, 2020). This space is generally high-dimensional, representing multiple topics or questions on which opinions can vary. To evaluate differences between opinions, individuals need a way to spatially represent the opinions of others. In contrast to the opinion dynamics literature (e.g. Schweighofer et al., 2020; Steiglechner et al., 2024) and many empirical studies (e.g. DiMaggio et al., 1996; Dunlap et al., 2016), however, we assume that this way to represent opinions is dynamic, because it co-evolves with the opinions themselves, and subjective, because individuals pay particular attention to the opinions of their in-group. Our second goal is to quantitatively estimate the mismatch between perceived and actual polarisation of opinions on climate change. Using Germany as an example, we apply our method to empirical data from the

European Social Survey European Research Infrastructure (ESS ERIC) (2017) wave 8 conducted in 2016/17 and ESS (2022) wave 10 conducted in 2021, described in detail in Section 4.3, which covers attitudes towards climate change among German citizens.

Our research questions are:

- Is there a mismatch between actual and perceived opinion polarisation in the climate change debate? If so, do people perceive more or less polarisation than the actual divergence of opinions suggests?
- How does the mismatch depend on the importance of social identities? Specifically, does a stronger identification with partisan groups, i.e. a more biased perception, amplify or attenuate the level of perceived opinion polarisation?
- Is the mismatch consistent for the different partisan identity groups?

After describing our results in Section 4.4, we discuss their implications for measuring polarisation and for the climate change debate and suggest how the ideas behind our method could be useful for theory development, survey or model design and the analysis of empirical or simulated data (Section 4.5).

To facilitate reproducibility and comparison with other survey data or simulation models, we make our method and analysis fully available as open source code[1]. The code can be readily applied to data from the European Social Survey (for any country and any topics) and can be easily modified to be applied to any survey with multi-wave cross-sectional or panel data in which the participants state (i) their opinions on one or multiple topics using a (relatively fine-grained) ordinal scale and (ii) a categorical identity group that they feel particularly close to, such as a political party, an ideology, or a demographic group. The method can be easily applied to simulation data obtained from opinion dynamics models that include some representation of a social identity.

4.2 Method

We present a method to disentangle actual from perceived opinion polarisation in empirical data. To define indices for perceived disagreement and perceived opinion polarisation, we operationalise (i) opinions as mathematical objects and (ii) how people represent the opinions of others and measure opinion differences based on their social identities. The opinion of an individual i at time t is a numerical vector, $x_i(t)$, in the m-dimensional opinion space, which represents m topics of interest. In the following, we will often omit the time argument, t, for better readability. This definition of opinions follows a large part of the literature on continuous opinion dynamics (Noorazar, 2020; Sobkowicz, 2020) and it has a very intuitive connection to surveys on political opinions that use ordinary scales, such as Likert scales. Since opinion polarisation is based on opinion differences, we define the distance seen by an individual i between the own opinion, x_i, and the opinion of another individual x_j at time t via the function $\delta\left(x_i, x_j | T_i\right)$. T_i denotes the subjective representation of the opinion space used by an individual i at time t (and $\mathcal{T}_t = \{T_j(t)\}$ denotes the set of representations for all individuals j). We will describe in the next paragraphs how these representations are shaped. Using these subjective distance functions, we define an index for perceived disagreement, d, as the average of the pairwise subjective distances between the opinions $\mathcal{X}_t = \{x_i(t)\}$ of all n individuals depending on their subjective representations $\mathcal{T}_t$ (a similar index was used by Enders and Armaly (2019)):

[1] Available at https://github.com/PeterSteiglechner/perceived-polarisation

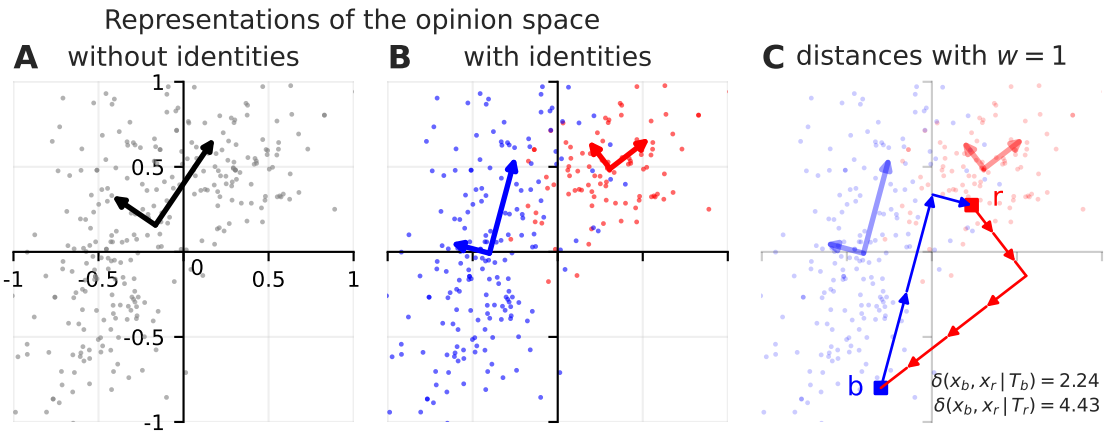


Figure 4.1: Illustrations of opinion spaces and opinion distances using fake data. (A) Sample opinions in a two-dimensional space and the corresponding inferred spatial representation of the opinion space given that the individuals have no identities (or no identity bias, $w = 0$). (B) The same opinions but for individuals that affiliate with a red or a blue identity group. In the case of large in-group bias, $w = 1$, there is a significant difference between the inferred spatial representations of the opinion space used by the red group, which is more homogeneous in their opinions, and the representation used by the blue group, which exhibits a larger variance, especially along the y-axis. (C) This affects, for example, the asymmetric distance seen by an individual r, who uses the red, and an individual b, who uses the blue spatial representation of the opinion space between their two opinions x_b and x_r: $\delta(x_b, x_r \,|\, T_b) = \sqrt{2^2 + 1^2} \approx 2.24$ and $\delta(x_b, x_r \,|\, T_r) = 4.43$.

$$d(t) := d(\mathcal{X}_t, \mathcal{T}_t) = \frac{1}{n} \cdot \sum_i \frac{1}{n-1} \cdot \sum_{j \neq i} \delta\left(x_i, x_j \,|\, T_i\right) \ . \tag{4.1}$$

Finally, we define perceived opinion polarisation between t_1 and t_2 as $P_{\text{perc}}(t_1, t_2) = d(t_2) - d(t_1)$, where $P_{\text{perc}} > 0$ implies perceived divergence of opinions and $P < 0$ implies perceived convergence. Note that here polarisation P_{perc} refers to a dynamic process—the increase in disagreement (see Dandekar et al., 2013; DiMaggio et al., 1996)—rather than a property of a society at a given time.

The focus of our study lies mainly on how individuals perceive distances between their own opinion and the opinions of others, i.e. the function δ and its dependence on the subjective representation of the opinion space, T_i, of an individual i. To measure distances between opinions, one needs to have a spatial representation of these objects. If such representations are objective, they are independent from the individuals and their social environment (e.g. friends, peer groups, or family with their respective opinions). In particular, the objective distance between opinions x_i and x_j is their standard Cartesian distance, $|x_i - x_j|^2 = (x_i - x_j)^T \cdot (x_i - x_j)$ in the m-dimensional space, and thus the same when measured by individual i or j. In reality, people have subjective perceptions of the opinion space which are shaped by their respective social environments. These perceptions can differ between individuals, i.e. $T_i(t) \neq T_j(t)$, and can vary over time if the opinions within the individual's social environment change, i.e. $T_i(t) \neq T_i(t')$, such that $\delta(x_i, x_j \,|\, T_i(t))$ may not coincide with $\delta(x_i, x_j \,|\, T_i(t'))$ or $\delta(x_i, x_j \,|\, T_j(t))$. We conceptualise these aspects according to the following assumptions: (i) the subjective representation of the opinion space of an individual i, $T_i(t)$, is shaped dynamically to best represent the distribution of the opinions in the individual's social environment and (ii) individuals have fixed social identities which separate them into distinct identity groups, implying that $T_i(t)$ depends primarily on the opinions of the in-group members of individual i.

The first assumption considers that spatial representations of opinions are derived from the (dynamic) distribution of opinions in society. For now, we assume that there is only one identity group such that individuals consider all others as in-group members. The way one represents the opinion space is, in principle, an arbitrary choice. For example, it comes quite natural to people to collapse multi-dimensional opinions, such as opinions on immigration, economic issues, or environmental protection, onto one primary axis describing the political spectrum between 'left' and 'right'. We assume here that individuals use this simple representation of the opinion space, which is derived from the present distribution of opinions. A space is characterised by its bases (e.g. the standard Cartesian space has the bases $(1, 0, 0, \ldots)$, $(0, 1, 0, \ldots)$, etc.). The bases of the dynamic and subjective representation of the opinion space are defined by a principal component analysis (PCA), such that the first principal component (PC1) represents the direction of the largest variance of opinions in society, the second principal component (PC2) represents the direction of the second largest variance and so on (Figure 4.1A). In particular, at a given time t, we derive the covariance matrix $cov(\{x_j\})$ of the opinions $\{x_j\}$ of all individuals j. From this, we obtain rotated axes pointing in the direction of the eigenvectors $v_{1\ldots m}$ of the covariance matrix (the principal components), and scale each axis by the square root of the respective eigenvalue, $\sqrt{\lambda_{1\ldots m}}$, which represents the variance explained by the corresponding axis (or the loading of the principal components). This procedure yields a distorted representation of the opinion space defined by m rotated and scaled axes as seen by an individual at time t (see Figure 4.1A for an example). We construct a matrix containing these distorted axes, which represent the base vectors of the distorted opinion space and refer to this as the subjective bases $T_i(t)$ of individual i:

$$T_i(t) = T^0(t) = \left(\sqrt{\lambda_0} \cdot \vec{v_0}, \ldots, \sqrt{\lambda_m} \cdot \vec{v_m}\right) \quad \text{with eigenvalues } \lambda_k \text{ and -vectors } v_k \text{ of } cov\left(\{x_j|\ j\}\right)\ . \tag{4.2}$$

In the case when individuals do not distinguish between in- and out-group members, they all use the same spatial representation of opinions, i.e. the matrix $T^0(t)$, at a given time t.

In reality, there is more than just one social identity group. Our second assumption considers that these identities play an important role in shaping the group's subjective representation of the opinion space. In general, people pay more attention to in-group members than to out-group members. Thus, we assume that in-group opinions are more dominantly represented in the individuals' subjective perceptions. We define the in-group bias parameter, $w \in [0, 1]$, which reflects the importance of social identities in political debates[2]. Zero in-group bias, $w = 0$, implies that social identities do not affect the spatial representation of opinions (equation 4.2 and Figure 4.1A). Maximum in-group bias, $w = 1$, implies that social identities fully determine the spatial representation of opinions (Figure 4.1B). The subjective bases T_i^1 for $w = 1$ is obtained by replacing $cov\left(\{x_j|j\}\right)$ with $cov\left(\{x_j|j \in \text{ in-group of } i\}\right)$ for an individual i in equation 4.2. For intermediate in-group bias, w, the subjective bases used by individual i is the weighted combination of the two extreme cases, $T_i^w(t) = w \cdot T_i^1(t) + (1 - w) \cdot T^0(t)$. In sum, at a given time t, the members of each identity group g use a distinct representation of the opinion space from the set $\{T_g^w(t)\} =: \mathcal{T}_t$ to perceive the opinions of others.

Distorted representations of opinions leads individuals to also see opinion distances in relation to this distorted space (Figure 4.1C). Specifically, individual i sees an opinion x (with coordinates x_1, x_2, ... in the standard space) as x' by transforming the original coordinates into the subjective bases spanned by the columns in T_i^w:

$$x = x'_1 \cdot \left(\sqrt{\lambda_1} \cdot \vec{v_1}\right)_j + x'_2 \cdot \left(\sqrt{\lambda_2} \cdot \vec{v_2}\right)_j + \ldots = T_i^w \cdot x' \qquad \Longleftrightarrow \qquad x' = (T_i^w)^{-1} \cdot x\,. \tag{4.3}$$

[2]We call this 'bias' here because, for $w > 0$, individuals weigh some opinions higher than others. However, we mean to imply no negative connotation with the term 'bias'. In fact, from a socio-psychological point of view, it may be very rational to weigh in-group members higher (Bayes & Druckman, 2021).

An individual i, thus, sees the distance between its own opinion, x_i, and the opinion of another individual, x_j, as:

$$\delta(x_i, x_j|T_i)^2 = (x_i - x_j)^T \cdot (T_i^w)^{-T} (T_i^w)^{-1} \cdot (x_i - x_j) \tag{4.4}$$

with $T_i^w = w \cdot T_i^1 + (1 - w) \cdot T^0$ for individuals with in-group bias w. Note that $T_i^w = 1$ the original bases of the m dimensional opinion space and $\delta(x_i, x_j|1)$ gives the standard Cartesian distance between the actual opinions.

Combining equations 4.4 and 4.1, we obtain the overall perceived disagreement d and, thus, a measure for perceived opinion polarisation dependent on the in-group bias w:

$$P_{\text{perc}}(t_1, t_2) = d(\mathcal{X}_{t_2}, \mathcal{T}_{t_2}) - d(\mathcal{X}_{t_1}, \mathcal{T}_{t_1}) \ . \tag{4.5}$$

Perceived opinion polarisation comes in part from the divergence of opinions, $\mathcal{X}_{t_2}$ vs. $\mathcal{X}_{t_1}$, which we will refer to as actual opinion polarisation:

$$P_{\text{actual}}(t_1, t_2) = d(\mathcal{X}_{t_2}, \mathcal{T}_{t_1}) - d(\mathcal{X}_{t_1}, \mathcal{T}_{t_1}) \ , \tag{4.6}$$

assuming that the individuals perceive opinions in the same, fixed way after t_1. We denote the mismatch between perceived and actual opinion polarisation as P_Δ. This mismatch (or the 'perception gap', Andre et al., 2024) accounts for changes in the individuals' subjective representations of the opinion space between t_1 and t_2, $\mathcal{T}_{t_2}$ vs. $\mathcal{T}_{t_1}$:

$$\begin{aligned} P_\Delta &= P_{\text{perc}}(t_1, t_2) - P_{\text{actual}}(t_1, t_2) \\ &= [d(\mathcal{X}_{t_2}, \mathcal{T}_{t_2}) - d(\mathcal{X}_{t_1}, \mathcal{T}_{t_1})] \ - \ [d(\mathcal{X}_{t_2}, \mathcal{T}_{t_1}) - d(\mathcal{X}_{t_1}, \mathcal{T}_{t_1})] \\ &= d(\mathcal{X}_{t_2}, \mathcal{T}_{t_2}) - d(\mathcal{X}_{t_2}, \mathcal{T}_{t_1}) \ . \end{aligned} \tag{4.7}$$

P_Δ captures perceived polarisation between t_1 and t_2 independent of any actual opinion change.

Before applying this method to empirical data, we describe how differences in the representations of opinions may affect the perceived disagreement using the fake data shown in Figure 4.1 as an example. Assuming maximum bias, $w = 1$ (as in Figure 4.1B), the subjective bases to represent opinions used by the blue and red identity groups differ quite strongly. Our method considers that perceived disagreement depends on the typical range of opinions in the individual's in-group and how far opinions tend to deviate from this range. For example, since the blue individuals are characterised by higher opinion variance along the y-axis compared to their opinion variance along the x-axis, they systematically weigh opinion differences in the y-direction less than opinion differences in the x-direction. As a consequence, the distance between two opinions is judged very differently from a red and a blue perspective (Figure 4.1C). Note that social identity has no effect beyond defining the spatial representation of opinions, i.e. the subjective bases $T_i^w(t)$ of individual i. Specifically, the distance between x_i and x_j seen by individual i is independent of whether individual j is an in- or out-group member. In sum, when an in-group becomes increasingly homogeneous in their opinions (along a certain, rotated axis), the subjective bases of the opinion space used by the group members to represent opinions of others are characterised by 'shorter' axes. Thus, even if the average objective distance between opinions remains the same, the group perceives a higher level of disagreement (see Appendix C.2.1 for an illustrative example of such a dynamic scenario).

4.3 Data

Our measure of perceived ideological (or opinion) polarisation can be applied to survey data of a representative sample of the population that captures (i) the respondents' opinions coded as numerical values and (ii) social identities that play a relevant role for the topic. Here, we use the German subset of the European Social Survey (ESS, 2017) from wave 8 in 2016/17 with $n = 2852$ and ESS (2022) wave 10 in 2021 with $n = 8725$ [3]. Focusing on climate change, we selected two survey questions: *wrclmch* ("How worried are you about climate change?") with response options on a five-point Likert-scale from 1 ("not at all worried") to 5 ("extremely worried") and *ccnthum* ("Do you think climate change is mainly natural/anthropogenic or both?") from 1 ("entirely by natural processes") to 5 ("entirely by human activity"). These two questions define a two-dimensional, discrete opinion space $x_i \in \{1, 2, 3, 4, 5\}^2$. Assuming that the discrete response options, $\{1, 2, \ldots\}$, correspond to the respective values on a continuous scale, $\mathbb{R}$, we can apply our measure to discrete data as well. We excluded those respondents from our analysis who did not give a valid answer to the questions, *ccnthum* and *wrclmch* (57 in wave 8 and 277 in wave 10), or who answered 'climate is not changing' for *ccnthum* (3 in wave 8 and 74 in wave 10). The participants' responses to the two questions are highly correlated (Pearson correlation coefficient $r = 0.46$).

Given that political orientation is a critical driver of climate change attitudes (Hornsey et al., 2016), we use affiliation with a political party as a salient social identity category. We consider only the six most popular German parties as identity groups: respondents identifying with (1) 'Union', the union of the conservative parties 'CDU' and 'CSU', (2) 'SPD', the social democrat party, (3) 'Bündnis 90/Die Grünen', the green party, (4) 'Die Linke', the left party, (5) 'FDP', the economic liberal party, and (6) 'AfD', the right-wing extremist party. We define partisan identity as follows: an individual affiliates with a partisan identity if his/her answer to the question "Is there a particular political party you feel closer to than all the other parties? Which one?" (*prtclede*/*prtclfde* in waves 8 and 10) was one of the six parties listed above and if his/her answer to "How close are you to the particular party?" (*prtdgcl*) was "very close" (1), "quite close" (2), or "not particularly close" (3). We excluded participants who affiliated with a party different than the six considered above (i.e. 59 participants in wave 8 and 177 in wave 10) or who refused to state the party (55 in wave 8 and 498 in wave 10). We classified the rest, i.e. individuals who did not feel close to a particular party (or feel "not at all" close to the stated party, i.e. response option 4 for *prtdgcl*), as individuals without in-group bias ($w = 0$) and refer to them as having the identity 'None'. This 'None' group contained roughly half of the respondents (see Appendix Table C.1 for the relative sizes of each identity group included in our analysis). Note that our analysis likely underrepresents the size of the right-wing extremist group, 'AfD': while $7.9\,\%$ of the eligible population voted for the 'AfD' in the 2021 national election ($10.3\,\%$ of the votes with a turnout of $76.6\,\%$), their group represents only $2.2\,\%$ of the ESS respondents in wave 10 (Appendix Table C.1). The final sample size used for the analysis is 2681 in wave 8 and 7807 in wave 10. To calculate representative averages over the sample population, we used the analysis weight recommended by the ESS (see equations (C.1) and (C.2) in Appendix C.2).

Figure 4.2 shows the degree of concern (*wrclmch*) among German respondents in the ESS in 2016/17 (wave 8) and 2021 (wave 10). If shown as an aggregate (Figure 4.2A), the respondents have visibly moved towards higher degrees of concern on climate change, representative of a collective cultural shift. If shown as separate distributions for each partisan identity group in wave 8 (Figure 4.2B) and wave 10 (4.2C), the interpretation of a collective cultural shift is less clear. In the following we show that our method of conceptualising (identity-

[3] In Germany, the ESS wave 10 was conducted online due to COVID-19 restrictions.

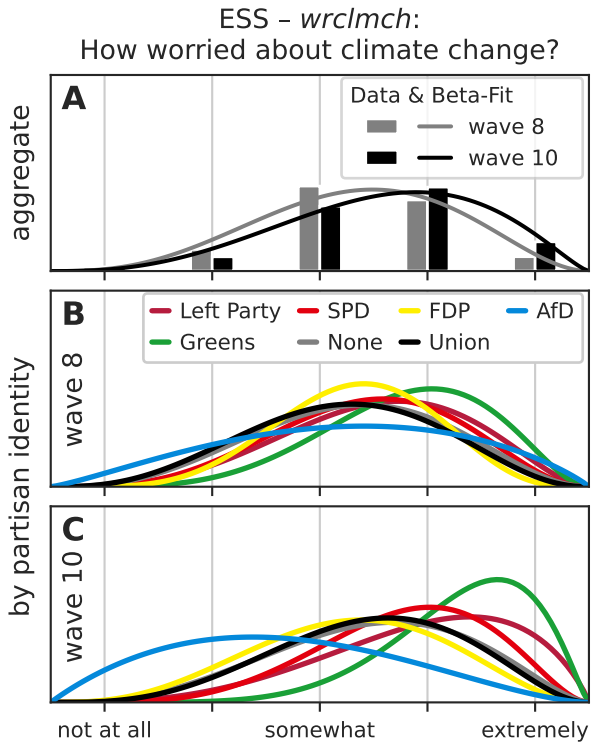


Figure 4.2: Frequency of responses to "How worried are you about climate change?" (*wrclmch*) among German respondents in the European Social Survey (ESS) in 2016/17 (wave 8) and 2021 (wave 10). Panel A compares responses in waves 8 and 10 shown as an aggregate histogram (bars) and the corresponding (normalised) beta-distribution fit over the continuous opinion space (grey and black lines). German citizens seem to increasingly worry about climate change. Panels B and C show the data separately for each partisan identity group (coloured lines). While the (normalised) distributions are quite similar across groups in wave 8 (panel B), both the mean and the shape of the distributions differ more significantly across groups in wave 10 (C).

biased) perception captures these contrasting views and has implications for the level of perceived opinion polarisation.

4.4 Results

4.4.1 Subjective partisan representations of the opinion space

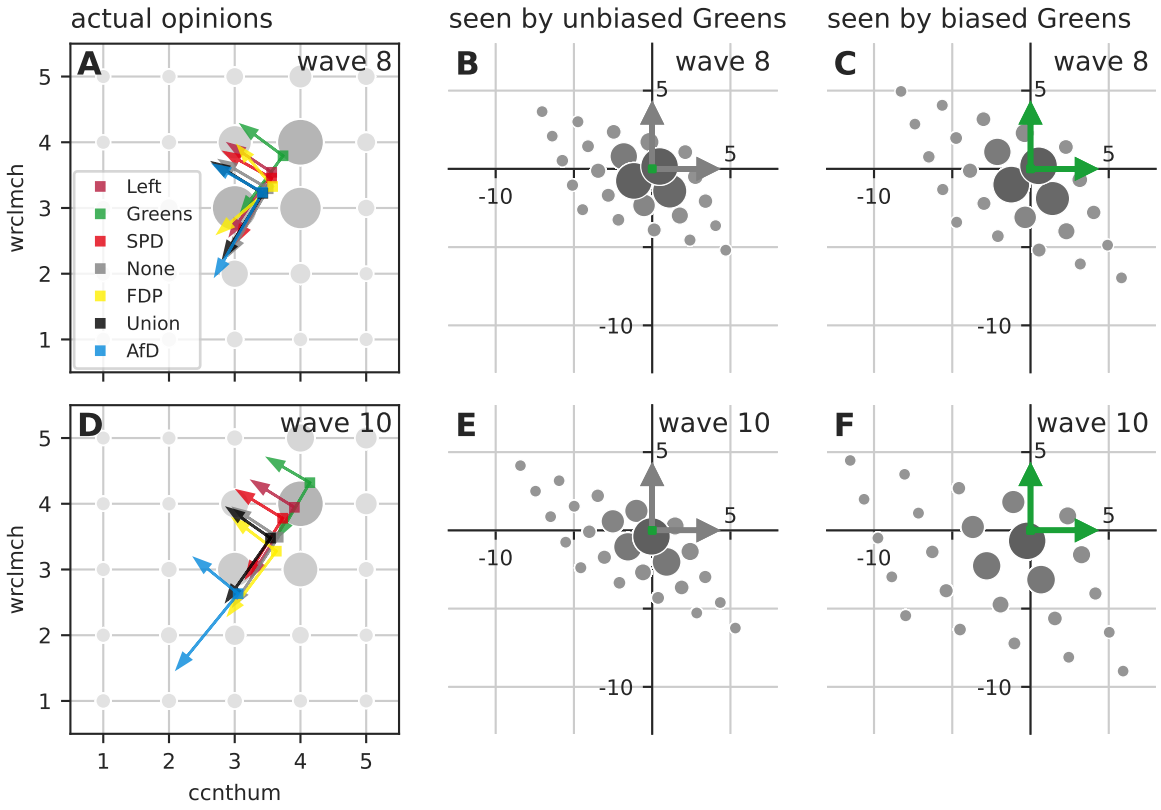


Figure 4.3: Distributions of climate-related opinions seen from different perspectives. Panels A and D show the reported opinions on the two-dimensional 5×5 grid of possible response options to the questions *ccnthum* and *wrclmch* in wave 8 and wave 10, respectively. The sizes of the grey circles represent the relative frequencies of the answers of all respondents. The coloured arrows, located at the mean opinion of each partisan identity group, indicate the subjective bases that each group uses to represent opinions in the case of maximum in-group bias $w = 1$. Note that participants classified as having no partisan identity (labelled as 'None' and comprising roughly half of the participants) are assumed to have no in-group bias, $w = 0$. That is, they use bases (indicated as grey arrows) to represent opinions that are inferred from the opinions of all individuals, rather than of one identity group. Changes in the bases from wave 8 to wave 10 affect the perception of opinions. For example, panels B, C, E and F show the same opinion distribution as panels A and D but seen from the perspective of an average individual with a 'Green' identity. When the Greens exhibit no in-group bias, $w = 0$, they represent opinions in the identical way as the 'None' group (panels B and E for wave 8 and wave 10). When the Greens are characterised by maximum in-group bias, $w = 1$, they represent the opinions using their subjective bases, which depend only on the opinions of the other Greens (panels C and F for both waves). The perceived distribution of opinions becomes notably broader between wave 8 (panel C) and wave 10 (panel F), indicating that the Greens perceive high levels of polarisation in the case of maximum in-group bias. Appendix Figure C.2 shows the opinion distributions (and the corresponding subjective bases) for all of the seven identity groups separately for both waves.

Figure 4.3 shows the distribution of opinions in the original space spanned by the two survey questions (*wrclmch* vs. *ccnthum*) and the different subjective bases to represent the opinion space as seen by each partisan group, assuming maximum in-group bias $w = 1$ (except for the group 'None', which has no partisan identity and is thus fixed to $w = 0$). In wave 8 (2016/17), the mean opinions (square points) and the respective subjective bases (arrows) are strikingly similar across the groups. Five years later, in wave 10 (2021), the mean opinions

of the groups have diverged somewhat and the subjective base vectors are slightly rotated and scaled compared to those in wave 8. Most notably, the subjective bases used by those identifying with the 'Green' party have become shorter—with lengths 0.55 and 0.84 in wave 8 and lengths 0.50 and 0.74 in wave 10—indicating that the Greens have become more homogeneous about the two questions in general. The bases of the right wing extremist group, 'AfD', conserve their lengths (0.61 and 1.21 in wave 8 and 0.61 and 1.23 in wave 10). Although the changes in the subjective bases between wave 8 and 10 appear small, they can affect how individuals perceive the opinion changes over time. To illustrate this, we show how the average 'Green' individual perceives the opinion distribution in wave 8 and wave 10 in the case of no in-group bias, that is when they are identical to the 'None' group in terms of opinion representation (Figure 4.3B and 4.3E), and in the case of maximum in-group bias (Figure 4.3C and 4.3F).

4.4.2 Perceived opinion polarisation

Table 4.1: Actual and perceived opinion polarisation and their mismatch for no in-group bias ($w = 0$) and maximum in-group bias ($w = 1$).

	subjective bases, $w = 0$		
opinions	$\mathcal{T}_8$	$\mathcal{T}_{10}$	P_Δ
$\mathcal{X}_8$	1.71	—	—
$\mathcal{X}_{10}$	1.76	1.74	-0.02
P_{actual}	0.05	—	$P_{\text{perc}} = 0.03$

	subjective bases, $w = 1$		
opinions	$\mathcal{T}_8$	$\mathcal{T}_{10}$	P_Δ
$\mathcal{X}_8$	1.73	—	—
$\mathcal{X}_{10}$	1.79	1.82	0.03
P_{actual}	0.06	—	$P_{\text{perc}} = 0.09$

Our method, we remind the reader, aims at quantifying the perceived opinion polarisation, P_{perc}, in the German climate debate depending on the in-group bias of partisan individuals. In all cases, we obtain positive values for P_{perc} between wave 8 and wave 10 (Table 4.1)[4]. We find that individuals in a society characterised by maximal in-group bias ($w = 1$) perceive three times more polarisation, $P_{\text{perc}} = 0.09$, than those in a society without in-group bias ($w = 0$), $P_{\text{perc}} = 0.03$. As described in Section 4.2, $P_\Delta = P_{\text{perc}} - P_{\text{actual}}$ measures the mismatch between perceived and actual opinion polarisation. If the subjective representations of the opinion space do not change, there is no mismatch, $P_\Delta = 0$ and $P_{\text{perc}} = P_{\text{actual}}$. In our data, however, the perceived opinion polarisation deviates from the actual opinion polarisation, as shown in Table 4.1, and the mismatch can have a similar order of magnitude as P_{actual}. Interestingly, the direction of the mismatch depends on the degree of in-group bias characterising the individuals with a partisan identity (roughly half of the respondents). The actual opinion polarisation is nearly equivalent regardless of in-group bias ($P_{\text{actual}} = 0.05$ and for $w = 0$ and $P_{\text{actual}} = 0.06$ for $w = 1$). Without in-group bias, $w = 0$, the changes in perception of opinions between wave 8 and wave 10 attenuate the perceived polarisation with $P_\Delta = -0.02$ and $P_{\text{perc}} < P_{\text{actual}}$. With strong in-group bias, the changes in perception amplify the perceived polarisation, $P_{\text{perc}} > P_{\text{actual}}$, for example, $P_\Delta = 0.03$ for $w = 1$. We find that the perceived opinion polarisation P_{perc} (as well as P_Δ and P_{actual}) remains relatively constant even if we randomly exclude up to a half of the participants from our analysis (Appendix Figure C.4), thus attesting to the robustness of the estimates.

[4]Note that we also find a positive polarisation index for $T_i = 1$, i.e. assuming that perception is objective and fixed, but the magnitude of this objective polarisation can not be directly compared to the perceived polarisation because it depends on the conversion from survey response options to numeric values.

4.4.3 Perceived opinion polarisation by identity group

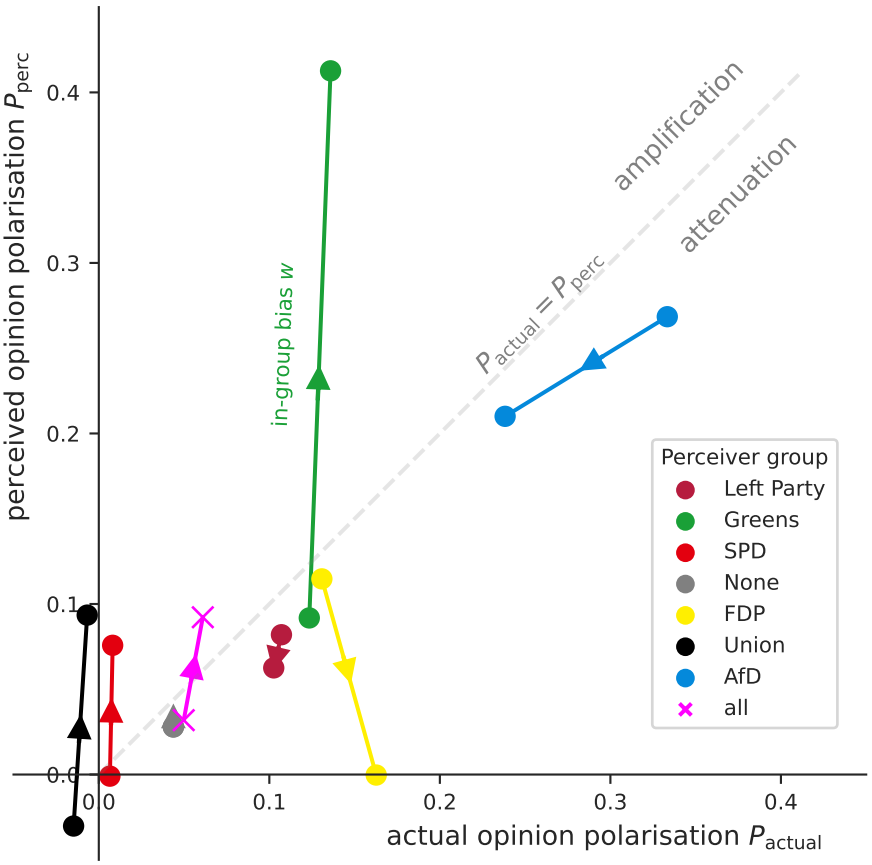


Figure 4.4: Perceived opinion polarisation, P_{perc} (y-axis), vs. actual opinion polarisation, P_{actual} (x-axis), in the German climate debate between wave 8 (2016/17) and wave 10 (2021) data of the ESS from the perspective of different partisan identity groups. For each group, the two dots denote the perceived/actual opinion polarisation (averaged over the group) in the extreme cases of no in-group bias, $w = 0$, and maximum in-group bias, $w = 1$. The arrows connecting the dots indicate the direction of increasing bias w. The diagonal (grey dashed line) represents the case in which perceived and actual opinion polarisation match, $P_{\text{perc}} = P_{\text{actual}}$ and $P_{\Delta} = 0$. For $P_{\text{perc}} > P_{\text{actual}}$ (upper left), the group members perceive, on average, amplified levels of opinion polarisation. For $P_{\text{perc}} < P_{\text{actual}}$ (lower right half), group members perceive attenuated levels of polarisation. The magenta crosses (and arrow) indicate the perceived/actual opinion polarisation for the whole society as shown in Table 4.1. Appendix Figure C.3 shows the relation between perceived and actual opinion polarisation of all pairs of identity groups, for example, the degree of opinion polarisation between Greens and Liberals as seen by the Greens and as seen by the Liberals.

In the previous subsection, we have considered perceived opinion polarisation as a single index for the whole society. Figure 4.4 shows the degrees of perceived and actual opinion polarisation as seen by each partisan identity group separately. If perceived and actual opinion polarisation match ($P_{\Delta} = 0$), all the points should lie on the diagonal, $P_{\text{perc}} = P_{\text{actual}}$. However, this is not the case and the direction and magnitude of P_{Delta} for the society as a whole depends on the in-group bias (Table 4.1 and magenta crosses in Figure 4.4). This is also true for each group individually. In the absence of in-group bias, $w = 0$, the perceived opinion polarisation is consistently smaller than the actual opinion polarisation for all groups—the attenuating effect of perception $P_{\Delta} < 0$ under small in-group bias is shared by all groups. In the presence of strong in-group bias, $w \to 1$, however, the mismatch between perceived and actual opinion polarisation, P_{Δ}, exhibits significant variations in both magnitude and direction between the groups. For $w = 1$, the Greens stand out for perceiving the

highest level of polarisation among all groups, far surpassing their perceived polarisation with $w = 0$ (see also Figure 4.3). This phenomenon arises mainly from the narrowing of the subjective bases used by the Greens to represent the opinion space under stronger in-group bias, rather than from an actual divergence of opinions from the Greens. In fact, actual divergence only makes up for roughly a third of the perceived polarisation by the Greens for $w = 1$. The Social Democrats ('SPD') and the Conservatives ('Union') also perceive higher levels of polarisation with increasing in-group bias. In particular, the Conservatives would see opinions converging towards them in the absence of in-group bias, $P_{\text{perc}} < 0$ for $w = 0$. However, in the presence of strong in-group bias, they see a positive opinion polarisation even though $P_{\text{actual}} < 0$. The in-group bias has a negligible effect on the perceived opinion polarisation seen by those affiliating with the Left party or the large group that does not affiliate with any party ('None'). Finally, the Liberals ('FDP') and especially the Right-wing Extremists ('AfD') are characterised by the highest levels of actual opinion polarisation. Yet, in the case of strong in-group bias, they perceive notably less opinion polarisation—in-group bias attenuates the polarisation (in contrast to the effect of in-group bias on opinion polarisation perceived by the Greens, the Social Democrats, and the Conservatives). Strikingly, Liberals perceive no polarisation for $w = 1$, $P_{\text{perc}} \approx 0$, despite their relatively high level of actual opinion divergence. In sum, while the in-group bias amplifies the overall perceived polarisation, this can vary substantially for the different groups: the bias has a much stronger amplification effect on some identity groups (e.g. the Greens) and a strong opposite attenuation effect for others (e.g. the Liberals). Consequently, perceived opinion polarisation is quite ambiguous when compared on a group-level.

4.5 Discussion

We have presented a new method to estimate ideological (or opinion) polarisation in a society. Crucially, and in contrast to many other studies on polarisation, our measure factors in how people perceive opinions and how this perception is driven by social influence and social identity. This allows us to disentangle perceived from actual ideological polarisation and, thus, to quantify their mismatch. We applied the measure to empirical data of opinions on the fundamental characteristics of climate change among German citizens in 2016/17 and 2021. Although there is some actual divergence of opinions, the polarisation that people perceive can be much higher (here, by a factor of 1.5), given that partisan individuals use this identity as a lens to perceive political opinions. We find in addition that the direction and magnitude of the mismatch between perceived and actual opinion polarisation varies for the different groups. Notably, those identifying with the 'Green' party perceive a much higher ideological polarisation than the actual opinions suggest, in line with empirical evidence (Herold et al., 2023). In contrast, those identifying with the right-wing extremist party ('AfD') or the liberal party ('FDP') perceive less polarisation on climate related questions than the actual divergence of opinions would suggest, owing to the fact that climate change is less central to these groups' respective identities or that they have become indifferent or flexible about the topic, which is in line with other empirical evidence (Falkenberg et al., 2022). In general, our results suggest a more nuanced view of ideological polarisation in the context of climate change and indicate a possible driver for the mismatch between perceived and actual opinion polarisation.

Our method lies somewhere between a fully relational, complex system representation of opinions and a (traditional) spatial representation of opinions (see Homer-Dixon et al., 2013, for an overview of these two contrasting conceptualisations). Although we use a spatial framework, the way individuals perceive this space is shaped by (in-group) opinions and is thus a product of the social environment. This is more in line with

a relational approach like 'Belief Network Analysis' (BNA; Boutyline & Vaisey, 2017), which assumes that belief systems are defined by the correlations between different belief dimensions in socio-political surveys. Our method, however, goes beyond BNA in four important ways. First, opinion dimensions are not fixed a priori but are a product of the opinion distribution, which partly removes the arbitrariness in the selection of such dimensions in the surveys. Second, we assume that perceptions are dynamic and co-evolve with the opinions. The more a group converges along one dimension, the more important differences in this dimension are to its members. This co-evolution of opinions and (relational) belief system has also been highlighted as an important but understudied factor in social dynamics (Brandt & Sleegers, 2021; Galesic et al., 2021). Third, besides merely describing a belief system, our method provides a straight-forward way to measure perceived opinion differences and polarisation (whereas BNA typically measures internal cognitive consistency, e.g. Dalege et al., 2018). Fourth, we allow different (identity) groups to have different perceptions or belief systems (in line with Baldassarri & Goldberg, 2014; Brandt & Sleegers, 2021). While BNA highlights statistical cleavages in society (Brandt & Sleegers, 2021)—a belief dimensions is more central if it explains a lot of the between-individual variance, such as the liberal–conservative dimension for US citizens (Boutyline & Vaisey, 2017)—our approach focuses on within-group similarities of opinions. Consequently, an opinion can be important for the perceptions of a group irrespective of whether other groups have a substantially different position. Using our concept of subjectively perceived opinion space polarisation, we have demonstrated that polarisation occurs not only when opinions diverge along a particularly central dimension, as suggested by BNA (Borsboom et al., 2021; Dalege et al., 2018), but additionally also when perceptions of opinions (i.e. belief systems) change differently for different partisan identity groups, even if these changes appear to be relatively small.

Our assumption that perceptions (or belief systems) vary across individuals is closely represented by an extension of belief networks developed by Carpentras et al. (Carpentras et al., 2021; Lüders et al., 2024). Their framework—'Response Item Network (ResIN)'—infers belief networks by determining the correlations between all possible belief item responses from social surveys (i.e. all the Likert-scale options of belief item 1 with all the options for belief item 2 and so on) rather than correlations between the belief items themselves, as in BNA for instance. For example, a strong concern about climate change might be correlated to the belief that climate change is mostly human-made, but a weak concern about climate change might be fully uncorrelated to what one believes about the causes of climate change. Our method captures this asymmetry in a similar way by allowing different identity groups to hold different representations of the opinion space such that actual opinion distances mean something different for each identity group. Lüders et al. (2024) applied the ResIN framework to US survey data and included a Democrat or Republican identity as one of the correlates (i.e. one of the opinion dimensions). They found that identity groups can be confidently distinguished by their belief item-response networks, which confirms our core argument that identity groups have subjective (and dynamic) perceptions of opinion distances.

There are many ways to measure actual opinion divergence (Bramson et al., 2016; DiMaggio et al., 1996; Dunlap et al., 2016), but our results show that it is crucial to take the perceptions of opinions into account, which echoes the results of many empirical studies on ideological polarisation (e.g. Enders & Armaly, 2019; Lees & Cikara, 2021; Levendusky & Malhotra, 2016; Sherman et al., 2009; Westfall et al., 2015) and, in particular, how identity shapes these perceptions, which echoes also empirical evidence for identity-related biases in human information processing (e.g. Flores et al., 2022; Macy et al., 2019; Powell et al., 2023). The surveys that distinguished between perceived and actual polarisation so far, however, all asked participants explicitly to estimate the opinions of the out-group and thereby conflated affective and ideological polarisation. While affective evaluations may be closely connected to ideological differences (Armaly & Enders, 2021), the

approach of these surveys tells us little about whether and how much individuals really mis-perceive *ideological* polarisation (Baldassarri & Page, 2021).

Our method disentangles perceived from actual ideological polarisation irrespective of affective or politically motivated drivers. Our assumptions, however, need empirical support. How do people really see opinion differences and how do they evaluate them in the context of polarisation? One way to investigate this would be to conduct a survey in which participants are presented with opinion profiles of prototype individuals in the context of the climate change debate (e.g. a dismissive, a disengaged, and an alarmed person with stereotypical opinions on the causes of climate change and the level of concern about it; see Leiserowitz et al., 2021) and then ask the participants to place these prototypes on a continuous (multi-dimensional) opinion map (similar to Keijzer et al., 2024; Sharman & Howarth, 2017) in order to evaluate spatial distances. Conducting such surveys over multiple years would allow to investigate whether perceptions and evaluations change throughout social and political debates. This survey design could directly validate (i) how relevant certain opinion dimensions are in the perception of ideological polarisation, (ii) whether this centrality differs between identity groups (i.e. whether there are significant asymmetries in the evaluation of opinion distances between different partisan groups), and (iii) whether these evaluations change over time—all aspects that are suggested to be present by our analyses.

Our study is also relevant for the analysis and interpretation of computational models of opinion dynamics. While many of these models aim to better understand the emergence of ideological polarisation, they neglect that the changes in peoples' opinions may not be the (only) driver of apparent ideological polarisation, but changes in their perceptions may be similarly important. Modellers could apply our method to analyse the opinion patterns obtained from their simulations allowing for subjective perceptions of opinions and the co-evolution of opinions and perceptions. There are some models (e.g. Dalege et al., 2023; Friedkin et al., 2016; Galesic et al., 2021; Huet & Deffuant, 2010) that account for relations between different belief dimensions. Yet, these belief systems are typically assumed fixed a priori rather than co-evolving and the studies are more focused on how people can navigate the compromise between internal consistency and (external) social influence. Our method considers both changes in opinions and perceptions, but we are agnostic as to how these opinions change. We thus believe that there is a lot of merit in combining this complexity perspective of dynamic, identity-driven perception with opinion dynamics modelling (Homer-Dixon et al., 2013) in order to increase the plausibility of models aiming to explain perceived ideological polarisation.

While our method can be applied to study polarisation on any political topic (or set of topics), we have focused here on how German citizens might overestimate their disagreement on climate change. Attitudes towards climate action have shifted substantially in the recent past (Andre et al., 2024; Leiserowitz et al., 2021; Pew Research Center, 2022). Yet, perceiving the political world through a partisan identity lens—a realistic assumption with rich empirical support (e.g. Flores et al., 2022; Macy et al., 2019; Mason, 2015; Powell et al., 2023)—can still exaggerate the level of perceived ideological polarisation. Our analysis covers climate change opinions in Germany before and after the emergence of the activist group 'Fridays for Future' in 2019, which likely contributed to politicise society with respect to climate change. We speculate that ideological, as well as affective, polarisation and partisan identity are today even more intertwined in the German climate change debate given that the support for the right-wing extremist and climate-change sceptic party 'AfD' nearly doubled in recent election polls (infratest dimap, 2024) and the debate seems to have become increasingly heated following more radical forms of protests, such as the street blockades by the climate activist group 'Letzte Generation' (Last Generation) in 2022/23. One of the dangers of perceiving greater ideological differences with respect to climate change attitudes is that it may lead to a spiral of collective

inaction as people will cooperate less if they believe that others do too (Andre et al., 2024). Another danger is that perceived ideological polarisation may reinforce affective and structural polarisation and vice versa, thus creating a positive feedback loop of amplified polarisation (Baldassarri & Page, 2021).

Our results, however, also entail an optimistic outlook. There is likely less ideological polarisation on the fundamental aspects of climate change in Germany than what people perceive. This might be an encouraging sign for policy makers trying to establish effective measures to tackle climate change. Our method highlights one way to foster social cohesion by reducing the importance of identity on peoples' perceptions. This might involve, for example, blurring identity signals (e.g. by removing hashtags that act as identity signals in social media; Powell et al., 2023), weakening the notion that partisan affiliation represents a salient group identity (Sherman et al., 2009) or strengthening shared identities (e.g. the recent alliance of the German climate activists from 'Fridays for Future' with local transport workers from the labour union 'ver.di'; Lucht & Liebig, 2023) in order to establish pro-climate norms across different groups (Hornsey & Lewandowsky, 2022). What is also reassuring for climate advocates is that our results confirm previous findings (e.g. Falkenberg et al., 2022; Jenkins-Smith et al., 2020; Lüders et al., 2024) that those partisan groups with more sceptical or unconcerned opinions on climate change, may actually not put that much emphasis on disagreeing with others—climate change attitudes are not very central to their political identity. This understanding of identity-driven perception and its consequences for perceived ideological polarisation might prove useful in the difficult task to foster climate action while simultaneously maintaining social cohesion.

Data accessibility. The model is coded in python[5]. To foster reproducibility, transparency, and flow of ideas, we make the code publicly available at github.com/PeterSteiglechner/perceived-polarisation.

Authors' contributions. P.S.: conceptualisation, formal analysis, visualisation, writing—original draft, writing—review and editing; P.E.S.: conceptualisation, writing—review and editing; A.M.: conceptualisation, project administration, writing—review and editing.

Conflict of interest declaration. We declare we have no competing interests.

Funding. P.S. and A.M. acknowledge funding from the German Research Foundation (DFG) through the project SEATRAC (project number: 423711127) as part of the Priority Programme 1889: Regional Sea Level Change and Society (SeaLevel).

[5]Python Software Foundation. Python Language Reference, version 3.9.5. Available at http://www.python.org.

Chapter 5

Summary and Concluding Remarks

5.1 Summary

In the present thesis, with the support of co-authors, I presented and analysed mathematical models of opinion formation and perception with the aim of better understanding when and how certain distortions of perception produce disagreement, polarisation or consensus in society. The studies build on previous works in opinion formation modelling, but they extend these works with a 'post-truth' perspective, providing novel insights into the non-trivial consequences of these distortions for collective opinion dynamics in social environments. In Chapter 2, I focused on noise in communication and its interplay with confirmation bias. Noise can induce consensus, but different types of noise induce quite different patterns. In particular, applying the model to the climate change debate, I find that the chance of reaching a pro-environmental agreement is high when agents are moderately biased and exchange messages with a specific, relatively narrow amount of ambiguity. This does not occur with other types of noise. In Chapter 3, I focused on social identity and the impact of related in-group biases on individual perception and, subsequently, on collective opinion patterns. In-group bias can foster or can impede consensus, depending on the network structure: in highly random networks, the bias impedes consensus; in highly clustered networks, the bias fosters consensus. This positive effect on consensus formation is surprising because an assimilation pressure that is biased towards in-group members tends to keep groups apart (especially in combination with homophily) rather than to align their opinions. The effect emerges because the bias helps to spread an opinion across a dispersed society by fostering consensus within one identity group. Finally, in Chapter 4, I focused on the mismatch between perceived and actual ideological polarisation. I developed a method to measure perceived polarisation, assuming that individuals perceive opinions (and distances between opinions) subjectively and that these perceptions change over time. In particular, the way an individual perceives opinions is shaped predominantly by the (co-evolving) opinions in the individual's in-group. Applying this framework to data on the German climate change debate, I find that perceived polarisation is up to 1.5 times greater than actual ideological polarisation, assuming maximal in-group bias of German citizens that identify with different political parties. This shows that amplified perceived polarisation does not require affective or structural drivers. Interestingly, the effect varies both in magnitude and sign across the different partisan groups. This is important because it is the perceived rather than the actual polarisation that undermines social cohesion and reinforces affective polarisation between partisan groups with potentially devastating consequences for effective policy making on climate change.

Taken together, the results imply that cognitive biases and noise, which both distort the way individuals perceive social influence, can have non-trivial and surprising effects on collective opinion patterns. The studies highlight that evaluations about biases, such as the negative assessment of cognitive biases in "The Tragedy of Cognition: Psychological Biases and Environmental Inaction" by Johnson and Levin (2009), may be too hasty and can actually depend on many other aspects, including the level of ambiguous communication (Chapter 2), the topology of the interaction network (Chapter 3), or the distribution of opinions among different partisan groups (Chapter 4). The studies demonstrate, in particular, that a polarisation of opinions, which is a great concern to many people and experts in the current social and political debates, may appear larger than it is in reality (Chapter 4) and may not be necessarily amplified by cognitive biases (Chapters 2 and 3).

This notion—that seemingly adverse mechanisms like bias or noise may have surprisingly diverse and complex effects—is fully consistent with recent literature suggesting that mechanisms initially delaying a convergence of opinions within a group may improve the chances of that group to successfully reach consensus in the long run (Bak-Coleman et al., 2021; Galesic et al., 2023; Smaldino et al., 2023). Mechanisms responsible for this include psychological factors (such as cognitive biases when adapting to others' opinions; Boroomand & Smaldino, 2023; Gabriel & O'Connor, 2024), properties of the communication channels (such as noise in the transmission of signals between the agents; Boroomand & Smaldino, 2023; Kahneman et al., 2022), or group structure (such as sparse social networks or networks with inefficient, dispersed topologies; Keijzer & Mäs, 2021; Moser & Smaldino, 2023). The chapters in this thesis contribute to this literature by strengthening the evidence for the notion that distorted perceptions are non-trivial, complex factors in collective opinion formation and illustrating further mechanisms through which such distortions can enhance agreement or disagreement among individuals under certain conditions.

5.2 Challenges in modelling opinion dynamics

Compared to the more traditional methods adopted in the social sciences, mathematical modelling of opinion dynamics is a relatively new field. It should be no surprise that mathematical modelling faces a number of challenges that limit its broader integration into the social sciences (Sobkowicz, 2020). These challenges include: (i) opinion dynamics models often struggle to distinguish between different types of opinions and the respective processes acting on them, (ii) models tend to exaggerate the frequency and amplitude of opinion changes, (iii) modellers are influenced by their own cultural histories and biases, and (iv) model validation or forecasting is made difficult by the fact that most models neglect crucial features of opinion formation in real environments. While some of these shortcomings are inherent to the studied social system and are unlikely to be eliminated, others can be addressed by developing, refining, and comparing different models or by further exploring the integration of empirical data. In the remainder of the section, I will describe the challenges listed above in more detail because I deem these as particularly relevant to the general topics of the present thesis.

Opinions are, in general, extremely vague concepts. For example, some opinions are related to value-based assessments (such as the importance of nature), some describe beliefs about the state of the world (such as whether anthropogenic greenhouse gas emissions cause climate change), some describe preferences (such as a preference for carbon emission taxation over investment in green energy), and some others describe affective judgements (such as anger about an offensive statement). Social influence affects different types of opinions in different ways (Kendal et al., 2018). For example, value-based opinions may be less susceptible to social influence than emotionally driven opinions regarding a particular topic. Moreover, different types of opinions

also depend on each other (Dalege et al., 2016). A person's concern about climate change may depend on the value that person ascribes to the environment. Opinion formation models have largely ignored this variety of possibilities (for some exceptions, see Dalege et al., 2023; Sobkowicz, 2012). In the present thesis, I formalised an opinion on climate change as a single numerical value on a continuous, bounded scale (Chapter 2), as a vector of such values representing different, interrelated opinion dimensions (Chapter 4), and as a distribution over a continuous one-dimensional space (Chapter 3). While these formalisations capture some aspects of the complexity of opinions, such as the interrelation of opinion dimensions in Chapter 4 or opinion strength and uncertainty in Chapter 3, they inevitably neglect others, such as the differences in types of opinions and their causal relations. Future research could investigate how different social influence mechanisms that act on different types of opinions affect opinion patterns. Meanwhile, when interpreting the results of opinion dynamic models, one should be aware of the abstract nature of opinions.

Opinion formation models typically produce dynamics in which the opinions of individual agents are characterised by frequent and quite significant fluctuations. Some agents change their opinions from one extreme to another and back while their network position, social identity, and degree of bias are all assumed to remain fixed (see Figure 3.2 or the dynamics present in the classic bounded confidence model by Deffuant et al., 2000). Such strong fluctuations are rarely observed in real public debates. Many models, however, focus only on the macro-level distribution of opinions (see the models in Chapters 2 and 3 or, for example, by Axelrod et al., 2021; Schweighofer et al., 2020). The issue with this approach is that many different trajectories can lead to outcomes like polarisation or consensus—this issue is known as equifinality (Elsenbroich & Badham, 2023)—but not all of these trajectories represent realistic dynamics. One route to address the discrepancy between individual-level opinion fluctuations and the simultaneous stability of other features is to follow an adaptive system approach (Schill et al., 2019), for example, by allowing the interaction network to co-evolve with opinions (Edmonds, 2020; Will et al., 2020). This approach, however, tends to generate very complex dynamics and is sensitive to small variations in the assumptions. In Chapter 4, I take a different route to address the issue of exaggerated opinion fluctuations: I fix the opinions at different snapshots by extracting them from multi-wave empirical data and analyse how individuals perceive the change in opinions. In general, however, we need further efforts to formalise how individuals actually form opinions in a way such that both macro- and micro-level trajectories plausibly reflect realistic patterns.

The ability to generalise the conclusions drawn from opinion formation models is limited by the fact that opinion formation models are designed nearly exclusively by WEIRD (western, educated, industrialised, rich, and democratic) modellers—a characteristic that modelling shares with most other fields in the social and psychological sciences (Henrich et al., 2010)[1]. In support of this point, I collected the abstracts of all published articles in the *Journal of Artificial Societies and Social Simulation* (JASSS) since 2010 and plotted the geographic distribution of the first affiliation of each (co-)author (Figure 5.1). JASSS is one of the key journals for opinion dynamics modellers, yet only a fraction of the studies have contributions from authors affiliated at institutes or universities in non-WEIRD countries (with China being a notable exception). This poses an important limitation, because modellers from the same cultural background will likely have similar culturally-driven experiences about social phenomena and the underlying processes. But social phenomena are not universal across societies. For example, the effect of cognitive biases varies across cultures (Moser et al., 2022). Given the over-representation of researchers from Europe and the US in JASSS, I suspect that social processes, phenomena, or influence relations characteristic of societies in Africa or South East Asia are, in

[1]Here, different to Henrich et al., the term 'WEIRD' does not relate to the study subjects, but to the researchers themselves.

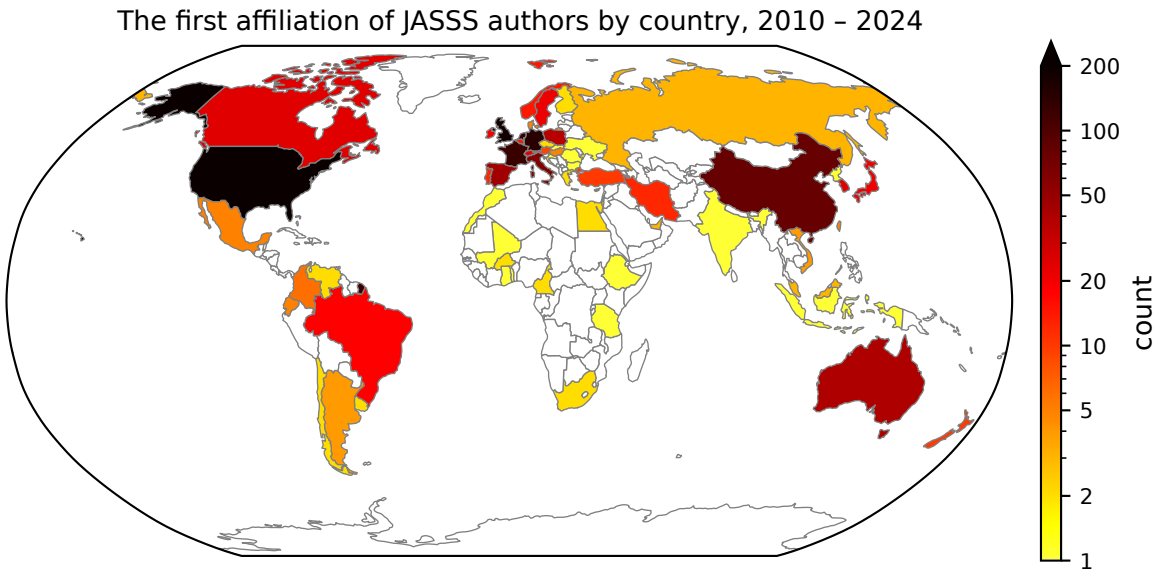


Figure 5.1: Countries associated with the authors of JASSS (*Journal of Artificial Societies and Social Simulation*) publications from 2010 to March 2024. The association is the country of the first affiliation stated by each author. A vast majority of authors in JASSS are located in the US (280) or Central/Western Europe (182 in UK, 146 in Germany, 126 in the Netherlands, and 124 in France). There are barely any authors in Africa or South East Asia. I suspect that most models (unconsciously) adopt a US or eurocentric perspective. This systematic underrepresentation of certain communities is an issue because many social processes are not universal but manifest in different ways depending on the culture of the people.

general, understudied in current models of opinion dynamics[2]. After all, modellers are biased and culturally embedded humans and so are their models (Hämäläinen, 2015). The models presented in this thesis are no exception and suffer from the limitations of being designed by WEIRD authors. This should encourage us to incentivise or strengthen international research collaborations with underrepresented scientific communities, especially for studies about social influence and opinion formation.

Another important consideration concerns the purpose of modelling. There is a trend in the field trying to make models increasingly complex and comprehensive in order to improve their match with reality (see the recent special issue on prediction with agent-based models of social phenomena in Elsenbroich & Polhill, 2023). Models with this purpose often aim to forecast social phenomena, or at least, to provide coarse predictions of future trends. This is a daunting challenge limited not only by the factors described above (the abstract nature of opinions in current models, the mismatch of individual-level trajectories with empirical evidence, and the issue of underrepresented communities) but also by ontological properties of social systems that prevent prediction (Elsenbroich & Polhill, 2023). After all, the world is a radically uncertain place and peoples' behaviours are heterogeneous, stochastic, and path-dependent. This world is also characterised by a variety of intertwined emergent phenomena arising from coupled, non-linear feedback processes (Bookstaber, 2017). There is also a different trend that focuses on different modelling purposes, such as theory improvement, visualisation of processes, suggestion of new empirical experiments, or even teaching about complexity of social systems (Edmonds, 2023; Epstein, 2008). For a model to be useful in this sense, it is essential to narrow its scope, that is to describe only those aspects relevant to the phenomenon of interest (Smaldino, 2023).

[2] I assume that 'affiliation' indicates something about the representation of cultural backgrounds of the authors, although the authors at US and European institutes may of course have very diverse cultural backgrounds.

But this comes at the cost of oversimplifying the complexity of real-world opinion dynamics. For example, in the present thesis, I did not consider the influence of public media, political ideology, personal experience, or actors with vested interests in the climate debate, although these aspects have crucial impacts on peoples' opinions. This poses a challenge because model outcomes deviate, by design, from empirical observations. It is difficult to validate a model with a relatively narrow scope, which aims to provide mechanistic understanding of particular processes, against empirical data, which incorporates the variety and complexity of real-world processes. Opinion dynamics modelling can nevertheless be a useful addition to the portfolio of tools available to the social sciences, for example, by helping to enforce transparency about theories of social processes, illustrating complex feedbacks, illuminating gaps in our understanding of empirical data, or suggesting new aspects to be investigated empirically (as discussed before in Section 1.3). Each model is ultimately a trade-off between the complexities of the real world and the necessity to abstract from such complexities.

5.3 Implications

In this final paragraph, I discuss the lessons learned from this thesis for opinion formation, in general, and polarisation, in particular. These lessons concern the issue of public disagreement on questions related to climate change. Referring to the impact of cognitive biases on climate change attitudes, Daniel Kahneman—whose research on biases was one of the key inspirations for this thesis—said: "I really see no path to success on climate change" (Marshall, 2015). There is certainly a lot of truth in this statement. However, people *are* increasingly aware and concerned about the climate crisis. Moreover, I presented how the very cognitive biases that may cause pessimism among social scientists can, under certain conditions, counter-intuitively help to foster agreement about the fundamental questions on climate change and, thereby, increase social cohesion in non-obvious ways. Chapter 2 suggests that a combination of moderate confirmation bias and moderate ambiguity in communication can foster pro-climate consensus. For example, if the design of a social media platform would encourage its users to write more generic (and thus ambiguous) messages (by removing clearly identifiable hashtags, for instance), this might increase the chances of reaching a pro-climate consensus. Chapter 3 suggests that in-group bias related to generational identity may help to form consensus under certain conditions. In the recent past, young activists, such as Greta Thunberg, have put strong emphasis on generational identities in the climate debate. Counterintuitively, in-group bias related to such stronger identification among youngsters and elders does not need to drive polarisation. In dispersed societies, a moderate level of in-group bias may foster agreement about climate change among one group—the youngsters, for example—which helps to spread pro-climate attitudes more widely and, eventually, also across identity groups. Finally, Chapter 4 suggests that people in Germany may perceive more polarisation about climate-related issues than what actually exists, assuming that their perceptions are influenced predominantly by their partisan in-groups. However, such a false impression of high polarisation is especially pronounced among those groups with strong agreement about climate change. Correcting for such perceptions by making these individuals aware of the fact that the attitudes towards climate change have shifted also in other groups, for example, may reduce overall perceived polarisation and, thereby, relieve policy-makers from concerns about deteriorating social cohesion or creating public resistance. To be clear, I do not argue that the mechanisms that increase agreement or cohesion in the specific models describe realistic opinion dynamics or that climate advocates should aim to exploit these biases in real politics. For example, a combination of social identity bias with the vested interests of the public media might be quite detrimental for reducing disagreement. More research is needed to investigate such feedback mechanisms. However, the results show that there are paths

through which even something seemingly negative like socio-cognitive biases can actually have a positive effect on reaching pro-climate agreement. As polarisation creates an environment of mistrust and misinformation and poses a barrier for effective policy making in the face of a looming climate crisis, reducing polarisation may be critical and the present thesis has mainly focused on this aspect. However, greater agreement about climate-related issues by itself is only useful if we do not compromise the strong, science-based concern about the consequences of climate change and the call for rapid and thorough action to mitigate these impacts. To end with Greta Thunberg, we should '[...] put [our] differences aside and start acting as [one] would in a crisis' (Thunberg, 2019).

Bibliography

Abelson, R. P. (1967, January). Mathematical Models in Social Psychology. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (pp. 1–54, Vol. 3). Academic Press. https://doi.org/10.1016/S0065-2601(08)60341-X.

Abramowitz, A. I., & Saunders, K. L. (2008). Is Polarization a Myth? *The Journal of Politics*, *70*(2), 542–555. https://doi.org/10.1017/s0022381608080493.

Abrams, D., Wetherell, M., Cochrane, S., Hogg, M. A., & Turner, J. C. (1990). Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization*. *British Journal of Social Psychology*, *29*(2), 97–119. https://doi.org/10.1111/j.2044-8309.1990.tb00892.x.

Acemoglu, D., & Ozdaglar, A. (2011). Opinion Dynamics and Learning in Social Networks. *Dynamic Games and Applications*, *1*(1), 3–49. https://doi.org/10.1007/s13235-010-0004-1.

Alizadeh, M., Cioffi-Revilla, C., & Crooks, A. (2015). The effect of in-group favoritism on the collective behavior of individuals' opinions. *Advances in Complex Systems*, *18*(01n02), 1550002. https://doi.org/10.1142/S0219525915500022.

Andre, P., Boneva, T., Chopra, F., & Falk, A. (2024). Globally representative evidence on the actual and perceived support for climate action. *Nature Climate Change*, 1–7. https://doi.org/10.1038/s41558-024-01925-3.

Armaly, M. T., & Enders, A. M. (2021). The role of affective orientations in promoting perceived polarization. *Political Science Research and Methods*, *9*(3), 615–626. https://doi.org/10.1017/psrm.2020.24.

Axelrod, R. (1997). The Dissemination of Culture: A Model with Local Convergence and Global Polarization. *Journal of Conflict Resolution*, *41*(2), 203–226. https://doi.org/10.1177/0022002797041002001.

Axelrod, R., Daymude, J. J., & Forrest, S. (2021). Preventing extreme polarization of political attitudes. *Proceedings of the National Academy of Sciences*, *118*(50), e2102139118. https://doi.org/10.1073/pnas.2102139118.

Baccelli, F., Chatterjee, A., & Vishwanath, S. (2017). Pairwise Stochastic Bounded Confidence Opinion Dynamics: Heavy Tails and Stability. *IEEE Transactions on Automatic Control*, *62*(11), 5678–5693. https://doi.org/10.1109/TAC.2017.2691312.

Bak-Coleman, J. B., Alfano, M., Barfuss, W., Bergstrom, C. T., Centeno, M. A., Couzin, I. D., Donges, J. F., Galesic, M., Gersick, A. S., Jacquet, J., Kao, A. B., Moran, R. E., Romanczuk, P., Rubenstein, D. I., Tombak, K. J., Van Bavel, J. J., & Weber, E. U. (2021). Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, *118*(27), e2025764118. https://doi.org/10.1073/pnas.2025764118.

Baldassarri, D., & Goldberg, A. (2014). Neither Ideologues nor Agnostics: Alternative Voters' Belief System in an Age of Partisan Politics. *American Journal of Sociology*, *120*(1), 45–95. https://doi.org/10.1086/676042.

Baldassarri, D., & Page, S. E. (2021). The emergence and perils of polarization. *Proceedings of the National Academy of Sciences*, *118*(50), e2116863118. https://doi.org/10.1073/pnas.2116863118.

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.

Banisch, S., & Shamon, H. (2024). Validating Argument-Based Opinion Dynamics with Survey Experiments. *Journal of Artificial Societies and Social Simulation*, *27*(1), 17.

Barkoczi, D., & Galesic, M. (2016). Social learning strategies modify the effect of network structure on group performance. *Nature Communications*, *7*(1), 13109. https://doi.org/10.1038/ncomms13109.

Barnes, M. L., Wang, P., Cinner, J. E., Graham, N. A. J., Guerrero, A. M., Jasny, L., Lau, J., Sutcliffe, S. R., & Zamborain-Mason, J. (2020). Social determinants of adaptive and transformative responses to climate change. *Nature Climate Change*, *10*(9), 823–828. https://doi.org/10.1038/s41558-020-0871-4.

Bartels, L. M. (2002). Beyond the Running Tally: Partisan Bias in Political Perceptions. *Political Behavior*, *24*(2), 117–150. https://doi.org/10.1023/A:1021226224601.

Bavel, J. J. V., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J. H., Gelfand, M., Han, S., Haslam, S. A., Jetten, J., ... Willer, R. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, *4*(5), 460–471. https://doi.org/10.1038/s41562-020-0884-z.

Bayes, R., & Druckman, J. N. (2021). Motivated reasoning and climate change. *Current Opinion in Behavioral Sciences*, *42*, 27–35. https://doi.org/10.1016/j.cobeha.2021.02.009.

Beattie, G., & McGuire, L. (2018, October). *The Psychology of Climate Change* (1st ed.). Routledge.

Bliuc, A.-M., McGarty, C., Thomas, E. F., Lala, G., Berndsen, M., & Misajon, R. (2015). Public division about climate change rooted in conflicting socio-political identities. *Nature Climate Change*, *5*(3), 226–229. https://doi.org/10.1038/nclimate2507.

Bodrunova, S. S., Blekanov, I., Smoliarova, A., & Litvinenko, A. (2019). Beyond Left and Right: Real-World Political Polarization in Twitter Discussions on Inter-Ethnic Conflicts. *Media and Communication*, *7*, 119–132. https://doi.org/10.17645/mac.v7i3.1934.

Bolsen, T., & Shapiro, M. A. (2018). The US News Media, Polarization on Climate Change, and Pathways to Effective Communication. *Environmental Communication*, *12*(2), 149–163. https://doi.org/10.1080/17524032.2017.1397039.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(suppl 3), 7280–7287.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421.

Bookstaber, R. (2017, May). *The End of Theory: Financial Crises, the Failure of Economics, and the Sweep of Human Interaction*. Princeton University Press.

Borick, C. P., & Rabe, B. G. (2014). Weather or Not? Examining the Impact of Meteorological Conditions on Public Opinion regarding Global Warming. *Weather, Climate, and Society*, *6*(3), 413–424. https://doi.org/10.1175/WCAS-D-13-00042.1.

Boroomand, A., & Smaldino, P. E. (2023). Superiority Bias and Communication Noise Can Enhance Collective Problem Solving. *Journal of Artificial Societies and Social Simulation*, *26*(3). https://doi.org/10.18564/jasss.5154.

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., van Borkulo, C. D., van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*(1), 1–18. https://doi.org/10.1038/s43586-021-00055-w.

Boutyline, A., & Vaisey, S. (2017). Belief Network Analysis: A Relational Approach to Understanding the Structure of Attitudes. *American Journal of Sociology*, *122*(5), 1371–1447. https://doi.org/10.1086/691274.

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. https://doi.org/10.1111/pops.12337.

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., & Holman, B. (2017). Understanding Polarization: Meanings, Measures, and Model Evaluation. *Philosophy of Science*, *84*(1), 115–159. https://doi.org/10.1086/688938.

Bramson, A., Grim, P., Singer, D. J., Fisher, S., Berger, W., Sack, G., & Flocken, C. (2016). Disambiguation of social polarization concepts and measures. *The Journal of Mathematical Sociology*, *40*(2), 80–111. https://doi.org/10.1080/0022250X.2016.1147443.

Brandt, M. J., & Sleegers, W. W. A. (2021). Evaluating belief system networks as a theory of political belief system dynamics. *Personality and Social Psychology Review*, *25*, 159–185. https://doi.org/10.1177/1088868321993751.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, *86*(2), 307–324. https://doi.org/10.1037/0033-2909.86.2.307.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love or outgroup hate? *Journal of Social Issues*, *55*(3), 429–444. https://doi.org/10.1111/0022-4537.00126.

Brown, R., Vivian, J., & Hewstone, M. (1999). Changing attitudes through intergroup contact: The effects of group membership salience. *European Journal of Social Psychology*, *29*(5-6), 741–764. https://doi.org/10.1002/(SICI)1099-0992(199908/09)29:5/6<741::AID-EJSP972>3.0.CO;2-8.

Carpentras, D., Lueders, A., & Quayle, M. (2021, September). A method for exploring attitude systems by combining Belief Network Analysis and Item Response Theory (ResIN). https://doi.org/10.31234/osf.io/uzdcg.

Carpentras, D., Maher, P. J., O'Reilly, C., & Quayle, M. (2022). Deriving an Opinion Dynamics Model from Experimental Data. *Journal of Artificial Societies and Social Simulation*, *25*(4), 4. https://doi.org/10.18564/jasss.4947.

Carro, A., Toral, R., & San Miguel, M. (2013). The Role of Noise and Initial Conditions in the Asymptotic Solution of a Bounded Confidence, Continuous-Opinion Model. *Journal of Statistical Physics*, *151*(1), 131–149. https://doi.org/10.1007/s10955-012-0635-2.

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics*, *81*(2), 591–646. https://doi.org/10.1103/RevModPhys.81.591.

Centola, D. (2022). The network science of collective intelligence. *Trends in Cognitive Sciences*, *26*(11), 923–941. https://doi.org/10.1016/j.tics.2022.08.009.

Centola, D., & Macy, M. (2007). Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*, *113*(3), 702–734. https://doi.org/10.1086/521848.

Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., Leon-Villagra, P., & Sanborn, A. (2020). Probabilistic Biases Meet the Bayesian Brain. *Current Directions in Psychological Science*, *29*(5), 506–512. https://doi.org/10.1177/0963721420954801.

Chen, X., Zhang, X., Xie, Y., & Li, W. (2017). Opinion Dynamics of Social-Similarity-Based Hegselmann–Krause Model. *Complexity*, *2017*, 1–12. https://doi.org/10.1155/2017/1820257.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 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.

Clayton, S. D., & Opotow, S. (Eds.). (2003). *Identity and the natural environment: The psychological significance of nature*. MIT Press.

Coglianese, C. (2001). Is Consensus an Appropriate Basis for Regulatory Policy? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.270488.

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.

Cook, B. R., & Overpeck, J. T. (2019). Relationship-building between climate scientists and publics as an alternative to information transfer. *WIREs Climate Change*, *10*(2), e570. https://doi.org/10.1002/wcc.570.

Cook, J., & Lewandowsky, S. (2016). Rational Irrationality: Modeling Climate Change Belief Polarization Using Bayesian Networks. *Topics in Cognitive Science*, *8*(1), 160–179. https://doi.org/10.1111/tops.12186.

Couzin, I. D., Ioannou, C. C., Demirel, G., Gross, T., Torney, C. J., Hartnett, A., Conradt, L., Levin, S. A., & Leonard, N. E. (2011). Uninformed Individuals Promote Democratic Consensus in Animal Groups. *Science*. https://doi.org/10.1126/science.1210280.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376. https://doi.org/10.1007/s11229-006-9097-x.

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: The Causal Attitude Network (CAN) model. *Psychological Review*, *123*, 2–22. https://doi.org/10.1037/a0039802.

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. J. (2018). The Attitudinal Entropy (AE) Framework as a General Theory of Individual Attitudes. *Psychological Inquiry*, *29*(4), 175–193. https://doi.org/10.1080/1047840X.2018.1537246.

Dalege, J., Galesic, M., & Olsson, H. (2023, April). *Networks of Beliefs: An Integrative Theory of Individual- and Social-Level Belief Dynamics* (Preprint). Open Science Framework. https://doi.org/10.31219/osf.io/368jz.

Dandekar, P., Goel, A., & Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, *110*(15), 5791–5796. https://doi.org/10.1073/pnas.1217220110.

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

De Sanctis, L., & Galla, T. (2009). Effects of noise and confidence thresholds in nominal and metric Axelrod dynamics of social influence. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, *79*(4 Pt 2), 046108. https://doi.org/10.1103/PhysRevE.79.046108.

Deffuant, G. (2006). Comparing extremism propagation patterns in continuous opinion models. *Journal of Artificial Societies and Social Simulation*, *9*(3), 8.

Deffuant, G., Carletti, T., & Huet, S. (2013). The leviathan model: Absolute dominance, generalised distrust, small worlds and other patterns emerging from combining vanity with opinion propagation. *Journal of Artificial Societies and Social Simulation*, *16*(1), 5. https://doi.org/10.18564/jasss.2070.

Deffuant, G., Keijzer, M. A., & Banisch, S. (2023, May). Regular access to constantly renewed online content favors radicalization of opinions. https://doi.org/10.48550/arXiv.2305.16855.

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.

DellaPosta, D., Shi, Y., & Macy, M. (2015). Why Do Liberals Drink Lattes? *American Journal of Sociology*, *120*(5), 1473–1511. https://doi.org/10.1086/681254.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*(3), 629–636. https://doi.org/10.1037/h0046408.

DiMaggio, P., Evans, J., & Bryson, B. (1996). Have American's Social Attitudes Become More Polarized? *American Journal of Sociology*, *102*(3), 690–755. https://doi.org/10.1086/230995.

Dixit, A. K., & Weibull, J. W. (2007). Political polarization. *Proceedings of the National Academy of Sciences*, *104*(18), 7351–7356. https://doi.org/10.1073/pnas.0702071104.

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021). Affective polarization, local contexts and public opinion in America. *Nature Human Behaviour*, *5*(1), 28–38. https://doi.org/10.1038/s41562-020-01012-5.

Druckman, J. N., & Leeper, T. J. (2012). Is Public Opinion Stable? Resolving the Micro/Macro Disconnect in Studies of Public Opinion. *Daedalus*, *141*(4), 50–68. https://doi.org/10.1162/DAED_a_00173.

Dunlap, R. E., McCright, A. M., & Yarosh, J. H. (2016). The Political Divide on Climate Change: Partisan Polarization Widens in the U.S. *Environment: Science and Policy for Sustainable Development*, *58*(5), 4–23. https://doi.org/10.1080/00139157.2016.1208995.

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y.

Edmonds, B. (2020). Co-developing beliefs and social influence networks—towards understanding socio-cognitive processes like Brexit. *Quality & Quantity*, *54*(2), 491–515. https://doi.org/10.1007/s11135-019-00891-9.

Edmonds, B. (2023). The practice and rhetoric of prediction – the case in agent-based modelling. *International Journal of Social Research Methodology*, *26*(2), 157–170. https://doi.org/10.1080/13645579.2022.2137921.

Edmonds, B., Le Page, C., Bithell, M., Chattoe-Brown, E., Grimm, V., Meyer, R., Montañola-Sales, C., Ormerod, P., Root, H., & Squazzoni, F. (2019). Different Modelling Purposes. *Journal of Artificial Societies and Social Simulation*, *22*(3), 6.

Edmonds, B., & ní Aodha, L. (2019). Using Agent-Based Modelling to Inform Policy – What Could Possibly Go Wrong? In P. Davidsson & H. Verhagen (Eds.), *Multi-Agent-Based Simulation XIX* (pp. 1–16, Vol. 11463). Springer International Publishing. https://doi.org/10.1007/978-3-030-22270-3_1.

Eisenberg, E. M. (1984). Ambiguity as strategy in organizational communication. *Communication Monographs*, *51*(3), 227–242. https://doi.org/10.1080/03637758409390197.

Elsenbroich, C., & Badham, J. (2023). Negotiating a Future that is not like the Past. *International Journal of Social Research Methodology*, *26*(2), 207–213. https://doi.org/10.1080/13645579.2022.2137935.

Elsenbroich, C., & Polhill, J. G. (2023). Agent-based modelling as a method for prediction in complex social systems. *International Journal of Social Research Methodology*, *26*(2), 133–142. https://doi.org/10.1080/13645579.2023.2152007.

Enders, A. M., & Armaly, M. T. (2019). The Differential Effects of Actual and Perceived Polarization. *Political Behavior*, *41*(3), 815–839. https://doi.org/10.1007/s11109-018-9476-2.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, *4*(5), 41–60. https://doi.org/10.1002/(SICI)1099-0526(199905/06)4:5<41::AID-CPLX9>3.0.CO;2-F.

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), 12.

Erdos, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, *5*(1), 17–60.

Esposo, S. R., Hornsey, M. J., & Spoor, J. R. (2013). Shooting the messenger: Outsiders critical of your group are rejected regardless of argument quality. *British Journal of Social Psychology*, *52*(2), 386–395. https://doi.org/10.1111/bjso.12024.

Estrada, M., Schultz, P. W., Silva-Send, N., & Boudrias, M. A. (2017). The Role of Social Influences on Pro-Environment Behaviors in the San Diego Region. *Journal of Urban Health : Bulletin of the New York Academy of Medicine*, *94*(2), 170–179. https://doi.org/10.1007/s11524-017-0139-0.

European Social Survey European Research Infrastructure (ESS ERIC). (2017). ESS8 - integrated file, edition 2.3. https://doi.org/10.21338/ess8e02_3.

European Social Survey European Research Infrastructure (ESS ERIC). (2020). ESS8 - integrated file, edition 2.2.

European Social Survey European Research Infrastructure (ESS ERIC). (2022). ESS10 Self-completion - integrated file. edition 3.1. https://doi.org/10.21338/ess10sce03_1.

Falk, A., & Heckman, J. J. (2009). Lab Experiments Are a Major Source of Knowledge in the Social Sciences. *Science*, *326*(5952), 535–538. https://doi.org/10.1126/science.1168244.

Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., Mekacher, A., Pearce, W., Zollo, F., Quattrociocchi, W., & Baronchelli, A. (2022). Growing polarization around climate change on social media. *Nature Climate Change*, *12*(12), 1114–1121. https://doi.org/10.1038/s41558-022-01527-x.

Fazelpour, S., & Steel, D. (2022). Diversity, Trust, and Conformity: A Simulation Study. *Philosophy of Science*, *89*(2), 209–231. https://doi.org/10.1017/psa.2021.25.

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.

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202.

Fielding, K. S., & Hornsey, M. J. (2016). A Social Identity Analysis of Climate Change and Environmental Attitudes and Behaviors: Insights and Opportunities. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00121.

Fielding, K. S., Hornsey, M. J., Thai, H. A., & Toh, L. L. (2020). Using ingroup messengers and ingroup values to promote climate change policy. *Climatic Change*, *158*(2), 181–199. https://doi.org/10.1007/s10584-019-02561-z.

Fiorina, M. P., & Abrams, S. J. (2008). Political Polarization in the American Public. *Annual Review of Political Science*, *11*(1), 563–588. https://doi.org/10.1146/annurev.polisci.11.053106.153836.

Flache, A., & Macy, M. W. (2011). Local Convergence and Global Diversity: From Interpersonal to Social Influence. *Journal of Conflict Resolution*, *55*(6), 970–995. https://doi.org/10.1177/0022002711414371.

Flache, A., & Mäs, M. (2008). Why do faultlines matter? A computational model of how strong demographic faultlines undermine team cohesion. *Simulation Modelling Practice and Theory*, *16*(2), 175–191. https://doi.org/10.1016/j.simpat.2007.11.020.

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). https://doi.org/10.18564/jasss.352.

Flores, A., Cole, J. C., Dickert, S., Eom, K., Jiga-Boy, G. M., Kogut, T., Loria, R., Mayorga, M., Pedersen, E. J., Pereira, B., Rubaltelli, E., Sherman, D. K., Slovic, P., Västfjäll, D., & Van Boven, L. (2022). Politicians polarize and experts depolarize public support for COVID-19 management policies across countries. *Proceedings of the National Academy of Sciences*, *119*(3), e2117543119. https://doi.org/10.1073/pnas.2117543119.

Frankenhuis, W. E., Panchanathan, K., & Smaldino, P. E. (2023). Strategic ambiguity in the social sciences. *Social Psychological Bulletin*, *18*, e9923. https://doi.org/10.32872/spb.9923.

French, J. (1956). A formal theory of social power. *Psychological Review*, *63*(3), 181–194.

Friedkin, N. E., & Johnsen, E. C. (1990). Social influence and opinions. *The Journal of Mathematical Sociology*, *15*(3-4), 193–206. https://doi.org/10.1080/0022250X.1990.9990069.

Friedkin, N. E., Proskurnikov, A. V., Tempo, R., & Parsegov, S. E. (2016). Network science on belief system dynamics under logic constraints. *Science*, *354*(6310), 321–326. https://doi.org/10.1126/science.aag2624.

Gabriel, N., & O'Connor, C. (2024). Can Confirmation Bias Improve Group Learning? *Philosophy of Science*, *91*(2), 329–350. https://doi.org/10.1017/psa.2023.176.

Gächter, S. (2007). Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications. In *Economics and psychology: A promising new cross-disciplinary field*. (pp. 19–50). MIT Press. https://doi.org/10.7551/mitpress/2604.003.0006.

Galesic, M., Barkoczi, D., Berdahl, A. M., Biro, D., Carbone, G., Giannoccaro, I., Goldstone, R. L., Gonzalez, C., Kandler, A., Kao, A. B., Kendal, R., Kline, M., Lee, E., Massari, G. F., Mesoudi, A., Olsson, H., Pescetelli, N., Sloman, S. J., Smaldino, P. E., & Stein, D. L. (2023). Beyond collective intelligence: Collective adaptation. *Journal of The Royal Society Interface*, *20*(200), 20220736. https://doi.org/10.1098/rsif.2022.0736.

Galesic, M., Olsson, H., Dalege, J., van der Does, T., & Stein, D. L. (2021). Integrating social and cognitive aspects of belief dynamics: Towards a unifying framework. *Journal of The Royal Society Interface*, *18*(176), 20200857. https://doi.org/10.1098/rsif.2020.0857.

Geschke, D., Lorenz, J., & Holtz, P. (2019). The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, *58*(1), 129–149. https://doi.org/10.1111/bjso.12286.

Gestefeld, M., Lorenz, J., Henschel, N. T., & Boehnke, K. (2022). Decomposing attitude distributions to characterize attitude polarization in Europe. *SN Social Sciences*, *2*(7), 110. https://doi.org/10.1007/s43545-022-00342-7.

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, *1*(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x.

Gigerenzer, G., Fiedler, K., & Olsson, H. (2012). Rethinking cognitive biases as environmental consequences. In *Ecological rationality: Intelligence in the world* (pp. 80–110). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195315448.003.0025.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological review*, *103*(4), 650. https://doi.org/10.1037/0033-295X.103.4.650.

Gonyea, J. G., & Hudson, R. B. (2020). In an Era of Deepening Partisan Divide, What is the Meaning of Age or Generational Differences in Political Values? *Public Policy & Aging Report*, *30*(2), 52–55. https://doi.org/10.1093/ppar/praa003.

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, *83*(6), 1420–1443.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, *78*(6), 1360–1380.

Grauwin, S., & Jensen, P. (2012). Opinion group formation and dynamics: Structures that last from nonlasting entities. *Physical Review E*, *85*(6), 066113. https://doi.org/10.1103/PhysRevE.85.066113.

Gutmann, A., & Thompson, D. F. (2009, January). *Why Deliberative Democracy?* Princeton University Press. https://doi.org/10.1515/9781400826339.

Hall, M. P., Lewis, N. A., & Ellsworth, P. C. (2018). Believing in climate change, but not behaving sustainably: Evidence from a one-year longitudinal study. *Journal of Environmental Psychology*, *56*, 55–62. https://doi.org/10.1016/j.jenvp.2018.03.001.

Hämäläinen, R. P. (2015). Behavioural issues in environmental modelling – The missing perspective. *Environmental Modelling & Software*, *73*, 244–253. https://doi.org/10.1016/j.envsoft.2015.08.019.

Hankins, K., Muldoon, R., & Schaefer, A. (2023). Does (mis)communication mitigate the upshot of diversity? *PLOS ONE*, *18*(3), e0283248. https://doi.org/10.1371/journal.pone.0283248.

Hansen, J., Sato, M., & Ruedy, R. (2012). Perception of climate change. *Proceedings of the National Academy of Sciences*, *109*(37), E2415–E2423. https://doi.org/10.1073/pnas.1205276109.

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

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29. https://doi.org/10.1038/466029a.

Herdağdelen, A., Zuo, W., Gard-Murray, A., & Bar-Yam, Y. (2013). An exploration of social identity: The geography and politics of news-sharing communities in twitter. *Complexity*, *19*(2), 10–20. https://doi.org/10.1002/cplx.21457.

Herold, M., Joachim, J., Otteni, C., & Vorländer, H. (2023). *Polarization in Europe. An Analysis of ten European Countries* (tech. rep.). Mercator Forum Migration and Democracy (MIDEM). Dresden.

Herzog, S. M., & Hertwig, R. (2009). The Wisdom of Many in One Mind: Improving Individual Judgments With Dialectical Bootstrapping. *Psychological Science*, *20*(2), 231–237. https://doi.org/10.1111/j.1467-9280.2009.02271.x.

Hewstone, M., Lolliot, S., Swart, H., Myers, E., Voci, A., Al Ramiah, A., & Cairns, E. (2014). Intergroup contact and intergroup conflict. *Peace and Conflict: Journal of Peace Psychology*, *20*(1), 39–53. https://doi.org/10.1037/a0035582.

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup Bias. *Annual Review of Psychology*, *53*(1), 575–604. https://doi.org/10.1146/annurev.psych.53.100901.135109.

Hohmann, M., Devriendt, K., & Coscia, M. (2023). Quantifying ideological polarization on a network using generalized Euclidean distance. *Science Advances*, *9*(9), eabq2044. https://doi.org/10.1126/sciadv.abq2044.

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, *5*(2), 109–137. https://doi.org/10.1016/0378-8733(83)90021-7.

Holley, R. A., & Liggett, T. M. (1975). Ergodic Theorems for Weakly Interacting Infinite Systems and the Voter Model. *The Annals of Probability*, *3*(4), 643–663. https://doi.org/10.1214/aop/1176996306.

Holyst, J. A., Kacperski, K., & Schweitzer, F. (2001, April). Social impact models of opinion dynamics. In *Annual Reviews of Computational Physics IX* (pp. 253–273, Vol. Volume 9). WORLD SCIENTIFIC. https://doi.org/10.1142/9789812811578_0005.

Homer-Dixon, T., Maynard, J. L., Mildenberger, M., Milkoreit, M., Mock, S. J., Quilley, S., Schröder, T., & Thagard, P. (2013). A Complex Systems Approach to the Study of Ideology: Cognitive-Affective Structures and the Dynamics of Belief Systems. *Journal of Social and Political Psychology*, *1*(1), 337–363. https://doi.org/10.5964/jspp.v1i1.36.

Hornsey, M. J. (2008). Social Identity Theory and Self-categorization Theory: A Historical Review. *Social and Personality Psychology Compass*, *2*(1), 204–222. https://doi.org/10.1111/j.1751-9004.2007.00066.x.

Hornsey, M. J., Harris, E. A., Bain, P. G., & Fielding, K. S. (2016). Meta-analyses of the determinants and outcomes of belief in climate change. *Nature Climate Change*, *6*(6), 622–626. https://doi.org/10.1038/nclimate2943.

Hornsey, M. J., & Lewandowsky, S. (2022). A toolkit for understanding and addressing climate scepticism. *Nature Human Behaviour*, *6*(11), 1454–1464. https://doi.org/10.1038/s41562-022-01463-y.

Howe, P. D., Marlon, J. R., Mildenberger, M., & Shield, B. S. (2019). How will climate change shape climate opinion? *Environmental Research Letters*, *14*(11), 113001. https://doi.org/10.1088/1748-9326/ab466a.

Huet, S., & Deffuant, G. (2010). Openness leads to opinion stability and narrowness to volatility. *Advances in Complex Systems*, *13*(03), 405–423. https://doi.org/10.1142/S0219525910002633.

infratest dimap. (2024, February). ARD-DeutschlandTREND Februar 2024.

IPCC. (2023). *Summary for Policymakers. Climate Change 2023: Synthesis Report. A Report of the Intergovernmental Panel on Climate Change. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]* (tech. rep.). IPCC. Geneva, Switzerland.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, Not Ideology: A Social Identity Perspective on Polarization. *The Public Opinion Quarterly*, *76*(3), 405–431.

Jasny, L., Waggle, J., & Fisher, D. R. (2015). An empirical examination of echo chambers in US climate policy networks. *Nature Climate Change*, *5*(8), 782–786. https://doi.org/10.1038/nclimate2666.

Jenkins-Smith, H. C., Ripberger, J. T., Silva, C. L., Carlson, D. E., Gupta, K., Carlson, N., Ter-Mkrtchyan, A., & Dunlap, R. E. (2020). Partisan asymmetry in temporal stability of climate change beliefs. *Nature Climate Change*, *10*(4), 322–328. https://doi.org/10.1038/s41558-020-0719-y.

Johnson, D., & Levin, S. (2009). The tragedy of cognition: Psychological biases and environmental inaction. *Current Science*, *97*(11), 1593–1603.
Judge, M., Kashima, Y., Steg, L., & Dietz, T. (2023). Environmental Decision-Making in Times of Polarization. *Annual Review of Environment and Resources*, *48*(Volume 48, 2023), 477–503. https://doi.org/10.1146/annurev-environ-112321-115339.
Kahan, D. M. (2016). The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It. In *Emerging Trends in the Social and Behavioral Sciences* (pp. 1–16). American Cancer Society. https://doi.org/10.1002/9781118900772.etrds0417.
Kahan, D. M., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research*, *14*(2), 147–174. https://doi.org/10.1080/13669877.2010.511246.
Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*, *2*(10), 732–735. https://doi.org/10.1038/nclimate1547.
Kahneman, D., Sibony, O., & Sunstein, C. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.
Kahneman, D. (2012). *Thinking, fast and slow*. Penguin Books.
Kahneman, D., Krakauer, D. C., Sibony, O., Sunstein, C., & Wolpert, D. (2022). An exchange of letters on the role of noise in collective intelligence. *Collective Intelligence*, *1*(1), 26339137221078593. https://doi.org/10.1177/26339137221078593.
Karimi, F., Génois, M., Wagner, C., Singer, P., & Strohmaier, M. (2018). Homophily influences ranking of minorities in social networks. *Scientific Reports*, *8*(1), 11077. https://doi.org/10.1038/s41598-018-29405-7.
Keijzer, M. A. (2022, January). *Opinion Dynamics in Online Social Media* [Doctoral dissertation, University of Groningen]. https://doi.org/10.33612/diss.196882523.
Keijzer, M. A., & Mäs, M. (2021). The strength of weak bots. *Online Social Networks and Media*, *21*, 100106. https://doi.org/10.1016/j.osnem.2020.100106.
Keijzer, M. A., Mäs, M., & Flache, A. (2018). Communication in Online Social Networks Fosters Cultural Isolation. *Complexity*, *2018*, 1–18. https://doi.org/10.1155/2018/9502872.
Keijzer, M. A., Mäs, M., & Flache, A. (2024). Polarization on Social Media: Micro-Level Evidence and Macro-Level Implications. *Journal of Artificial Societies and Social Simulation*, *27*(1), 7.
Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social Learning Strategies: Bridge-Building between Fields. *Trends in Cognitive Sciences*, *22*(7), 651–665. https://doi.org/10.1016/j.tics.2018.04.003.
Kjeldahl, E. M., & Hendricks, V. F. (2018). The sense of social influence: Pluralistic ignorance in climate change. *EMBO reports*, *19*(11), e47185. https://doi.org/10.15252/embr.201847185.
Klemm, K., Eguíluz, V., Toral, R., & Miguel, M. S. (2003a). Nonequilibrium transitions in complex networks: A model of social interaction. *Physical review. E, Statistical, nonlinear, and soft matter physics*. https://doi.org/10.1103/PhysRevE.67.026120.
Klemm, K., Eguíluz, V. M., Toral, R., & Miguel, M. S. (2003b). Global culture: A noise-induced transition in finite systems. *Physical Review E*, *67*(4), 045101. https://doi.org/10.1103/PhysRevE.67.045101.
Kowalska-Styczeń, A., & Malarz, K. (2020). Noise induced unanimity and disorder in opinion formation. *PLOS ONE*, *15*(7), e0235313. https://doi.org/10.1371/journal.pone.0235313.
Kozlowski, A. C., & Murphy, J. P. (2021). Issue alignment and partisanship in the American public: Revisiting the 'partisans without constraint' thesis. *Social Science Research*, *94*, 102498. https://doi.org/10.1016/j.ssresearch.2020.102498.
Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111.
Kurahashi-Nakamura, T., Mäs, M., & Lorenz, J. (2016). Robust clustering in generalized bounded confidence models. *Journal of Artificial Societies and Social Simulation*, *19*(4), 7. https://doi.org/10.18564/jasss.3220.
Landrum, A. R., Lull, R. B., Akin, H., Hasell, A., & Jamieson, K. H. (2017). Processing the papal encyclical through perceptual filters: Pope Francis, identity-protective cognition, and climate change concern. *Cognition*, *166*, 1–12. https://doi.org/10.1016/j.cognition.2017.05.015.
Latané, B. (1981). The psychology of social impact. *American Psychologist*, *36*(4), 343–356. https://doi.org/10.1037/0003-066X.36.4.343.
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. https://doi.org/10.5465/amr.1998.533229.
Laukemper, A., Keijzer, M. A., & Bakker, D. (2020). defSim.
Lazarsfeld, P. F., & Merton, R. K. (1954). Friendship as a social process: A substantive and methodological analysis. In M. Berger, T. Abel, & C. H. Page (Eds.), *Freedom and control in modern society* (pp. 18–66). D. Van Nostrand Company.
Lees, J., & Cikara, M. (2021). Understanding and combating misperceived polarization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1822), 20200143. https://doi.org/10.1098/rstb.2020.0143.
Leiserowitz, A., Roser-Renouf, C., Marlon, J., & Maibach, E. (2021). Global Warming's Six Americas: A review and recommendations for climate change communication. *Current Opinion in Behavioral Sciences*, *42*, 97–103. https://doi.org/10.1016/j.cobeha.2021.04.007.
Lenton, T. M., Rockström, J., Gaffney, O., Rahmstorf, S., Richardson, K., Steffen, W., & Schellnhuber, H. J. (2019). Climate tipping points — too risky to bet against. *Nature*, *575*(7784), 592–595. https://doi.org/10.1038/d41586-019-03595-0.
Levendusky, M. S., & Malhotra, N. (2016). (Mis)perceptions of Partisan Polarization in the American Public. *Public Opinion Quarterly*, *80*(S1), 378–391. https://doi.org/10.1093/poq/nfv045.
Levin, S. A., Milner, H. V., & Perrings, C. (2021). The dynamics of political polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(50), e2116950118. https://doi.org/10.1073/pnas.2116950118.
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.

Liu, S., Mäs, M., Xia, H., & Flache, A. (2023). Job Done? Future Modeling Challenges After 20 Years of Work on Bounded-Confidence Models. *Journal of Artificial Societies and Social Simulation*, *26*(4), 8.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, *37*(11), 2098.

Lucht, K., & Liebig, S. (2023). Sozial-ökologische Bündnisse als Antwort auf Transformationskonflikte? Die Kampagne von ver.di und Fridays for Future im ÖPNV. *PROKLA. Zeitschrift für kritische Sozialwissenschaft*, *53*(210), 15–33. https://doi.org/10.32387/prokla.v53i210.2037.

Lüders, A., Carpentras, D., & Quayle, M. (2024). Attitude networks as intergroup realities: Using network-modelling to research attitude-identity relationships in polarized political contexts. *British Journal of Social Psychology*, *63*(1), 37–51. https://doi.org/10.1111/bjso.12665.

Lüders, A., Reiss, S., MacCarron, P., Dinkelberg, A., & Quayle, M. (2023, July). Not Our Kind of Crowd! How Partisan Bias Distorts Perceptions of Political Twitter Bots. https://doi.org/10.31219/osf.io/swkvr.

Maciel, M. V., & Martins, A. C. R. (2020). Ideologically motivated biases in a multiple issues opinion model. *Physica A: Statistical Mechanics and its Applications*, *553*, 124293. https://doi.org/10.1016/j.physa.2020.124293.

Mackie, D. M., Gastardo-Conaco, M. C., & Skelly, J. J. (1992). Knowledge of the advocated position and the processing of in-group and out-group persuasive messages. *Personality and Social Psychology Bulletin*, *18*(2), 145–151. https://doi.org/10.1177/0146167292182005.

Macy, M., Deri, S., Ruch, A., & Tong, N. (2019). Opinion cascades and the unpredictability of partisan polarization. *Science Advances*, *5*(8), eaax0754. https://doi.org/10.1126/sciadv.aax0754.

Macy, M., & Tsvetkova, M. (2015). The Signal Importance of Noise. *Sociological Methods & Research*, *44*(2), 306–328. https://doi.org/10.1177/0049124113508093.

Maibach, E. W., Leiserowitz, A., Roser-Renouf, C., & Mertz, C. K. (2011). Identifying Like-Minded Audiences for Global Warming Public Engagement Campaigns: An Audience Segmentation Analysis and Tool Development. *PLOS ONE*, *6*(3), e17571. https://doi.org/10.1371/journal.pone.0017571.

Marris, E. (2019). Why young climate activists have captured the world's attention. *Nature*, *573*(7775), 471–472. https://doi.org/10.1038/d41586-019-02696-0.

Marshall, G. (2015). *Don't even think about it: Why our brains are wired to ignore climate change*. Bloomsbury.

Martins, A. C. R. (2009). Bayesian updating rules in continuous opinion dynamics models. *Journal of Statistical Mechanics: Theory and Experiment*, *2009*(02), P02017. https://doi.org/10.1088/1742-5468/2009/02/P02017.

Martins, A. C. R. (2008). Continuous Opinions and Discrete Actions in Opinion Dynamics Problems. *International Journal of Modern Physics C*, *19*(04), 617–624. https://doi.org/10.1142/S0129183108012339.

Mäs, M., Flache, A., & Helbing, D. (2010). Individualization as Driving Force of Clustering Phenomena in Humans. *PLOS Computational Biology*, *6*(10), e1000959. https://doi.org/10.1371/journal.pcbi.1000959.

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. https://doi.org/10.1287/orsc.1120.0767.

Mason, L. (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, *59*(1), 128–145. https://doi.org/10.1111/ajps.12089.

Masson, T., & Fritsche, I. (2021). We need climate change mitigation and climate change mitigation needs the 'We': A state-of-the-art review of social identity effects motivating climate change action. *Current Opinion in Behavioral Sciences*, *42*, 89–96. https://doi.org/10.1016/j.cobeha.2021.04.006.

McCright, A. M., & Dunlap, R. E. (2011a). The Politicization of Climate Change and Polarization in the American Public's Views of Global Warming, 2001–2010. *The Sociological Quarterly*, *52*(2), 155–194. https://doi.org/10.1111/j.1533-8525.2011.01198.x.

McCright, A. M., & Dunlap, R. E. (2011b). Cool dudes: The denial of climate change among conservative white males in the United States. *Global Environmental Change*, *21*(4), 1163–1172. https://doi.org/10.1016/j.gloenvcha.2011.06.003.

McMahan, P., & Evans, J. (2018). Ambiguity and Engagement. *American Journal of Sociology*, *124*(3), 860–912. https://doi.org/10.1086/701298.

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. https://doi.org/10.1146/annurev.soc.27.1.415.

Meadows, D. H., Meadows, D. L., Randers, J., & Behrens III, W. W. (1972). *The limits to growth. A report for the club of rome's project on the predicament of mankind*. Universe Books.

Meleady, R., & Crisp, R. J. (2017). Redefining climate change inaction as temporal intergroup bias: Temporally adapted interventions for reducing prejudice may help elicit environmental protection. *Journal of Environmental Psychology*, *53*, 206–212. https://doi.org/10.1016/j.jenvp.2017.08.005.

Mewes, L., Tuitjer, L., & Dirksmeier, P. (2024). Exploring the variances of climate change opinions in Germany at a fine-grained local scale. *Nature Communications*, *15*(1), 1867. https://doi.org/10.1038/s41467-024-45930-8.

Milosh, M., Painter, M., Sonin, K., Van Dijcke, D., & Wright, A. L. (2021). Unmasking partisanship: Polarization undermines public response to collective risk. *Journal of Public Economics*, *204*, 104538. https://doi.org/10.1016/j.jpubeco.2021.104538.

Moser, C., & Smaldino, P. E. (2023). Innovation-facilitating networks create inequality. *Proceedings of the Royal Society B: Biological Sciences*, *290*(2011), 20232281. https://doi.org/10.1098/rspb.2023.2281.

Moser, D., Steiglechner, P., & Schlueter, A. (2022). Facing global environmental change: The role of culturally embedded cognitive biases. *Environmental Development*, *44*, 100735. https://doi.org/10.1016/j.envdev.2022.100735.

Moussaïd, M., Kämmer, J. E., Analytis, P. P., & Neth, H. (2013). Social Influence and the Collective Dynamics of Opinion Formation. *PLOS ONE*, *8*(11), e78433. https://doi.org/10.1371/journal.pone.0078433.

Myers, K. F., Doran, P. T., Cook, J., Kotcher, J. E., & Myers, T. A. (2021). Consensus revisited: Quantifying scientific agreement on climate change and climate expertise among Earth scientists 10 years later. *Environmental Research Letters*, *16*(10), 104030. https://doi.org/10.1088/1748-9326/ac2774.

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.
Nilsson, H., Juslin, P., & Winman, A. (2016). Heuristics Can Produce Surprisingly Rational Probability Estimates: Comment on Costello and Watts (2014). *Psychological Review*, *123*(1), 103–111. https://doi.org/10.1037/a0039249.
Noorazar, H. (2020). Recent advances in opinion propagation dynamics: A 2020 survey. *The European Physical Journal Plus*, *135*(6), 521. https://doi.org/10.1140/epjp/s13360-020-00541-2.
Nowak, A., Rychwalska, A., & Borkowski, W. (2011). Why Simulate? To Develop a Mental Model. *Journal of Artificial Societies and Social Simulation*, *16*(3), 12.
Nowak, A., Szamrej, J., & Latané, B. (1990). From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review*, *97*(3), 362–376. https://doi.org/10.1037/0033-295X.97.3.362.
Nyczka, P. (2011). A Model of Opinion Dynamics with Bounded Confidence and Noise. *arXiv:1106.0008 [nlin]*.
Nyczka, P., Sznajd-Weron, K., & Cislo, J. (2012). Phase transitions in the q-voter model with two types of stochastic driving. *Physical Review E*, *86*. https://doi.org/10.1103/PhysRevE.86.011105.
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.
Pearson, A. R., & Schuldt, J. P. (2018). Climate change and intergroup relations: Psychological insights, synergies, and future prospects. *Group Processes & Intergroup Relations*, *21*(3), 373–388. https://doi.org/10.1177/1368430217747750.
Pearson, A. R., Schuldt, J. P., & Romero-Canyas, R. (2016). Social Climate Science: A New Vista for Psychological Science. *Perspectives on Psychological Science*, *11*(5), 632–650. https://doi.org/10.1177/1745691616639726.
Pew Research Center. (2022, August). *Climate change remains top global threat across 19-Country survey* (tech. rep.). Pew Research Center.
Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*(3), 346–354. https://doi.org/10.1037/h0023653.
Pineda, M., Toral, R., & Hernández-García, E. (2009). Noisy continuous-opinion dynamics. *Journal of Statistical Mechanics: Theory and Experiment*, *2009*(08), P08001. https://doi.org/10.1088/1742-5468/2009/08/P08001.
Pineda, M., Toral, R., & Hernández-García, E. (2011). Diffusing opinions in bounded confidence processes. *The European Physical Journal D*, *62*(1), 109–117. https://doi.org/10.1140/epjd/e2010-00227-0.
Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838. https://doi.org/10.1073/pnas.0707192105.
Powell, M., Kim, A. D., & Smaldino, P. E. (2023). Hashtags as signals of political identity: #BlackLivesMatter and #AllLivesMatter (J. Galak, Ed.). *PLOS ONE*, *18*(6), e0286524. https://doi.org/10.1371/journal.pone.0286524.
Price, V. (1989). Social Identification and Public Opinion: Effects of Communicating Group Conflicts. *Public Opinion Quarterly*, *53*(2), 197–224. https://doi.org/10.1086/269503.
Proskurnikov, A. V., & Tempo, R. (2018). A tutorial on modeling and analysis of dynamic social networks. Part II. *Annual Reviews in Control*, *45*, 166–190. https://doi.org/10.1016/j.arcontrol.2018.03.005.
Reynolds, C. W. (1987). Flocks, herds and schools: A distributed behavioral model. *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, 25–34. https://doi.org/10.1145/37401.37406.
Richerson, P., Baldini, R., Bell, A. V., Demps, K., Frost, K., Hillis, V., Mathew, S., Newton, E. K., Naar, N., Newson, L., Ross, C., Smaldino, P. E., Waring, T. M., & Zefferman, M. (2016). Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behavioral and Brain Sciences*, *39*, e30. https://doi.org/10.1017/S0140525X1400106X.
Ripple, W. J., Wolf, C., Gregg, J. W., Levin, K., Rockström, J., Newsome, T. M., Betts, M. G., Huq, S., Law, B. E., Kemp, L., Kalmus, P., & Lenton, T. M. (2022). World Scientists' Warning of a Climate Emergency 2022. *BioScience*, *72*(12), 1149–1155. https://doi.org/10.1093/biosci/biac083.
Ross, A. D., Rouse, S. M., & Mobley, W. (2019). Polarization of Climate Change Beliefs: The Role of the Millennial Generation Identity. *Social Science Quarterly*, *100*(7), 2625–2640. https://doi.org/10.1111/ssqu.12640.
Sabherwal, A., Ballew, M. T., van der Linden, S., Gustafson, A., Goldberg, M. H., Maibach, E. W., Kotcher, J. E., Swim, J. K., Rosenthal, S. A., & Leiserowitz, A. (2021). The Greta Thunberg Effect: Familiarity with Greta Thunberg predicts intentions to engage in climate activism in the United States. *Journal of Applied Social Psychology*, *51*(4), 321–333. https://doi.org/10.1111/jasp.12737.
Schawe, H., Fontaine, S., & Hernández, L. (2021). When network bridges foster consensus. Bounded confidence models in networked societies. *Physical Review Research*, *3*(2), 023208. https://doi.org/10.1103/PhysRevResearch.3.023208.
Schelling, T. C. (1971). Dynamic models of segregation. *Journal of mathematical sociology*, *1*(2), 143–186.
Schill, C., Anderies, J. M., Lindahl, T., Folke, C., Polasky, S., Cárdenas, J. C., Crépin, A.-S., Janssen, M. A., Norberg, J., & Schlüter, M. (2019). A more dynamic understanding of human behaviour for the Anthropocene. *Nature Sustainability*, *2*(12), 1075–1082. https://doi.org/10.1038/s41893-019-0419-7.
Schweighofer, S., Garcia, D., & Schweitzer, F. (2020). An agent-based model of multi-dimensional opinion dynamics and opinion alignment. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *30*(9), 093139. https://doi.org/10.1063/5.0007523.
Sharman, A., & Howarth, C. (2017). Climate stories: Why do climate scientists and sceptical voices participate in the climate debate? *Public Understanding of Science*, *26*(7), 826–842. https://doi.org/10.1177/0963662516632453.
Sherman, D. K., Hogg, M. A., & Maitner, A. T. (2009). Perceived Polarization: Reconciling Ingroup and Intergroup Perceptions Under Uncertainty. *Group Processes & Intergroup Relations*, *12*(1), 95–109. https://doi.org/10.1177/1368430208098779.
Shirado, H., & Christakis, N. A. (2017). Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, *545*(7654), 370–374. https://doi.org/10.1038/nature22332.
Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99. https://doi.org/10.2307/1884852.
Smaldino, P. E. (2017, May). Models Are Stupid, and We Need More of Them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational Social Psychology* (1st ed., pp. 311–331). Routledge. https://doi.org/10.4324/9781315173726-14.

Smaldino, P. E. (2019). Better methods can't make up for mediocre theory. *Nature*, *575*(7781), 9–9. https://doi.org/10.1038/d41586-019-03350-5.
Smaldino, P. E. (2020). How to Translate a Verbal Theory Into a Formal Model. *Social Psychology*, *51*(4), 207–218. https://doi.org/10.1027/1864-9335/a000425.
Smaldino, P. E. (2022). Models of Identity Signaling. *Current Directions in Psychological Science*, *31*(3), 231–237. https://doi.org/10.1177/09637214221075609.
Smaldino, P. E. (2023). *Modeling social behavior: Mathematical and agent-based models of social dynamics and cultural evolution*. Princeton University Press.
Smaldino, P. E., Calanchini, J., & Pickett, C. L. (2015). Theory development with agent-based models. *Organizational Psychology Review*, *5*(4), 300–317. https://doi.org/10.1177/2041386614546944.
Smaldino, P. E., Janssen, M. A., Hillis, V., & Bednar, J. (2017). Adoption as a social marker: Innovation diffusion with outgroup aversion. *The Journal of Mathematical Sociology*, *41*(1), 26–45. https://doi.org/10.1080/0022250X.2016.1250083.
Smaldino, P. E., & Jones, J. H. (2021). Coupled dynamics of behaviour and disease contagion among antagonistic groups. *Evolutionary Human Sciences*, *3*, e28. https://doi.org/10.1017/ehs.2021.22.
Smaldino, P. E., Moser, C., Pérez Velilla, A., & Werling, M. (2023). Maintaining Transient Diversity Is a General Principle for Improving Collective Problem Solving. *Perspectives on Psychological Science*, 17456916231180100. https://doi.org/10.1177/17456916231180100.
Smaldino, P. E., & Turner, M. A. (2022). Covert signaling is an adaptive communication strategy in diverse populations. *Psychological Review*, *129*(4), 812–829. https://doi.org/10.1037/rev0000344.
Smith, E. K., Bognar, M. J., & Mayer, A. P. (2024). Polarisation of Climate and Environmental Attitudes in the United States, 1973-2022. *npj Climate Action*, *3*(1), 1–14. https://doi.org/10.1038/s44168-023-00074-1.
Sobkowicz, P. (2012). Discrete Model of Opinion Changes Using Knowledge and Emotions as Control Variables. *PLOS ONE*, *7*(9), e44489. https://doi.org/10.1371/journal.pone.0044489.
Sobkowicz, P. (2018). Opinion dynamics model based on cognitive biases of complex agents. *Journal of Artificial Societies and Social Simulation*, *21*(4), 8. https://doi.org/10.18564/jasss.3867.
Sobkowicz, P. (2020). Whither now, opinion modelers? *Frontiers in Physics*, *8*, 461. https://doi.org/10.3389/fphy.2020.587009.
Sobkowicz, P., Kaschesky, M., & Bouchard, G. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, *29*(4), 470–479. https://doi.org/10.1016/j.giq.2012.06.005.
Squazzoni, F., Jager, W., & Edmonds, B. (2014). Social Simulation in the Social Sciences: A Brief Overview. *Social Science Computer Review*, *32*(3), 279–294. https://doi.org/10.1177/0894439313512975.
Steiglechner, P. (2023, November). An opinion formation model with social identity and in-group bias. https://doi.org/10.5281/zenodo.10118407.
Steiglechner, P., Keijzer, M. A., Smaldino, P. E., Moser, D., & Merico, A. (2024). Noise and opinion dynamics: How ambiguity promotes pro-majority consensus in the presence of confirmation bias. *Royal Society Open Science*, *11*(4), 231071. https://doi.org/10.1098/rsos.231071.
Steiglechner, P., Smaldino, P. E., Moser, D., & Merico, A. (2023). Social identity bias and communication network clustering interact to shape patterns of opinion dynamics. *Journal of The Royal Society Interface*, *20*(209), 20230372. https://doi.org/10.1098/rsif.2023.0372.
Stern, S., & Livan, G. (2021). The impact of noise and topology on opinion dynamics in social networks. *Royal Society Open Science*, *8*(4), 201943. https://doi.org/10.1098/rsos.201943.
Su, W., Chen, G., & Hong, Y. (2017). Noise leads to quasi-consensus of Hegselmann–Krause opinion dynamics. *Automatica*, *85*, 448–454. https://doi.org/10.1016/j.automatica.2017.08.008.
Suldovsky, B. (2017, September). The Information Deficit Model and Climate Change Communication. In *Oxford Research Encyclopedia of Climate Science*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228620.013.301.
Swim, J. K., Aviste, R., Lengieza, M. L., & Fasano, C. J. (2022). OK Boomer: A decade of generational differences in feelings about climate change. *Global Environmental Change*, *73*, 102479. https://doi.org/10.1016/j.gloenvcha.2022.102479.
Sznajd-Weron, K., & Sznajd, J. (2000). Opinion evolution in closed community. *International Journal of Modern Physics C*, *11*(06), 1157–1165. https://doi.org/10.1142/S0129183100000936.
Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information*, *13*(2), 65–93. https://doi.org/10.1177/053901847401300204.
ten Broeke, G., van Voorn, G., & Ligtenberg, A. (2016). Which Sensitivity Analysis Method Should I Use for My Agent-Based Model? *Journal of Artificial Societies and Social Simulation*, *19*(1), 5.
Thunberg, G. (2019, April). Speech to the Members of Parliament in the UK.
Toomey, A. H. (2023). Why facts don't change minds: Insights from cognitive science for the improved communication of conservation research. *Biological Conservation*, *278*, 109886. https://doi.org/10.1016/j.biocon.2022.109886.
Turner, J. C., Wetherell, M. S., & Hogg, M. A. (1989). Referent informational influence and group polarization. *British Journal of Social Psychology*, *28*(2), 135–147. https://doi.org/10.1111/j.2044-8309.1989.tb00855.x.
Turner, M. A., Moya, C., Smaldino, P. E., & Jones, J. H. (2023). The form of uncertainty affects selection for social learning. *Evolutionary Human Sciences*, *5*, e20. https://doi.org/10.1017/ehs.2023.11.
Turner, M. A., & Smaldino, P. E. (2018). Paths to Polarization: How Extreme Views, Miscommunication, and Random Chance Drive Opinion Dynamics. *Complexity*, *2018*, e2740959. https://doi.org/10.1155/2018/2740959.
Vieira, A. R., & Crokidakis, N. (2016). Noise-induced absorbing phase transition in a model of opinion formation. *Physics Letters A*, *380*(34), 2632–2636. https://doi.org/10.1016/j.physleta.2016.06.014.
Vul, E., & Pashler, H. (2008). Measuring the Crowd Within: Probabilistic Representations Within Individuals. *Psychological Science*, *19*(7), 645–647. https://doi.org/10.1111/j.1467-9280.2008.02136.x.
Waldrop, M. M. (2021). Modeling the power of polarization. *Proceedings of the National Academy of Sciences*, *118*(37), e2114484118. https://doi.org/10.1073/pnas.2114484118.

Wallis, H., & Loy, L. S. (2021). What drives pro-environmental activism of young people? A survey study on the Fridays For Future movement. *Journal of Environmental Psychology*, *74*, 101581. https://doi.org/10.1016/j.jenvp.2021.101581.

Watts, D. J. (2004, March). *Six Degrees: The Science of a Connected Age*. WWNorton & Compnay.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918.

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.

Will, M., Groeneveld, J., Frank, K., & Müller, B. (2020). Combining social network analysis and agent-based modelling to explore dynamics of human interaction: A review. *Socio-Environmental Systems Modelling*, *2*, 16325–16325. https://doi.org/10.18174/sesmo.2020a16325.

Williams, H. T. P., McMurray, J. R., Kurz, T., & Hugo Lambert, F. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, *32*, 126–138. https://doi.org/10.1016/j.gloenvcha.2015.03.006.

Wimsatt, W. C. (1987). False models as means to truer theories. In M. Nitecki & A. Hoffman (Eds.), *Neutral models in biology* (pp. 23–55). Oxford University Press.

World Economic Forum. (2024). *The Global Risks Report 2024* (tech. rep. No. 19th edition).

Zhang, J., & Zhao, Y. (2018). The Robust Consensus of a Noisy Deffuant-Weisbuch Model. *Mathematical Problems in Engineering*, *2018*, e1065451. https://doi.org/10.1155/2018/1065451.

Zhao, Y., Zhang, L., Tang, M., & Kou, G. (2016). Bounded confidence opinion dynamics with opinion leaders and environmental noises. *Computers & Operations Research*, *74*, 205–213. https://doi.org/10.1016/j.cor.2015.07.022.

Appendix A

Supplementary Material for Chapter 2

This chapter contains supplementary material for

Steiglechner, P., Keijzer, M. A., Smaldino, P. E., Moser, D., & Merico, A. (2024). Noise and opinion dynamics: How ambiguity promotes pro-majority consensus in the presence of confirmation bias. *Royal Society Open Science*, *11*(4), 231071. https://doi.org/10.1098/rsos.231071.

It is published as Electronic Supplementary Material and available online at https://doi.org/10.6084/m9.figshare.c.7095862

A.1 Calibrating initial opinions on climate change to empirical data

To apply the model to the climate change debate, we use empirical data from Maibach et al. (2011) to determine the distribution of the initial opinions. In 2008, the authors surveyed more than 2000 US citizens about their global warming beliefs, environmental behaviours, policy preferences and issue engagement. They identified six distinct categories of attitudes towards climate change in the survey population: dismissive ($7\,\%$ of the sampled population), doubtful ($11\,\%$), disengaged ($12\,\%$), cautious ($18\,\%$), concerned ($33\,\%$) and alarmed ($19\,\%$). To reflect this distribution of opinions at the beginning of our simulations, we divide the continuous opinion space, $x \in [0, 1]$, into six equally sized segments, where each segment represents the opinion range of one of the six categories. For example, agents with $x \in [5/6,\, 1]$ are 'alarmed' and agents with $x \in [4/6,\, 5/6]$ are 'concerned'. We draw initial opinions from a smooth distribution, which we fitted to match the population shares from Maibach et al. (2011) within each segment (using a superposition of two Gaussian functions, see subpanel $t = 0$ in figure 2.3*b*).

A.2 Kurtosis

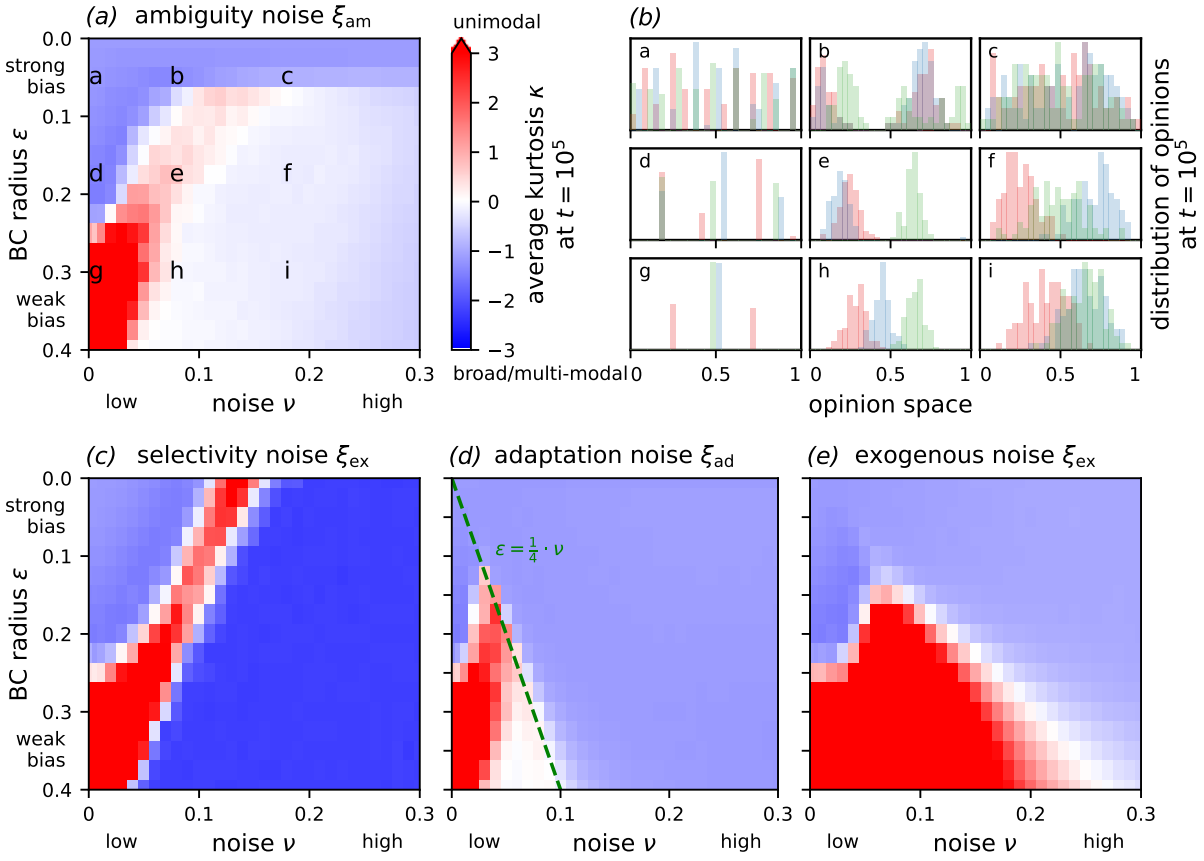


Figure A.1: Kurtosis κ of the agent opinion distribution for societies with different levels of bias and noise (see Figure 2 in the main manuscript). The kurtosis measures the peakedness of the opinion distribution and, indirectly, the number of peaks (or opinion clusters) in the distribution of agent opinions, $\{x_i|i\}$, at a specific time. Positive kurtosis indicates agreement with an opinion distribution characterised by a sharp, unimodal peak. Negative kurtosis indicates disagreement characterised by (1) a bi- or multi-modal or (2) a broad distribution of opinions. In general, positive kurtosis indicates regions with low dispersion, i.e. stronger agreement. However, kurtosis can lead to some spurious effects. For example, for high selectivity noise, a narrow consensus emerges, but the kurtosis is negative because the distribution resembles a 'broad', delta-like peak rather than a 'sharp', delta-like peak (panel *c*).

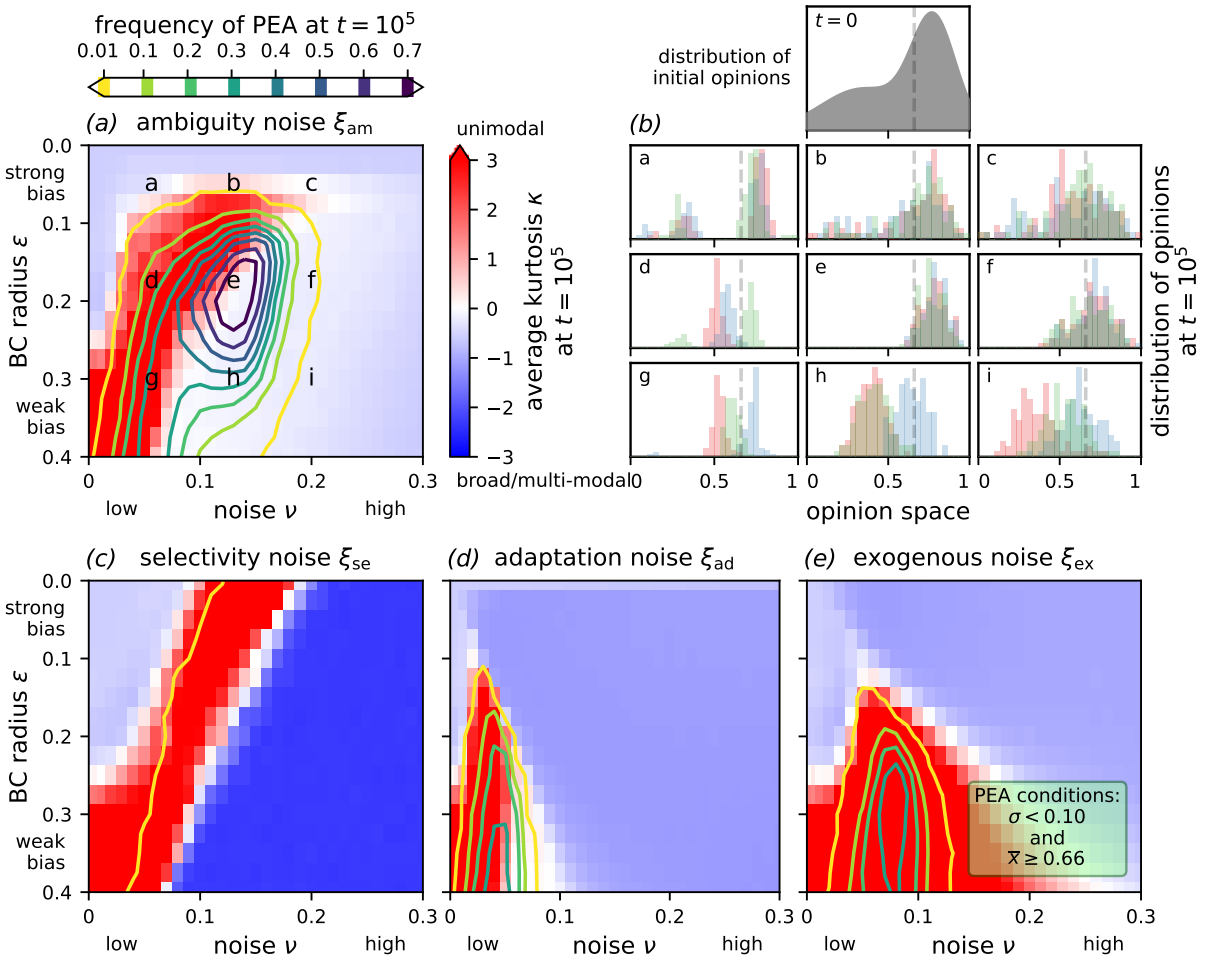


Figure A.2: Kurtosis κ of the agent opinion distribution for societies with different levels of bias and noise in the experiment with calibrated initial data (see Figure 3 in the main manuscript). The contour lines indicate the frequency of pro-environmental agreement ($\overline{x} > 0.66$ and $\sigma \leq 0.1$) for combinations of bias and noise as in Figure 3. Kurtosis measures the modality, but this is somewhat unrelated to agreement or disagreement. In this case, the regions with positive kurtosis, which indicates the sharpness of the distribution, are somewhat related to the regions with high PEA frequency especially for selectivity (panel *c*), adaptation (*d*), and exogenous noise (*e*). Yet, for ambiguity noise, kurtosis and PEA frequency do not match particularly well (panel *a*).

A.3 Sensitivity Analysis of the model parameters

In this supplementary section, we provide a sensitivity analysis of our results. We show how the dispersion, σ, of agent opinions over the range of noise, ν, and bias, expressed via ϵ, changes under different parameter choices. To explore the impacts of the parameter changes systematically, we use One-Factor-At-A-Time as recommended, ten Broeke et al., 2016. The parameters used for the results in the main manuscript (Figure 2) are the number of agents $n = 100$, the convergence parameter $\mu = 0.5$, and the time horizon $t = 10^5$. We provide the results for the parameter values $n = 50$ and $n = 1000$ (Figure A.3), $t = 10^4$ and $t = 10^6$ (Figure A.5) and $\mu = 0.1$ (Figure A.7), respectively.

Additionally, we provide figures showing the ensemble mean dispersion, σ, as a function of the three parameters, n (Figure A.4), t (Figure A.6), and μ (Figure A.8), for the nine configurations of bias and noise, a_1 to a_9, defined in Figure 2*a* and 2*b* in the main manuscript.

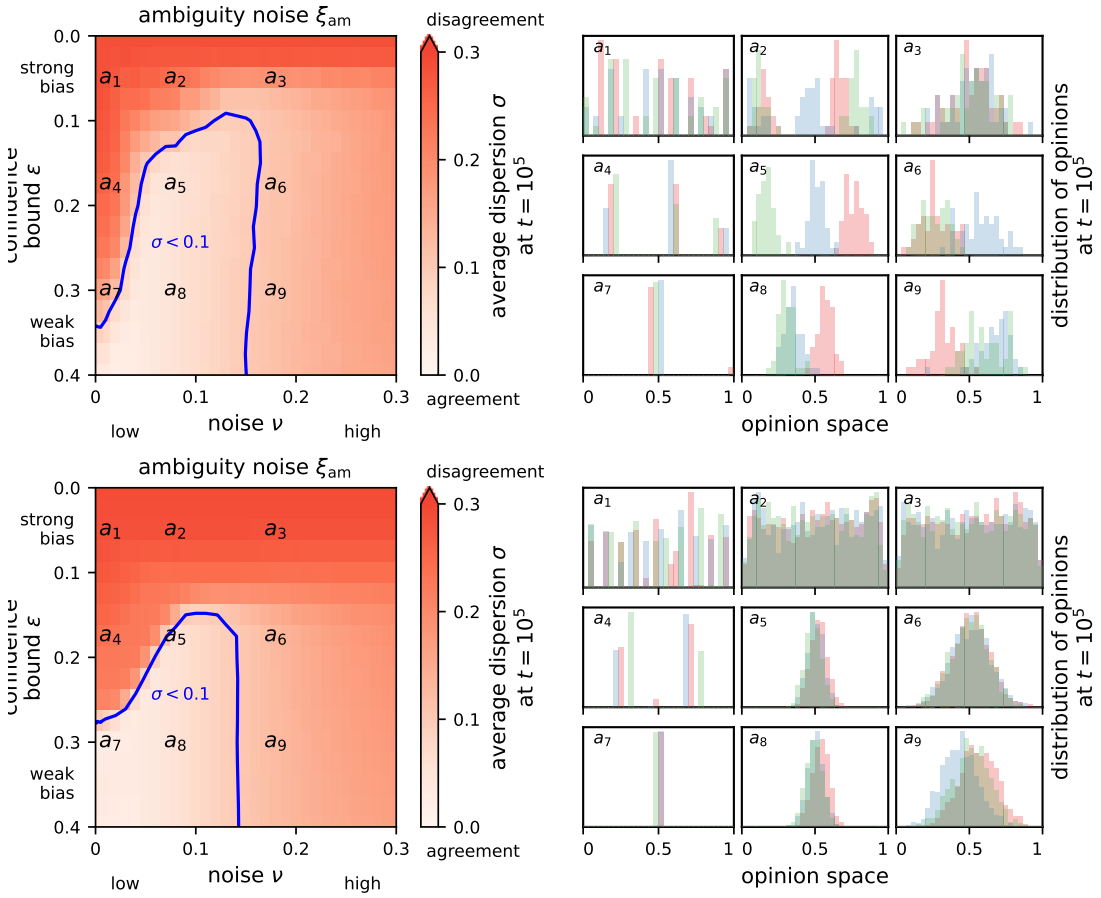


Figure A.3: Dispersion σ of the agent opinion distribution for societies with $n = 50$ (left) and $n = 1000$ agents (right) for different levels of bias and noise (see Figure 2 with $n = 100$ in the main manuscript).

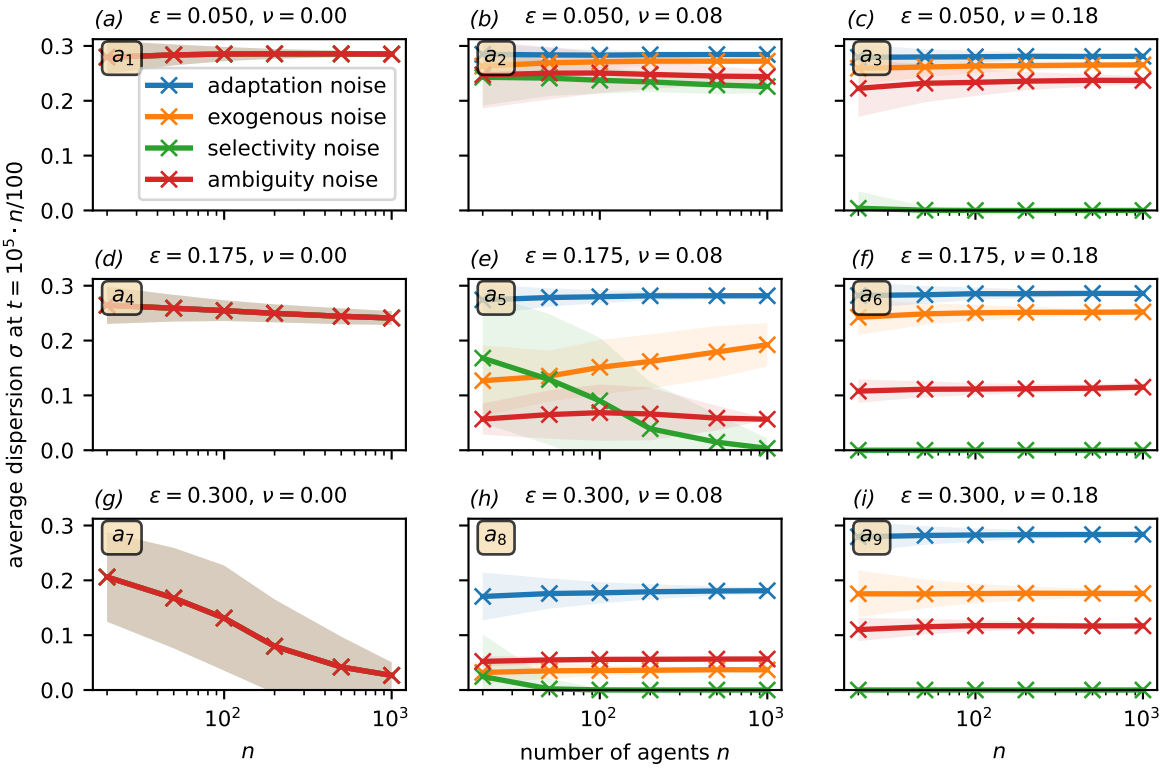


Figure A.4: Sensitivity analysis of the opinion dispersion σ for different numbers of agents, n. Each panel a to i corresponds to a specific location a_1 to a_9 in the parameter space of noise (ν) and bias (ϵ) as presented in Figure 2*a* and 2*b* in the main manuscript. At $t = 0$, the agents are initialised with opinions drawn from a uniform distribution. Simulations are run until an adjusted time $t = 10^5 \cdot n/100$ such that the agents have, on average, the same number of interactions regardless of n. The lines indicate the dispersion, σ at time $t = 10^5 \cdot n/100$, averaged over 1000 simulations for different types of noise (colours). The coloured areas indicate the standard deviation of the dispersion over the 1000 simulations. A large standard deviation typically implies that some simulated societies have reached agreement (low dispersion) while others remain in disagreement (high dispersion). The dispersion is robust over different orders of magnitude of n. For example, for the most relevant configuration of moderate ambiguity noise and moderate bias, a_5 (red line in panel *e*), agreement is very likely regardless of the number of agents. Some interesting effects are discernible at a higher level of detail. For moderate bias and moderate selectivity noise (green line in panel *e*), small populations vary substantially in the dispersion, which is the result of a sparsely populated opinion space in small societies and thus more factionalisation. In this case, ambiguity noise outperforms selectivity noise in terms of opinion convergence.

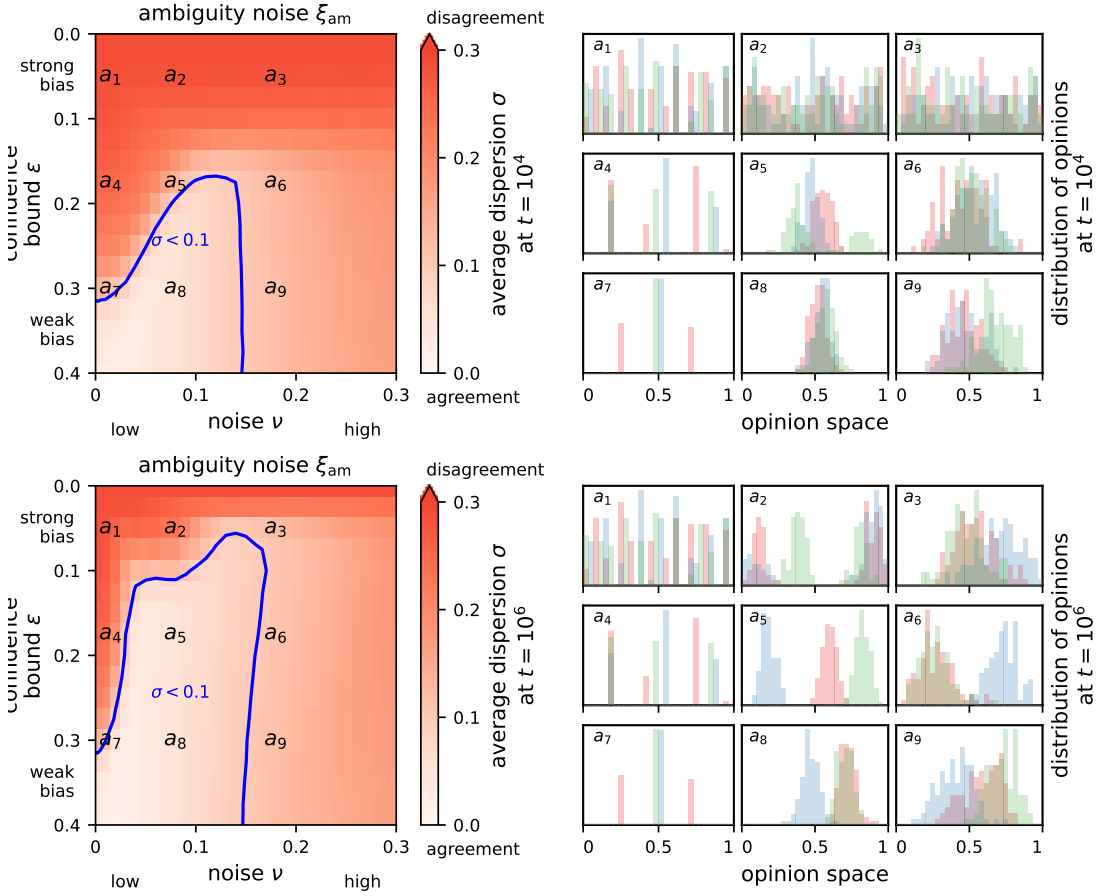


Figure A.5: Dispersion σ of the agent opinion distribution at time $t = 10^4$ (left) and $t = 10^6$ (right) for societies with different levels of bias and noise (see Figure 2 for $t = 10^5$ in the main manuscript).

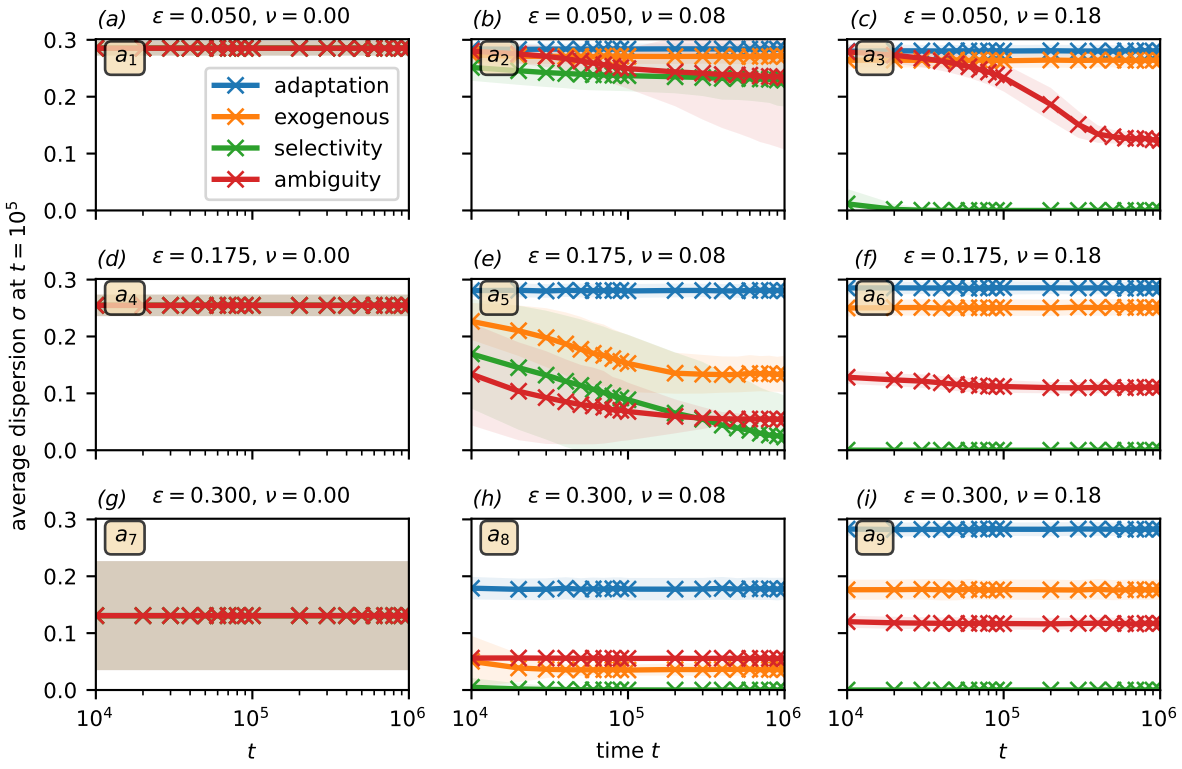


Figure A.6: Sensitivity analysis of the dispersion for varying simulation times t (see Figure A.4 for details). The dispersion is mostly robust over various orders of magnitude $10^4 \leq t \leq 10^6$. However, there are a few exceptions. Early in the simulations, there is more variation in the dispersion for moderate levels of bias and ambiguity, selectivity, or exogenous noise (red, green and orange lines in panel *e*). In the case of strong bias (panels *a*, *b*, and *c*), higher ambiguity noise (red line) implies more agreement, at least if the simulation is run sufficiently long (see the decrease in the mean dispersion for a_3 in panel *c* and the higher variance in the dispersion for a_2 in panel *b* after $t = 10^5$).

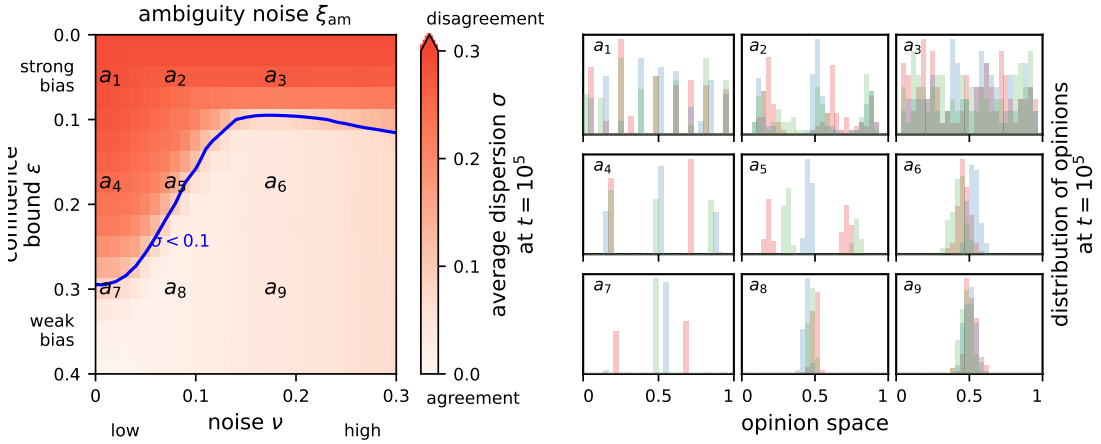


Figure A.7: Dispersion σ of the agent opinion distribution for societies with different levels of bias and noise (see Figure 2 in the main manuscript). Here, the convergence speed μ is smaller $\mu = 0.1$ instead of $\mu = 0.5$.

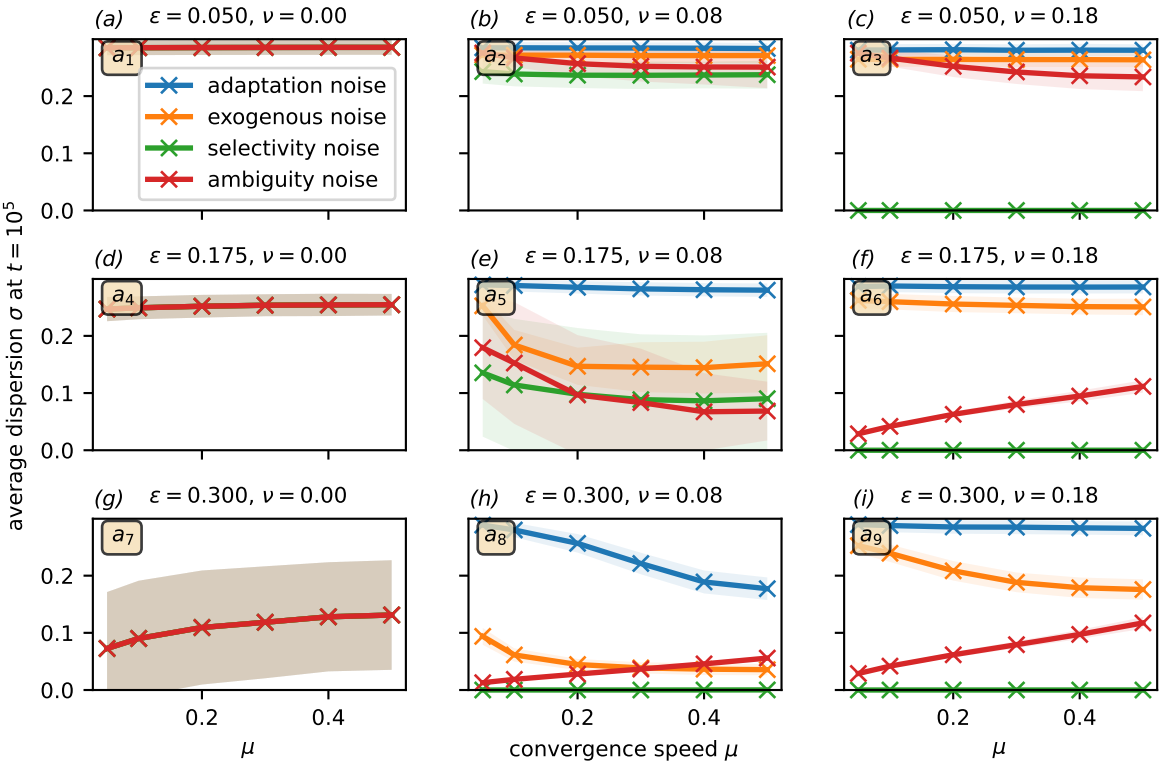


Figure A.8: Sensitivity analysis of the opinion dispersion for different convergence speeds μ (see Figure A.4 for details). The dispersion is quite robust to variation in μ. For ambiguity noise in the configurations a_6, a_8, and a_9 (panels *f*, *h*, and *i*), the mean dispersion, σ, grows linearly with μ. This is expected, as ambiguity noise in the message, $m_j = x_j + \xi_{\text{am}}$, leads to a noisy contribution in the receiver's adaptation rule $\mu \cdot \xi_{am}$ which scales linearly with μ. However, this is not at all a trivial result, for example, given that the dispersion declines with increasing μ for a_5, i.e. moderate bias and moderate ambiguity noise (red line in panel *e*).

A.4 Additional model analysis

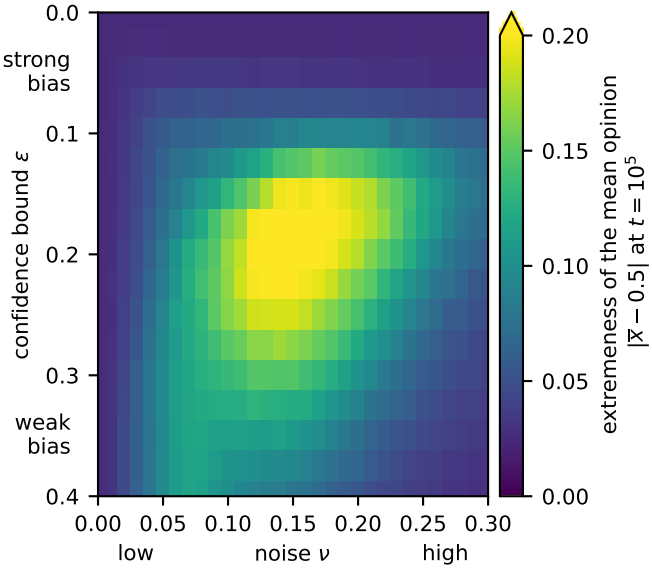


Figure A.9: The extremeness of the mean opinion under ambiguity noise, $|\overline{x} - 0.5|$ at $t = 10^5$, averaged over 1000 simulation runs (see Figure 2*a* in the manuscript). We assume uniformly distributed initial opinions and, therefore, the initial mean opinion is close to 0.5. For moderate ambiguity noise ($\epsilon \approx 0.2$) in societies with moderate bias ($\epsilon \approx 0.15$), drift causes the mean opinion to extremise, i.e. to get closer to the bounds of the opinion space, $\overline{x} = 0.5 \pm 0.2$. Such a drift is not observed for any of the other types of noise regardless of the level of noise or bias (not shown).

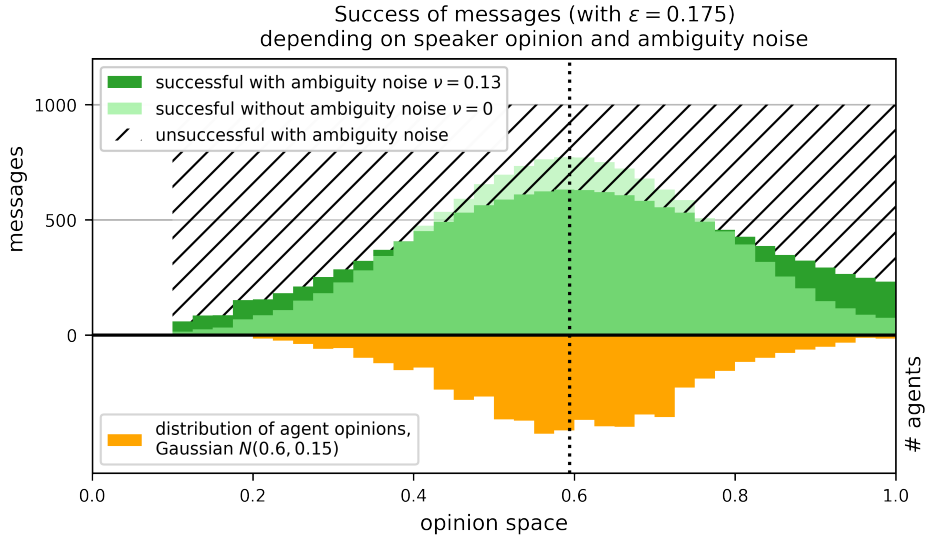


Figure A.10: This Figure shows how successful speakers are in sending messages with ambiguity noise $\nu = 0.13$ to recipients with confidence bounds $\epsilon = 0.175$. The opinion distribution (orange bars) is drawn from a Gaussian distribution centred at 0.6 and with a standard deviation of 0.15. The black dashed vertical line indicates the mean opinion. Each agent j sends 1000 messages and we track how frequently they would be accepted by randomly selected receivers as a function of the messenger opinion x_j for two cases: with ambiguity noise $\nu = 0.13$ (dark green bars) or without ambiguity noise $\nu = 0$ (light green bars). Agents with extreme opinions close to $x_j = 1$ benefit from ambiguity. These agents tend to be more successful as agents close to the centre of the opinion space.

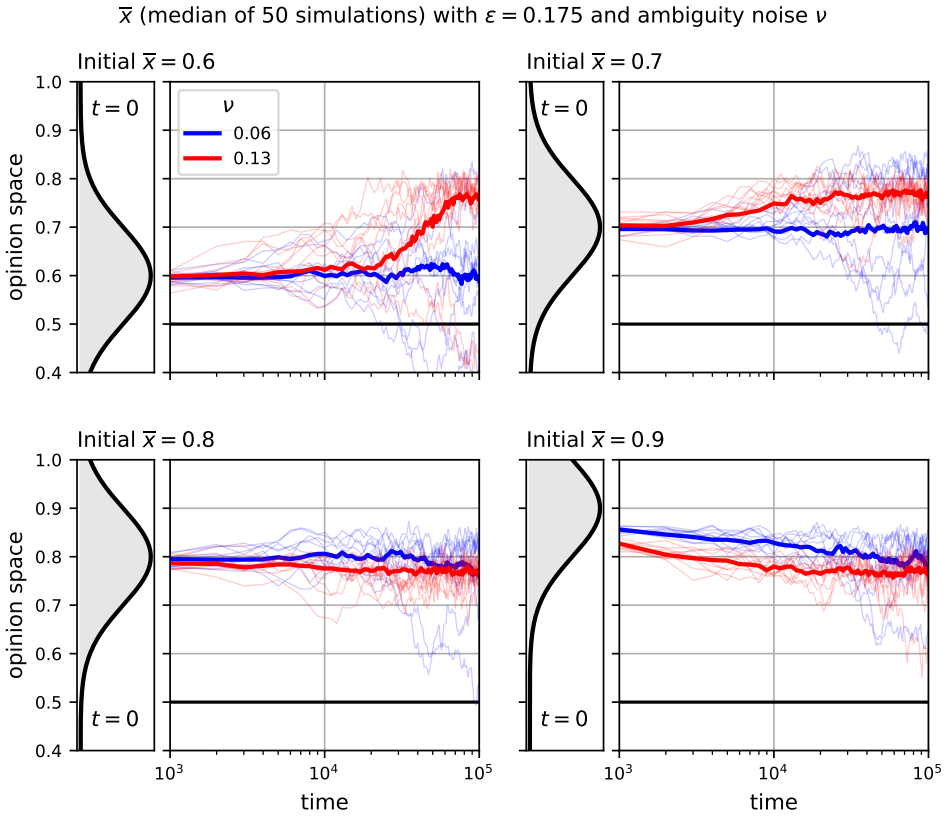


Figure A.11: Average opinions, $\overline{x}$, over time in simulations with low/moderate ambiguity noise ($\nu = 0.06$ in blue and $\nu = 0.13$ in red), fixed moderate bias $\epsilon = 0.175$ (see locations a_4 and a_5 in Figure 3*a* and 3*b* in the manuscript), and normally distributed initial conditions, $\mathcal{N}(\overline{x}(t = 0), 0.1)$, centred around $\overline{x}(t = 0) = 0.6$, 0.7, 0.8, and 0.9 (see shaded area in the corresponding panels attached to the left). The thick blue and red lines represent the median of 50 simulations (some of which are shown as thin lines). For $\overline{x}(t = 0) = 0.6$ and 0.7, the average opinion increases towards roughly $\overline{x}(t = 10^5) = 0.8$ for $\nu = 0.13$ but not for $\nu = 0.06$. We call this phenomenon drift, which occurs under moderate ambiguity noise and moderate bias. In the presence of ambiguity noise, the agents closer to the bounds of the opinion space in the initial distribution are more successful in transmitting messages to others than the agents in the centre of the opinion space (see section 4.2 in the main article) and, thus, they tend to be more successful in dragging others towards their more extreme opinions. For $\overline{x}(t = 0) = 0.8$, the average opinion remains roughly stable. For $\overline{x}(t = 0) = 0.9$, the average opinion decreases away from the bound of the opinion space both for low and moderate ambiguity noise. Very extreme agents can only move towards more moderate opinions and the agents collectively shift away from the extreme bound, $x = 1$.

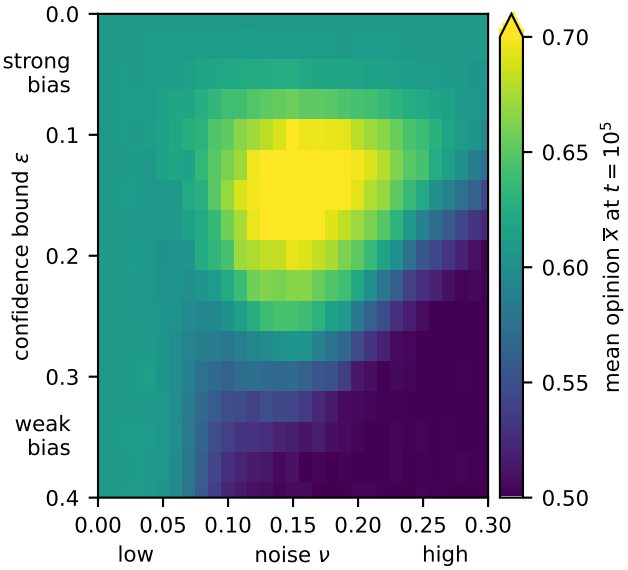


Figure A.12: Mean pro-environmental attitude, $\overline{x}$, calculated as the average of the focal opinions x_i of all agents i at $t = 10^5$ for different levels of bias and ambiguity noise. The agents are initialised with opinions on climate change calibrated to survey data (see Figure 3, panel $t = 0$ in the main manuscript). One condition for a pro-environmental agreement in Figure 3 is that the mean opinion $\overline{x} \geq 0.66$.

A.5 The behaviour of ambiguity noise under different communication regimes

We test the robustness of the effects of ambiguity noise on opinion dynamics to three influence regimes: one-to-one most commonly assumed in OD models, for example in Deffuant et al., 2000, many-to-one Flache and Macy, 2011 and one-to-many Keijzer et al., 2018 communication.

We operationalise these communication regimes as follows. In each time step:

- o2o – as in the main manuscript: one agent receives a message from one other agent and is influenced by it if this message is within its confidence bound.
- m2o – following Flache and Macy, 2011: all agents formulate a message with ambiguity noise. The opinion of a receiving focal agent is influenced by the average of all those messages within the confidence bound.
- o2m – following Keijzer et al., 2018: one focal agent sends a message with the ambiguity noise to all others. Those for whom this message falls with their BC are influenced by the message.

Figures A.13, A.14, and A.16 show example runs of the model under each of the three communication runs. The chosen combinations *a* to *i* correspond to the parameter values denoted by a_1 through a_9 in Figure 2*a* in the main manuscript.

On the full timescale modelled in Figures A.13, A.14, and A.16, the trajectories under the one-to-many regime with moderate to strong noise and moderate to weak bias are hard to see, as its behaviour is very dynamic and

quick. To that end, we zoom into these four panels and show the dynamic at a smaller timescale. Figure A.15 shows the first 500 time steps of the panels *e*, *f*, *h* and *i* of Figure A.14. We see that it only takes a few time steps for each of these systems to converge to the same position, after which this collective opinion starts to walk around the opinion space chaotically. This behaviour perpetuates for the entirety of the run.

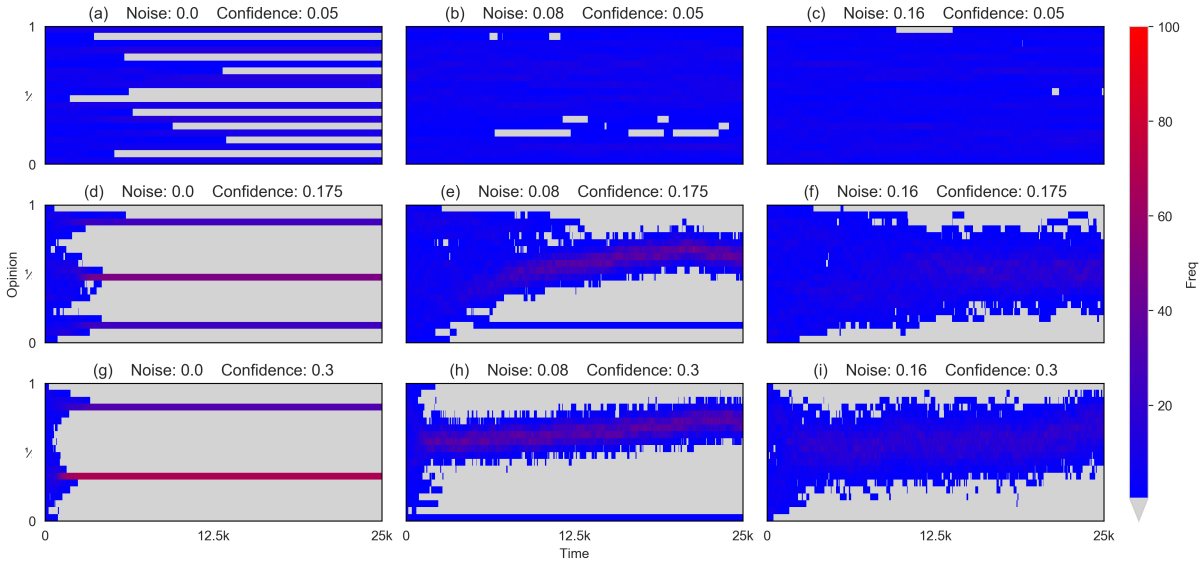


Figure A.13: Density of opinions over time for the ambiguity noise model with one-to-one communication. Each panel shows an example run for a given combination of ambiguity noise and confidence values.

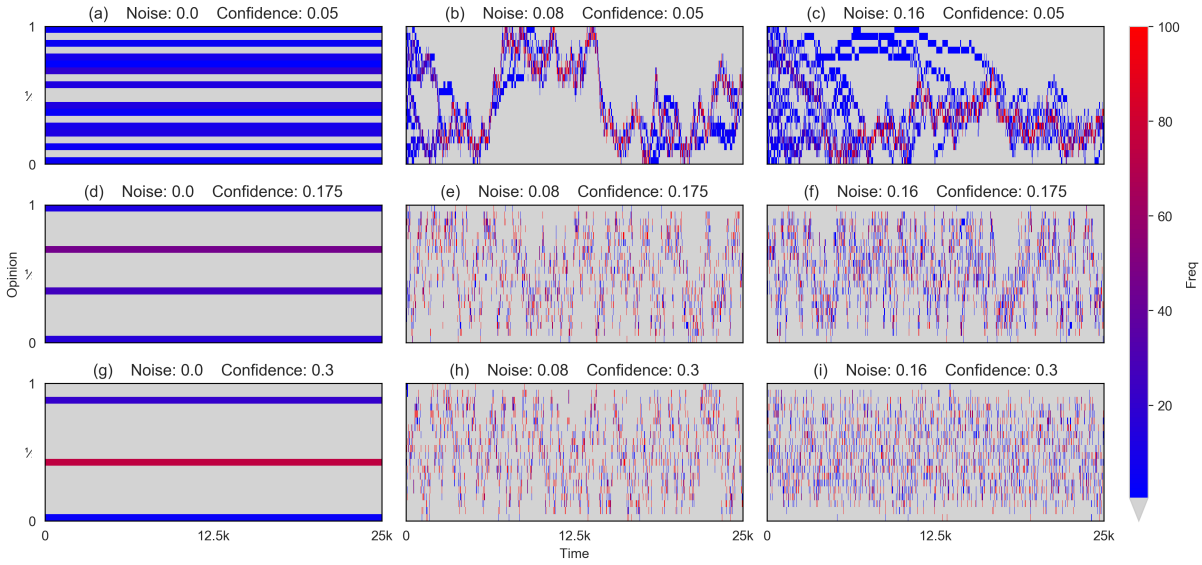


Figure A.14: Density of opinions over time for the ambiguity noise model with one-to-many communication. Each panel shows an example run for a given combination of ambiguity noise and confidence values.

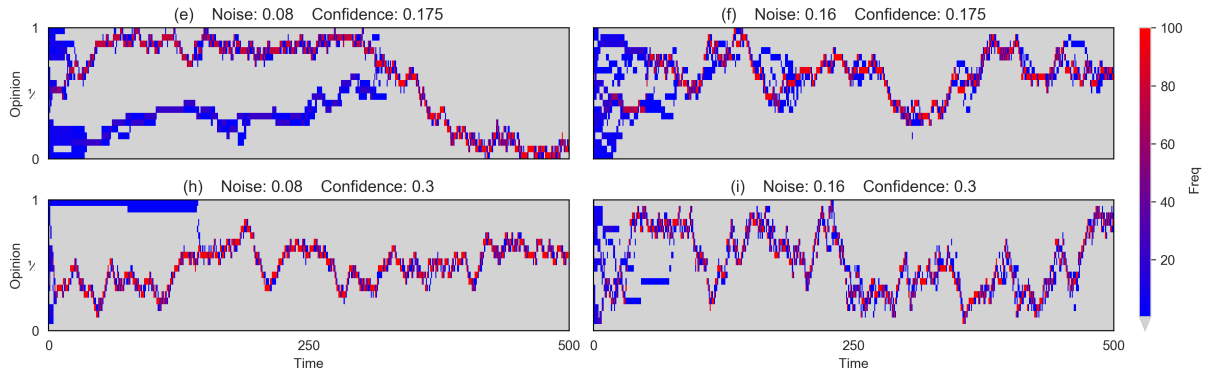


Figure A.15: Density of opinions over time for the ambiguity noise model with one-to-many communication. Each panel shows the first 500 time steps of the example runs from Figure A.14, panels *e*, *f*, *h*, and *i*.

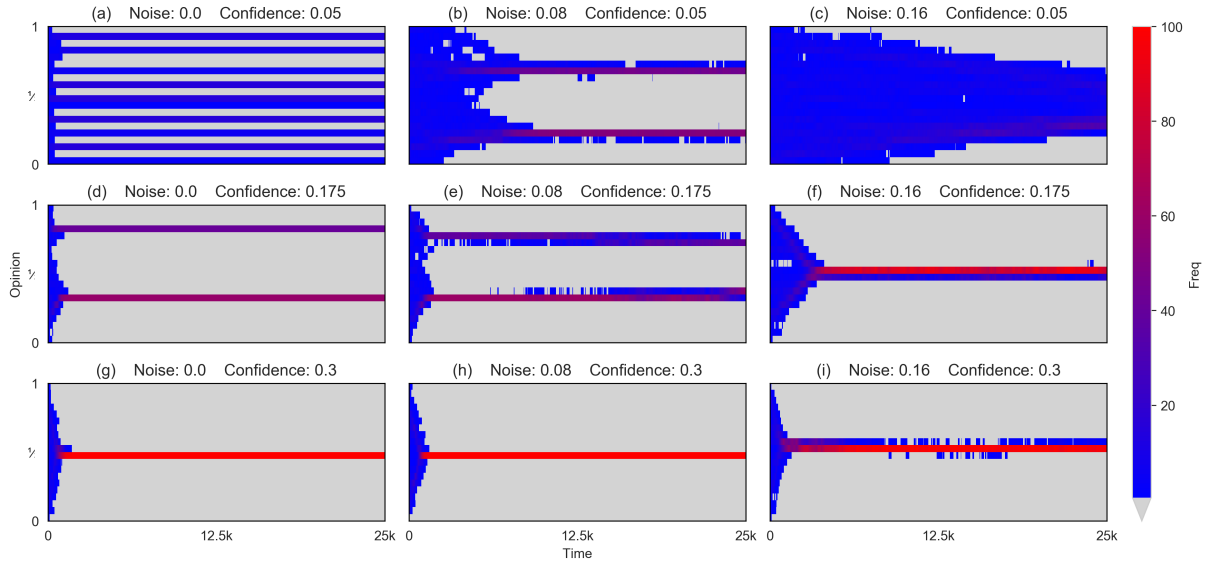


Figure A.16: Density of opinions over time for the ambiguity noise model with many-to-one communication. Each panel shows an example run for a given combination of ambiguity noise and confidence values.

A.6 Small-world network

In the main manuscript, we assume that every agent can interact with any other agent with equal probability. To test the sensitivity of our results to this assumption, we embed agents in a social network and restrict communication to those agents that are connected via a direct link. In particular, we create a stochastic network of edges between these agents following the Watts-Strogatz model Watts and Strogatz, 1998. We keep the network fixed throughout the simulation. Figures A.17 and A.18 show the dispersion, σ (as in Figure 2 in the main manuscript), for different average node degrees and different rewiring probabilities. Our goal is to show that the qualitative pattern is robust even under extreme assumptions about k and p in the small-world network.

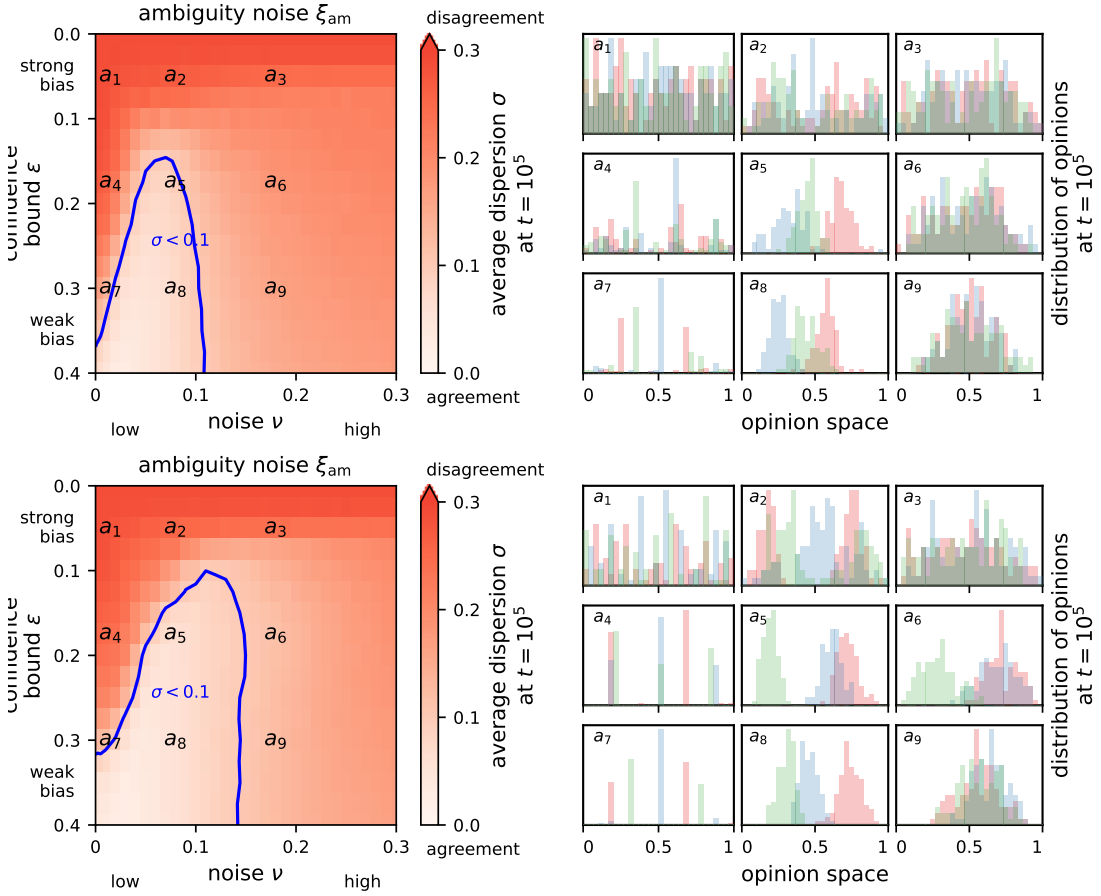


Figure A.17: Dispersion σ of the agent opinion distribution for societies where communication is restricted to a small-world network with $n = 100$ agents, average node degree $k = 4$ (left) and $k = 20$ (right), and randomness $p = 0.1$ for different levels of bias and noise (see Figure 2 in the main manuscript). A higher average node degree slightly promotes consensus but does not affect the opinion patterns substantially.

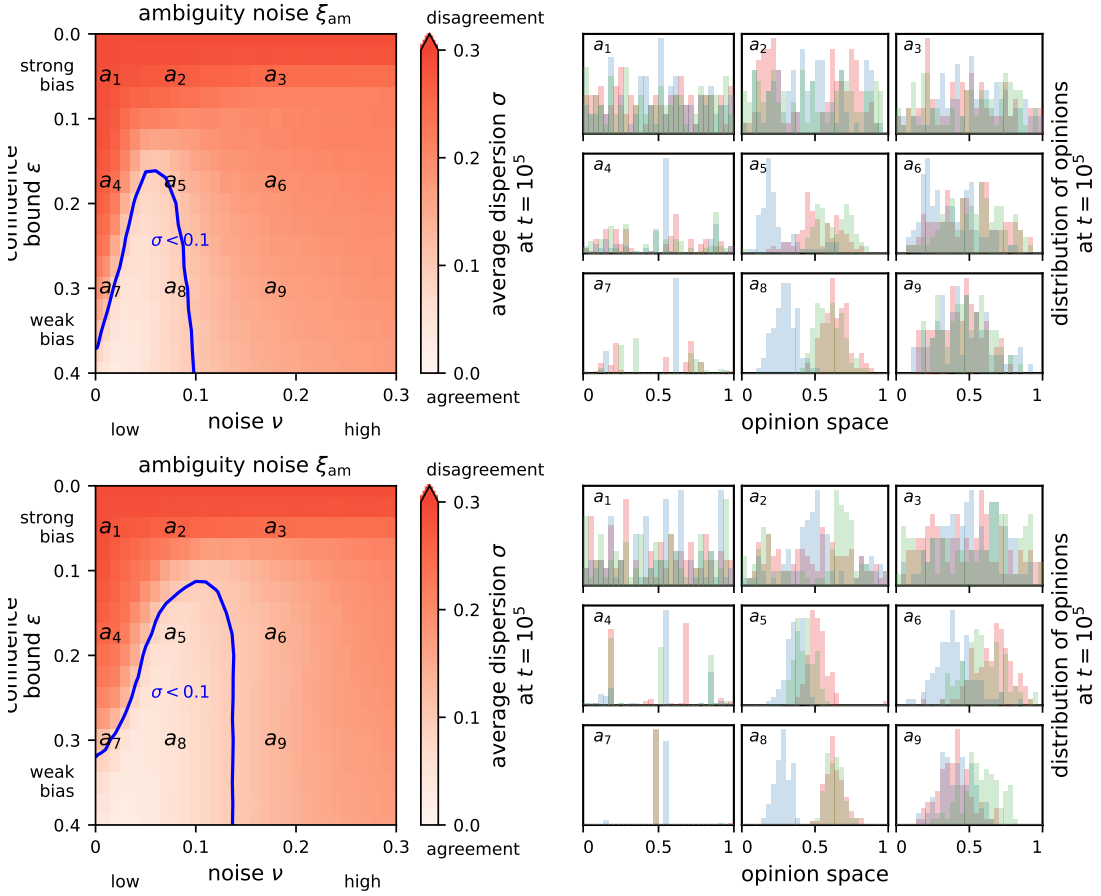


Figure A.18: Dispersion σ of the agent opinion distribution for societies where communication is restricted to a small-world network with $n = 100$ agents, average node degree $k = 6$, and randomness $p = 0.0$ (left) and $p = 1.0$ (right) for different levels of bias and noise (see Figure 2 in the main manuscript). Higher randomness slightly promotes consensus but does not affect the opinion patterns substantially.

A.7 Different noise distributions

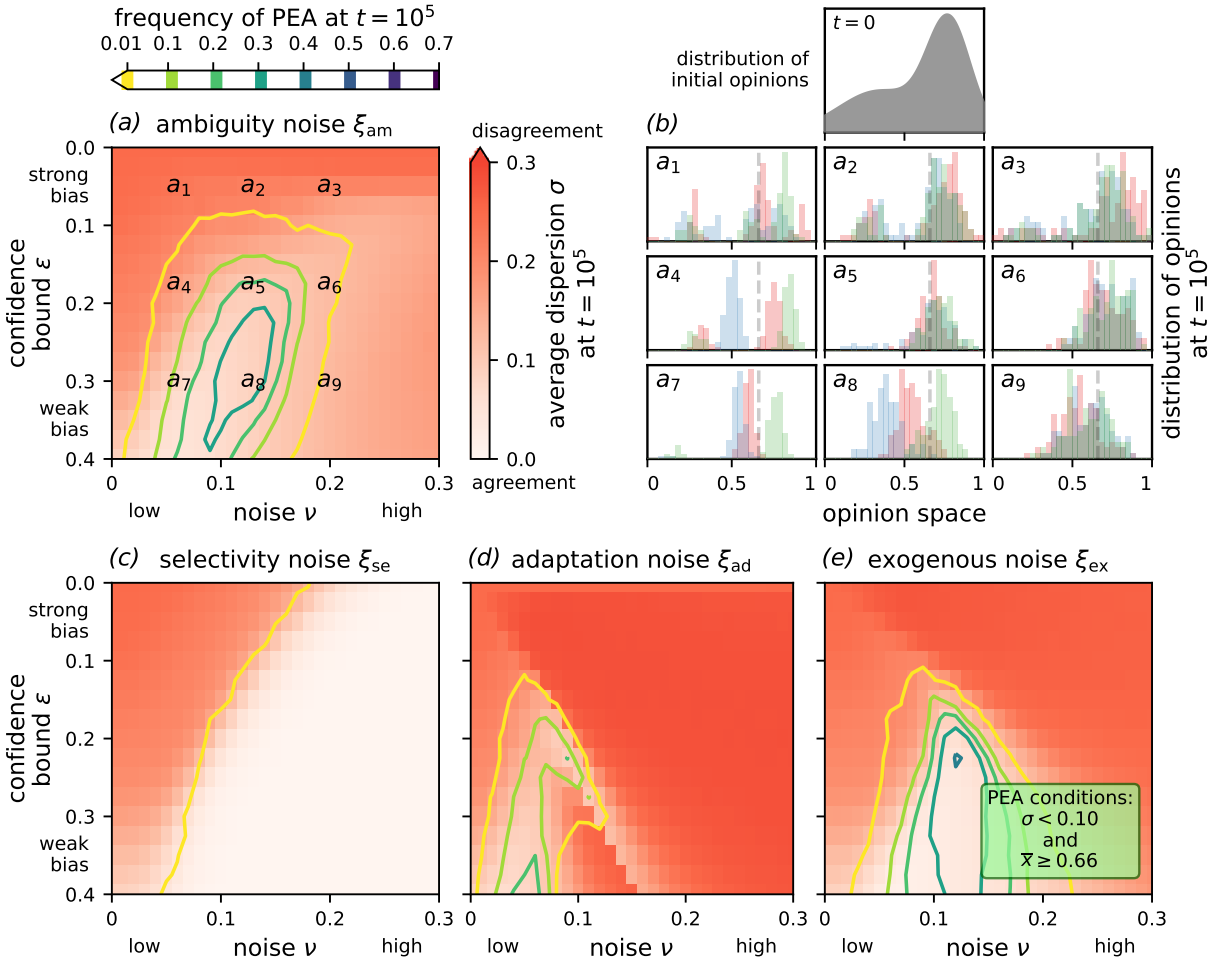


Figure A.19: Replication of Figure 3 in the main manuscript using a bounded uniform noise distribution $\xi \sim [-\nu, \nu]$, rather than a normal distribution with standard deviation ν as in Figure 3. The panels *a*, *c*, *d*, and *e* show the dispersion, σ, and the frequency of reaching a pro-environmental agreement (PEA, with $\sigma < 0.1$ and $\overline{x} \geq 0.66$) over different levels of noise ν and confidence bounds ϵ for the four types of noise. Panel *b* shows examples of opinion patterns under specific configurations of confidence bounds and ambiguity noise a_1 to a_9 (see panel *a*). While the combination of moderate ambiguity noise and moderate bias still outperforms most other types of noise in terms of reaching a PEA, the effect is less pronounced than in the case of normally distributed noise (compare panel *a* in this Figure with Figure 3*a*). Here, moderate exogenous noise, drawn from a bounded uniform distribution $[\nu, \nu]$, in combination with a moderate bias performs comparatively well in terms of reaching PEA. Interestingly, adaptation noise shows non-monotonic behaviour if the bias is relatively strong. This might be due to an interplay between the strict non-linearities of both confidence bounds and noise when assuming a bounded noise distribution.

Appendix B

Supplementary Material for Chapter 3

This chapter contains supplementary material for

Steiglechner, P., Smaldino, P. E., Moser, D., & Merico, A. (2023). Social identity bias and communication network clustering interact to shape patterns of opinion dynamics. *Journal of The Royal Society Interface*, *20*(209), 20230372. https://doi.org/10.1098/rsif.2023.0372.

It is published as Electronic Supplementary Material and available online at https://doi.org/10.6084/m9.figshare.c.6960070.

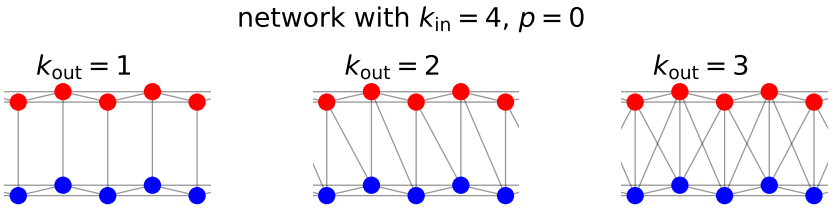


Figure B.1: The network generation algorithm for parameter values $p = 0$ (degree of randomness) and $k_{\text{in}} = 4$ (number of in-group links per agent) and $k_{\text{out}} = 1$, $k_{\text{out}} = 2$, or $k_{\text{out}} = 3$ (number of out-group links per agent). In-group networks (consisting of all red or all blue nodes) represent small-world networks following the Watts-Strogatz model. All in-group nodes are linked to their $k_{\text{in}} = 4$ nearest neighbours—two on the left and two on the right (red-red or blue-blue links). Then, each link to a neighbour on the right is rewired to any other in-group node with the probability p (here $p = 0$). The red and the blue in-group networks are then connected via out-group links. Specifically, each agent of the red group is linked to the k_{out} closest agents of the blue group. Then, each of these links is again rewired with probability p, such that a red node is replaced by another randomly selected red node or a blue node is replaced by another randomly selected blue node.

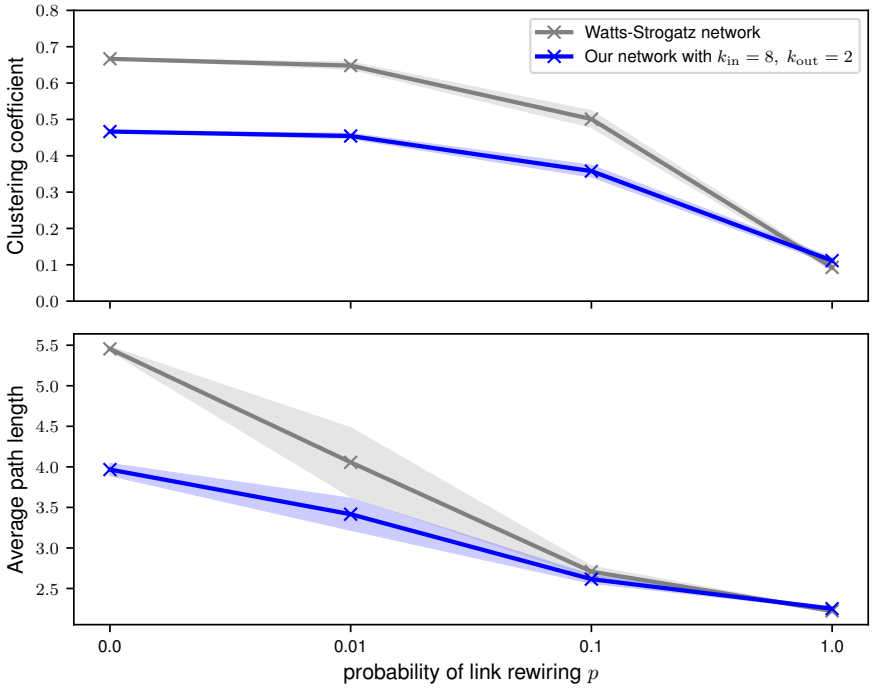


Figure B.2: The clustering coefficient (panel a) and the average path length (panel b) of the network used in the manuscript (blue line) compared to a standard Watts-Strogatz small world network Watts and Strogatz, 1998 (grey line) over the randomness p. Our network is characterised by moderate homophily, $k_{\text{in}} = 8$ and $k_{\text{out}} = 2$, while the small-world network without identities is constructed with $n = 100$ and $k = 10$. Our network exhibits similar clustering and path length behaviour as a standard small-world network, but less pronounced.

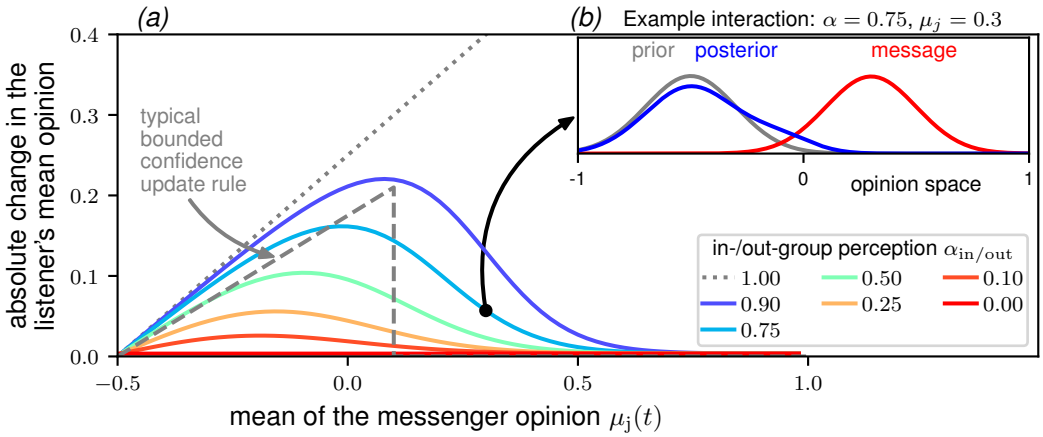


Figure B.3: Impact of a gaussian opinion distribution, $x_j \sim \mathcal{N}(\mu = \mu_j, \sigma = 0.2)$ (with μ_j on the x-axis), from a messenger j on the gaussian prior opinion, $x_i \sim \mathcal{N}(\mu = \mu_i = -0.5, \sigma = 0.2)$, of a listener i. The impact (y-axis) is measured as the change in the mean opinion of the listener i due to a single interaction with the messenger j for various values of the (in-group or out-group) perception parameter $\alpha_{\text{in/out}} \in [0, 0.99]$. To avoid division by 0 in the normalisation of the posterior opinion, we do not consider $\alpha_{\text{in/out}} = 1$ (dotted line), i.e. we neglect the case that agents see messages exactly as communicated through a fully transparent filter. In contrast to the typical bounded confidence update rule (assuming that the bounded confidence applies to the mean opinions of the communicating agents; grey dashed line), the threshold in our update rule is not sharp and it is endogenously derived from the interplay of an assimilative force and a conservative/preservative force (see manuscript). The assimilative force becomes stronger, the more discrepant the opinions of the messenger and listener are, i.e. assimilation increases with the distance $|\mu_j - \mu_i|$. This dominates the left part of the figure. However, similar to the assimilative force, the preservative force also increases with $|\mu_j - \mu_i|$ and this increase is non-linear. Preservation thus outweighs assimilation in the right part of the figure. When the messenger's opinion x_j is extremely discrepant to the listener's opinion x_i (e.g. $\mu_i = -0.5$ and $\mu_j = 1$), the a message has barely any impact on the listener. The inset panel (b) shows an example of an opinion update of a listener i—from a gaussian prior (grey) to a non-gaussian posterior (blue) opinion distribution—following the interaction with a messenger with $\mu_j = 0.3$ (red) whose message the listener perceives with $\alpha = 0.75$ (black dot in the main panel).

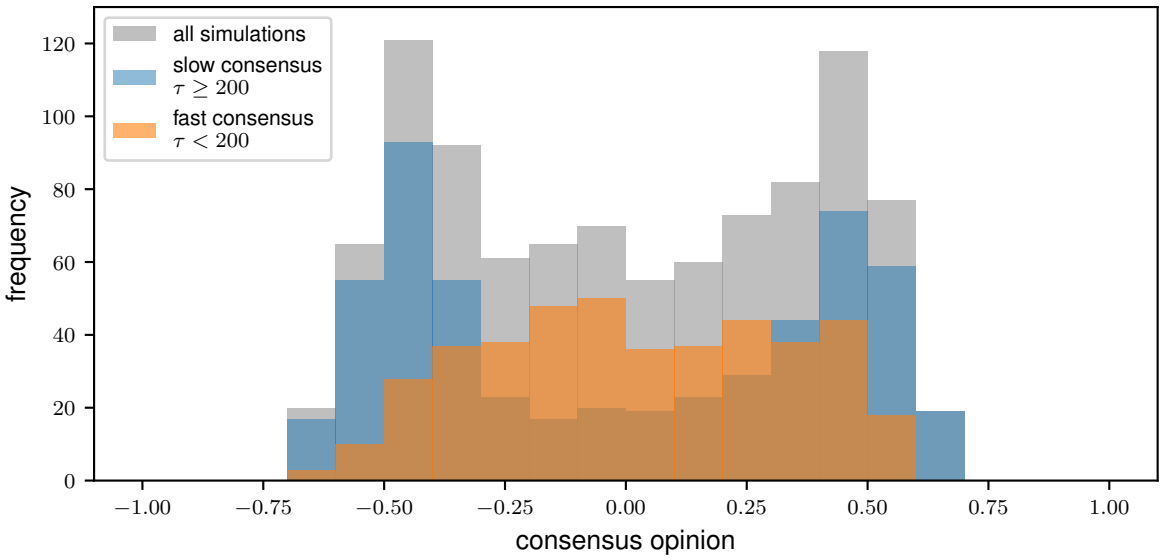


Figure B.4: The consensus opinion in simulations of a society B (with $\alpha_{\text{in}} = 0.75$ and $\alpha_{\text{out}} = 0.25$) in which consensus was reached before $t = 200$, i.e. 'fast' consensus (orange), or after $t = 200$, i.e. 'slow' consensus (blue). The simulated society is characterised by homophily ($k_{\text{in}} = 8$ and $k_{\text{out}} = 2$) and highly random network ($p = 1$). Fast consensus tends to be formed around moderate values. Late consensus emerges when the agents first converge to distinct opinion clusters and, then, one cluster absorbs the other. Thus, if consensus occurs later in the simulation, this consensus is often quite extreme (around $b = \pm 0.5$).

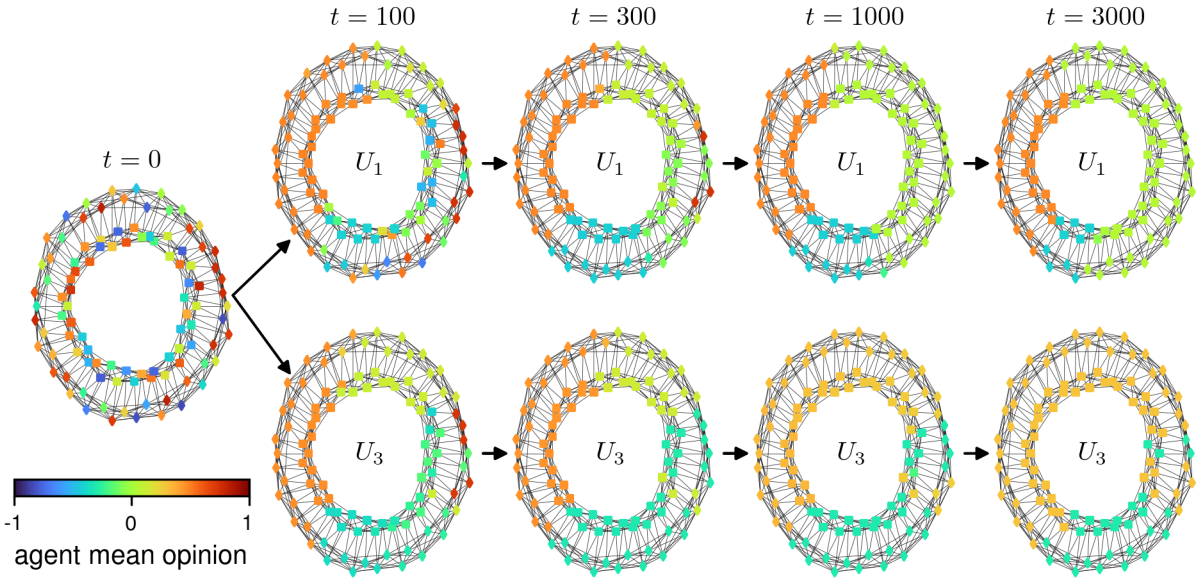


Figure B.5: Example simulation for unbiased societies U_1 (with $\alpha_{\text{in}} = \alpha_{\text{out}} = 0.25$) and U_3 (with $\alpha_{\text{in}} = \alpha_{\text{out}} = 0.75$) characterised by a homophilic ($k_{\text{in}} = 8$ and $k_{\text{out}} = 2$) and highly clustered network ($p = 0.0$). The shapes and positions of the nodes (the outer and inner circle) represent the agents' social identities, the colours of the nodes represent the agents' mean opinions at different times t. In contrast to society B (see figure 4 in the manuscript), consensus is not reached in both unbiased societies U_1 and U_3 in this example.

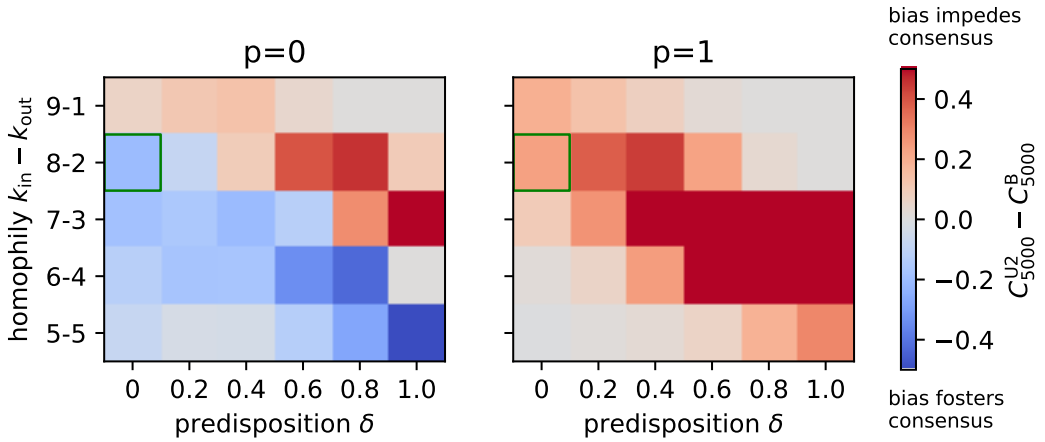


Figure B.6: Sensitivity of the effect of in-group bias under simultaneous variations of homophily (y-axis) and identity-based predisposition in the initial opinions (x-axis) on the simulated societies characterised by highly clustered ($p = 0$) and highly random networks ($p = 1$). The pixel colour denotes whether the consensus frequency C_{5000} (in 200 simulation runs) of an unbiased society U_2 is higher (red) or lower (blue) than that of a biased society B. In the red region, bias impedes consensus. In contrast, in the blue region where the consensus frequency C_{5000} of a biased society B is higher than that of an unbiased society U_2, the bias fosters consensus. The green square indicates our default parameter configuration (see figure 3 in the manuscript). Our main result—that bias impedes consensus in highly random networks (right panel) and fosters consensus in highly clustered networks (left panel)—is very robust unless homophily is extreme, predisposition is strong, or both combined.

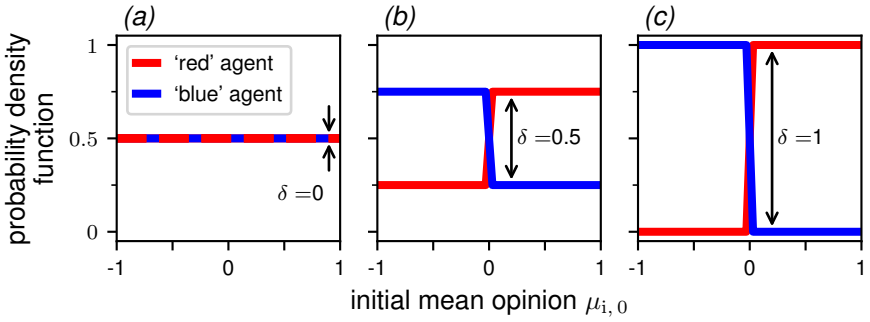


Figure B.7: The probability distribution from which we draw the initial mean opinions, $\mu_{i,0}$, of agents i with social identities s_i (blue/red). When predisposition is $\delta = 0$, the $\mu_{i,0}$ is uniformly distributed. As predisposition grows, for example with $\delta = 0.5$, agents with $s_i = blue$ are predisposed to have initial opinions centred in the right half of the opinion space at $t = 0$, i.e. $\mu_{i,0} > 0$ with a probability $0.5 + \delta/2$. For $\delta = 1$, the agents are entirely separated into opposite sides of the opinion space at $t = 0$ based on their social identities.

Appendix C

Supplementary Material for Chapter 4

This chapter contains supplementary material for Chapter 4.

C.1 Distribution of identities in the ESS data

Party	Left	Green	Social Democrat	Liberal	Conservative	Right-wing Extremist	No partisan identity	n
	Die Linke	Bündnis 90/Die Grünen	SPD	FDP	Union or CDU/CSU	AfD	None	
wave 8	5.6 %	8.2 %	14.6 %	1.9 %	18.0 %	3.6 %	48.0 %	2681
... parties	10.8 %	15.7 %	28.1 %	3.7 %	34.7 %	6.9 %	—	
wave 10	4.0 %	12.2 %	12.4 %	5.1 %	12.7 %	2.2 %	51.4 %	7807
... parties	8.2 %	25.1 %	25.5 %	10.4 %	26.1 %	4.5 %	—	

Table C.1: Shares of partisan identities in the ESS data included in our analysis for the two waves 8 (2016/17) and 10 (2021) with or without those individuals that do not feel close to any party ('None') and the overall size of the dataset included in the analysis in the column n.

Table C.1 shows the distribution of partisan identities as extracted according to the procedure described in the main article in Section 4.3. To set this into perspective, we compare it with the voting shares that each party received in the German general election in September 2021, which occurred shortly before wave 10. The 2021 voter shares are 4.9 % for 'Die Linke', 14.8 % for 'Bündnis 90/Die Grünen', 25.7 % for 'SPD', 11.5 % for 'FDP', 24.1 % for 'Union' (or 'CDU/CSU'), and 10.3 % for 'AfD' (with a voter turnout of 76.6 %). Note that the ESS wave 10 questionnaire also asks its participants which party they voted for in the last national election, i.e. in September 2021 (*prtvfde2*): 5.5 % for 'Die Linke', 20.8 % for 'Bündnis 90/Die Grünen', 27.3 % for 'SPD', 13.3 % for 'FDP', 21.4 % for 'Union', and 6.6 % for 'AfD'. These results are very similar to the question

which party the respondents felt closest to (see Table C.1). Comparing the national election results with the identities represented in the ESS sample, we conclude that the ESS sample seems to either underrepresent the share of those identifying or voting for the 'AfD' or those voting for the 'AfD' do not state the 'AfD' as a party that they feel particularly close to. This likely introduces a bias in our analysis.

C.2 Sampling weights

In the main article, we define perceived disagreement as:

$$d(\mathcal{X}_t, \mathcal{T}_t) = \overline{d_i(\{x_j | i \neq j\}), \mathcal{T}_t)} = \frac{1}{n \cdot (n-1)} \sum_{j \neq i} \delta_i \left(x_i, x_j | T_i\right) \ . \tag{C.1}$$

To account for the sampling weights of the European Social Survey (ESS), we adjusted equation (C.1) as follows:

$$d(\mathcal{X}_t, \mathcal{T}_t) = \frac{1}{\sum_i w_i} \cdot \sum_i \frac{1}{\sum_j w_j - w_i} \cdot \sum_j w_i \cdot w_j \cdot \delta_i \left(x_i, x_j | T_i\right) \ , \tag{C.2}$$

where w_i are the analysis weights for each individual as provided by the ESS dataset. Note that we did not account for these weights when inferring the bases of the subjective representation of the opinion space of each group. The weights are in fact all very close for the ESS data in wave 8 and wave 10. Thus, disregarding these weights yields nearly exactly the same results.

C.2.1 Illustrative example of a mismatch between actual and perceived opinion polarisation

To motivate our measure of perceived opinion polarisation and explain the implications of our assumptions, we present a very simple, artificial example. Consider a society of individuals separated into two identity groups (red and blue). The individuals hold one-dimensional opinions, $x(t)$, on a specific topic, such as climate change. The distributions of the red and blue identity groups differ (see Figure C.1). Note that in this one-dimensional case, the subjective bases representing individual opinion spaces can only be scaled, not rotated.

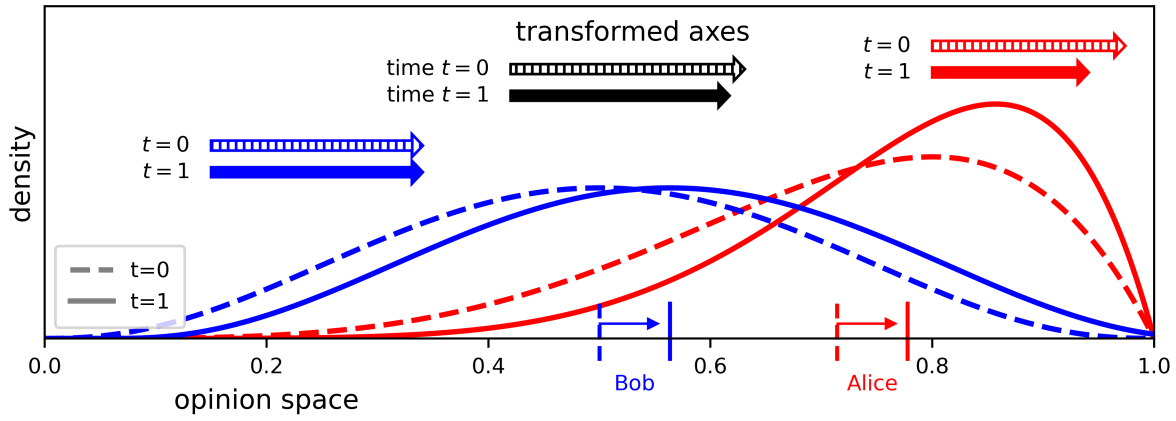


Figure C.1: Artificial one-dimensional opinion distribution at times t_0 (dashed) and t_1 (solid lines) of a red and a blue identity group and the corresponding bases defining the perceived opinion space at each point in time for each group (hatched arrows for t_0 and filled arrows for t_1). A subjective opinion distance of 'unit 1' corresponds to the length of the arrow for each group, respectively. The black arrows indicate the axes in the absence of identity or in-group bias, i.e. for $w = 0$. The opinions of the exemplary individuals, Alice (red) and Bob (blue), used in the main text are indicated as small vertical lines on the bottom axis for t_0 and t_1. We assume that from t_0 to t_1, the opinions of the red and the blue groups shift from $\overline{x}_{\text{blue}}(t_0) = 0.5$ and $\overline{x}_{\text{red}}(t_0) = 0.72$ to $\overline{x}_{\text{blue}}(t_1) = 0.56$ and $\overline{x}_{\text{red}}(t_1) = 0.78$, respectively, but the shapes of the distributions change as well. As a consequence, also the representation of the opinion space changes differently for the two groups. Specifically, the basis of the 'red opinion space' at t_1 is shorter—indicating more in-group homogeneity and, thus, larger subjective distances to other individuals—while the basis of the 'blue opinion space' remains the same.

Imagine an individual Alice with the 'red' identity. Alice is quite alarmed about climate change and her red in-group is similarly concerned. Alice's opinion is $x_A = 0.72$, which is also the mean of the opinion distribution of her in-group (red dashed line in Figure C.1). Now imagine Alice meets Bob, who has a 'blue' social identity. Bob is less worried about climate change, $x_B = 0.5$, and the topic of climate change is not central to the belief system of his 'blue' in-group, because the 'blue' group comprises a variety of climate change opinions. Bob's opinion is also the mean of the opinion distribution of his in-group (blue dashed line in Figure C.1). How different does Alice perceive Bob's opinion on climate change to hers? And does Bob see the same opinion difference? If perceptions were objective, the distance between Alice's and Bob's opinions would be $\delta^{\text{obj}}(x_A, x_B) = 0.22$. Our method implies, however, that subjective individuals with in-group bias, $w > 0$, see distances asymmetrically. Alice perceives a greater distance, for example, $\delta(x_A, x_B|T_A(t_0)) = 1.30$ for $w = 1$, than Bob, $\delta(x_A, x_B|T_B(t_0)) = 1.18$.

Now imagine that, after, for example, experiencing several heat waves linked to climate change, the opinions in both groups shift to higher degrees of concern about climate change. While the groups shift by the same average amount, 0.06, the distributions evolve in slightly different ways (solid lines in Figure C.1). With $x_A = 0.78$ and $x_B = 0.56$, the objective distance between Alice and Bob remains the same, $\delta^{\text{obj}} = 0.22$. In the absence of in-group bias, i.e. for $w = 0$ (black arrows in Figure C.1), the distance between Alice and Bob increases slightly from $\delta(x_A, x_B|T^0(t_0)) = 1.06$ to $\delta(x_A, x_B|T^0(t_1)) = 1.14$. But if social identity strongly affect their perceptions, Alice perceives Bob's opinion as much more distant to her own opinion than she did previously. For example, for $w = 1$, $\delta(x_A, x_B|T_A(t_1)) = 1.63$—an increase of 0.33 compared to t_0—while Bob perceives the same distance as before, $\delta(x_A, x_B|T_B(t_1)) = 1.18$.

Overall, assuming 1000 individuals in the red and 1000 individuals in the blue identity group with opinions drawn from the respective distributions in Figure C.1 and maximum in-group bias $w = 1$, our framework yields a positive perceived opinion polarisation $P_{\text{perc}} = d(\mathcal{X}_{t_1}, \mathcal{T}_{t_1}) - d(\mathcal{X}_{t_0}, \mathcal{T}_{t_0}) = 1.73 - 1.65 = 0.08$ and a negative actual opinion polarisation $P_{\text{actual}} = -0.12$ (i.e. opinion convergence)—a mismatch of $P\Delta = 0.2$. In other words, due to changes in how individuals in the red and blue identity group represent the opinions of others, they perceive that opinions polarise rather than converge. Moreover, there is also a significant asymmetry in how the groups perceive polarisation, with blues perceiving convergence $P_{\text{perc}}(\text{blue}) = -0.06$ and reds perceiving polarisation $P_{\text{perc}}(\text{red}) = 0.2$ for $w = 1$. This example illustrates that in the case that in-group bias shapes the perceptions of opinions, the perceived distances between opinions are asymmetric and polarisation in this example appears greatly amplified (here, even turning actual opinion convergence into perceived opinion polarisation).

C.3 Additional Analysis

C.3.1 Representations of the opinion space by partisan identity group

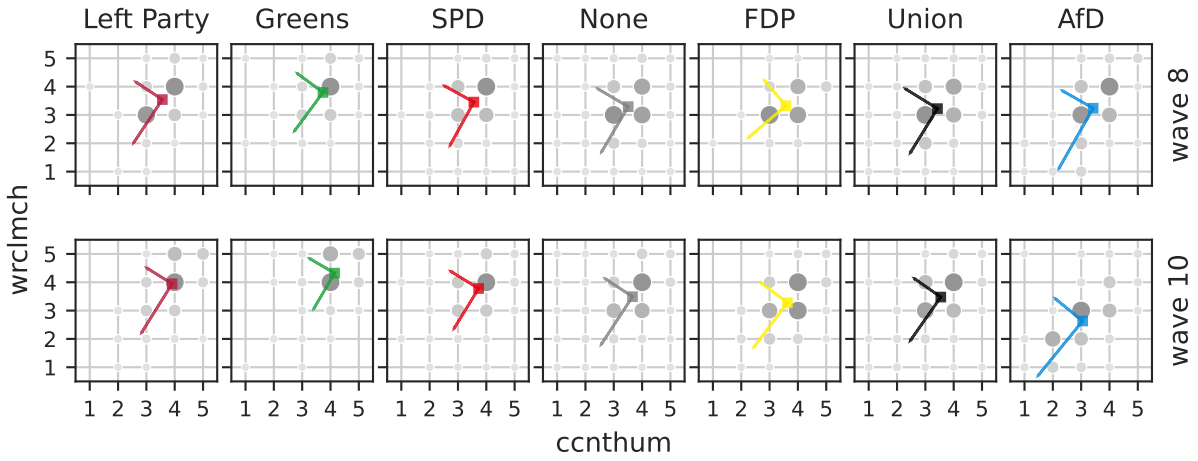


Figure C.2: The opinion distribution and the resulting subjective representation of the opinion space for each of the partisan identity groups in wave 8 and 10, assuming maximum in-group bias, $w = 1$. For details, see Figure 4.3.

C.3.2 Polarisation between partisan identity groups

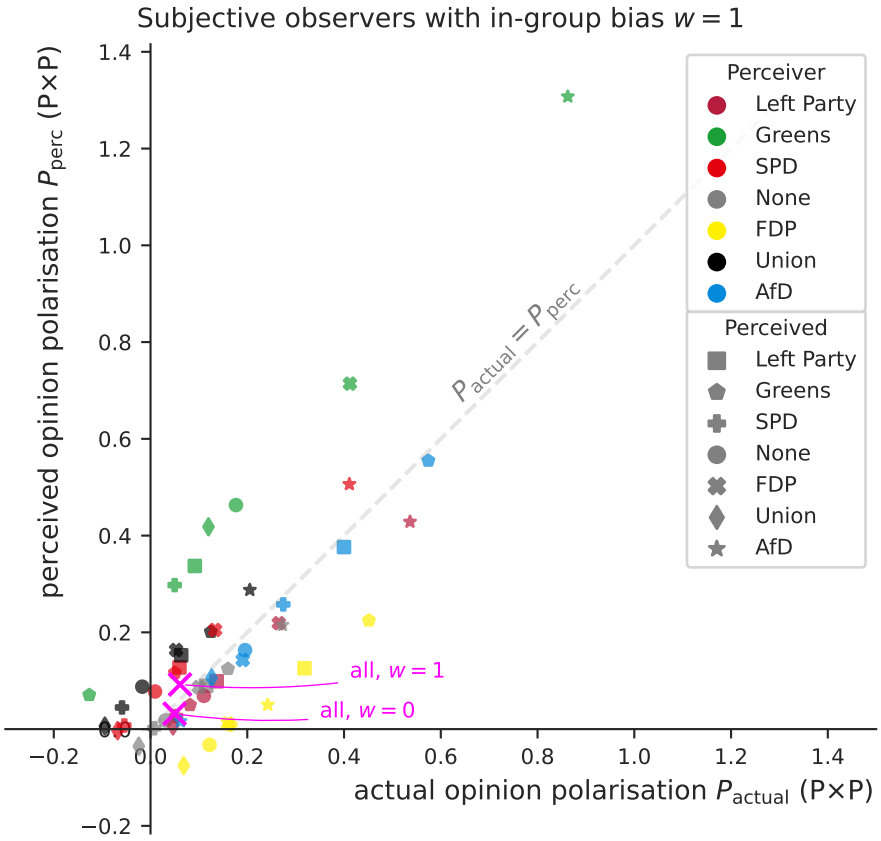


Figure C.3: Party-by-party polarisation: Each dot shows the relation between perceived and actual opinion polarisation between individuals in a perceiving group (colour) and individuals in a perceived group (shape of the dot) from the perspective of the perceiving group. For details, see Figure 4.4.

C.3.3 Robustness of the polarisation effects

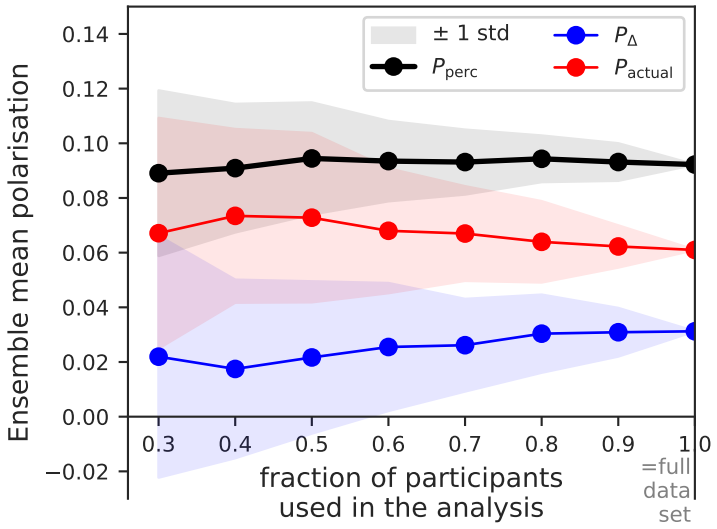


Figure C.4: Robustness test: Perceived opinion polarisation, P_{perc} (assuming $w = 1$), actual opinion polarisation P_{actual}, and P_Δ, the mismatch between the two, inferred from populations where we randomly excluded a fraction of the ESS survey respondents. In our main analysis, we consider the full dataset (fraction 1.0) with $n_8 = 2681$ (valid) participants in wave 8 and $n_{10} = 7807$ in wave 10. Here, we randomly selected a fraction of survey participants in each wave and derive the polarisation values from this reduced dataset. Performing this calculation 100 times for fixed fractions (x-axis), we report the ensemble mean polarisation values (lines) and their standard deviation (shaded areas). The obtained ensemble mean values for P_{perc}, P_Δ, and P_{actual} remain very robust regardless of the size of the used dataset. The standard deviation increases, which is expected because smaller samples cause greater variance (especially smaller identity groups, such as the Right-wing Extremists, are susceptible to great fluctuations). The stability of the polarisation indices for large enough datasets confirms the robustness of the main effect—that perceived opinion polarisation is greater than actual opinion polarisation assuming a strong in-group bias, $w = 1$.