
Electronic Thesis and Dissertation Repository

3-20-2023 3:30 PM

Using Formal Epistemology to Model Epistemic Injustice Against Neurodivergent People

Mackenzie Marcotte, *The University of Western Ontario*

Supervisor: Myrvold, Wayne C., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Philosophy

© Mackenzie Marcotte 2023

Follow this and additional works at: <https://ir.lib.uwo.ca/etd>



Part of the [Epistemology Commons](#)

Recommended Citation

Marcotte, Mackenzie, "Using Formal Epistemology to Model Epistemic Injustice Against Neurodivergent People" (2023). *Electronic Thesis and Dissertation Repository*. 9229.

<https://ir.lib.uwo.ca/etd/9229>

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.

Abstract

Neurodivergent people experience epistemic injustice, injustices that harm them in their capacity as knowers, but so far the epistemic injustice literature has mostly ignored this. This dissertation addresses this gap in knowledge in a novel way, using tools of formal epistemology. Bayesian network learning models that include modeled bias, communication style gaps, exclusion, and difference between people, are used to investigate testimonial injustice. Novel simultaneous Lewis-Skyrms signal games that include modeled bias, focus on success, gaps in way of thinking, exclusion, and difference in material interests are used to investigate hermeneutical injustice, the subset of epistemic injustice that involves concepts important to an identity group being obscured both in and out of that identity group, due to the model's ability to track formation of meaning over time. The model results indicate that improvement first requires neurodivergent people be integrated into social networks with mixed neurotypes, but that this must be done with care to not isolate neurodivergent people among neurotypical people, and without tokenizing. Additionally, the models give evidence of social evolutionary forces that would contribute toward the presence of ableism in norms of communication, so it is recommended that action to combat ableism should include actions that create countervailing cultural evolutionary pressure, and aim to benefit anyone whom the action hopes to win over.

Keywords: Epistemic Injustice, Formal Epistemology, Bayesian Network Learning Models, Lewis-Skyrms Signal Models, Disability Theory, Autism, ADHD, Neurodivergence, Computational Philosophy

Summary for Lay Audience

There is a specific kind of injustice that changes how good someone is at holding knowledge, and, in an unfair way, how good others consider them to be at holding knowledge. It is called “epistemic injustice.” Previous discussion of this kind of injustice has paid attention to how someone’s identity, such as race or gender, is at the root of the injustice, and how different identity groups experience it differently. However, researchers have mostly not paid enough attention to how people identified as “neurodivergent” have experienced it, such as autistic people and people with ADHD.

I aim to investigate the way this injustice works for neurodivergent people by creating computer simulations of unjust situations based on their experiences. There are two main kinds of simulation I look at. The first creates networks of individuals working on their own and communicating with each other. I use the networks to investigate epistemic injustice by changing how the individuals in the network communicate with each other, and observe the results when they become biased or have trouble communicating. The second kind of simulation has individuals sending signals to each other that initially do not have a meaning, but can gain meaning as the individuals learn to associate them with events. Again, I change how the individuals are able to signal each other and observe the results.

I then argue for some conclusions based on these simulations. For example, that we need to bring neurodivergent people into the mainstream, but in a way that helps them connect to each other too. If they are not connected to each other, there is a risk of not actually helping them because they are seen as tokens of progress instead of people to connect to. As well, I conclude that we should aim to make it easier to support neurodivergent people than not, either by making things easier for everyone, or getting in the way of people who want to harm them.

For uncle Andy. For all neurodivergent people who deserved more time.

Acknowledgements

This work suggests that neurodivergent people access and share knowledge much more effectively when they have meaningful connections with a large number of neurodivergent and neurotypical people. For this and the usual reasons, this project would not have been possible without my massive support network of varying neurotypes, including Rush Pill, Emily Chichocki, Amy Keating, Benjamin Formanek, Heather Stewart, Madeleine Marcotte, James Marcotte, Jean-Paul Marcotte, Sarah Gale, May Hoffos, Felix Cabezas, and a very neurodiverse discord server. As well, the advice and encouragement of my advisor Wayne C. Myrvold has been indispensable.

I owe a great debt to the developers of Python, and its Numpy, Pandas, Matplotlib, Pylab, Multiprocesing, Networkx, and Mesa libraries.

I am especially grateful to my mother Clare Scott and all others who have created the material foundation for neurodivergent people to create work like this, and who continue to work towards an even better world.

Table of Contents

Abstract and Keywords	i
Summary for Lay Audience	ii
Dedication	iii
Acknowledgements	iv
List of Figures	viii
List of Tables	ix
Introduction	1
1 Injustice in Epistemic Networks	4
1.1 Model Motivation: Epistemic Injustice	6
1.1.1 Fricker and Epistemic Injustice	6
1.1.2 Other Comments on Epistemic Injustice	8
1.1.3 How to Model Epistemic Injustice	11
1.1.4 Examples of Epistemic Injustice to Model	15
1.2 Network Learning Models	20
1.2.1 Bayesian Epistemic Network Models	21
1.2.2 Why These Models	27
1.2.3 Using Computer Models	33

2	Network Learning: Modeling Testimonial Injustice	37
2.1	Methodology and Results	39
2.1.1	The Code	39
2.1.2	Model Types	42
2.1.3	Results	48
2.2	Appraising Network Models of Epistemic Injustice	65
2.2.1	Discussing the Results	65
2.2.2	Shortcomings of Network Models	74
3	Injustice in Norms of Communication	79
3.1	Hermeneutical Injustice	80
3.1.1	Fricker and Hermeneutical Injustice	80
3.1.2	Examples of Hermeneutical Injustice to Model	84
3.2	Signal Games	86
3.2.1	A Brief Introduction to Game Theory	87
3.2.2	Lewis Signal Games	92
3.2.3	Lewis-Skyrms Signal Models	94
3.2.4	Simultaneous Signal Models	96
3.2.5	Hermeneutical Injustice in Signal Models	99
3.3	Equilibrium Analysis	103
3.3.1	Base Model Equilibrium Analysis	103
3.3.2	Altered Model Equilibrium Analyses	113
4	Signal Models of Hermeneutical Injustice	120
4.1	Methodology and Results	121
4.1.1	The Code	121
4.1.2	Results: Base Model	125
4.1.3	Results: Other Model Types	133

4.2	Appraising Signal Models	142
4.2.1	Discussion of Results	142
4.2.2	Shortcomings	149
	Conclusion	152
	Bibliography	156
	A Code and Data	162
	Curriculum Vitae	163

List of Figures

1.1	Example network graph	26
2.1	Network 1 starting arrangement	49
2.2	Network 1 with difference model type	51
2.3	Network 9 at start	52
2.4	Network 9 language model side-by-side	54
2.5	Network 1 l2 model	58
2.6	Network 1 with all factors	62
2.7	Network 6 divergent model	63
2.8	Network 6 mingling divergent model	65
3.1	Lewis signal game diagram	93
3.2	Lewis-Skyrms signal model diagram	94
3.3	Example simultaneous Lewis-Skyrms signal model step	100
4.1	Example alternating 4-agent signal model	138

List of Tables

2.1	Simple network model statistics	50
2.2	Language and mingling network statistics	57
2.3	Mingling model statistics	60
3.1	First outfit game payout matrix	88
3.2	Second outfit game payout matrix	88
3.3	Third outfit game payout matrix	90
3.4	Fourth outfit game payout matrix	90
3.5	Abstract game demonstrating that giving another player more points is not a bad thing.	90
3.6	Abstract game demonstrating that kindness is accounted for in points.	90
3.7	The prisoner’s dilemma payout matrix	91
3.8	Rock, paper, scissors payout matrix	92
3.9	Lewis signal game payout matrix	93
3.10	Example Lewis-Skyrms signal model step	95
4.1	Signal model parameter statistics	127
4.2	Continuation of table 4.1	128
4.3	Urn choice in 3-agent systems with and without bias	133
4.4	Statistics from select 4-agent signal models	134
4.5	Statistics from select 6-agent signal models.	135
4.6	Statistics from select 10-agent signal models	135
4.7	A simple coordination game	144

Introduction

As neurodivergence—the concept that people with neurological conditions like autism and ADHD should be treated as a marginalized identity group and not a series of unrelated deficient types of people—has become a more mainstream topic, there has been a spike in adult autism diagnoses [1]. Women are especially being diagnosed late; depending on the study, between 2-3 young boys are diagnosed for every girl, a gap that closes to 1.2 adult men diagnosed for every woman. Even when seeking it out, women have special barriers to being diagnosed [1] [2]. Additionally, black populations see less frequent and later diagnoses, with black women especially underrepresented [3]. Many autistic people¹ are surprisingly slow to learn the very fundamental fact about themselves that they are autistic, more so at intersections of other marginalized identities. An epistemic gap centered around an identity group is becoming much clearer.

The recent wave of adults diagnosed with autism have been referred to as a “lost generation,” and not without cause [1]. Growing up neurodivergent without knowing it can be incredibly disorienting. As an example, one common symptom of both ADHD and autism is executive dysfunction, a disconnect between desire and action. Someone with executive dysfunction can find themselves laying in bed for extended periods of time, focused on wanting to get up and begin work, but not getting up anyway [5]. Without other words to describe it, it is very easy for someone experiencing it to imagine that this experience is

¹In this dissertation I will intentionally not be using person-first language, e.g. “people with autism.” Person-first language is mostly used by non-disabled people to describe disabled people, and its minimizing intentions are in tension with both mainstream disability theory and the wishes of many neurodivergent people [4].

what other people mean when they talk about being “lazy,” or lacking discipline, especially if one has been called “lazy” by others. “Laziness” is supposed to describe intentional eschewing of effort, which executive dysfunction distinctly is not, but the value-neutral experience of executive dysfunction can nonetheless lead someone to believe they irrevocably lack virtue, and as a result never learn skills to deal with the root cause of the issue. This kind of lacuna of self-understanding is starkly reminiscent of the stories in Miranda Fricker’s landmark book *Epistemic Injustice*, in which people struggle to name their experiences due to not understanding the concepts that describe them, and wind up, as Fricker describes it, incoherent to themselves [6].

For this to be the result of epistemic injustice—systemic injustices that harm people in a way that impacts them as holders of knowledge, done on the basis of identity—it has to be the case that it is not mere coincidence that autistic people are not realizing they are autistic until late in life [6]. It would have to be that autism *could* be a topic that is better understood by autistic people, if they themselves were treated properly as knowers, or that autistic people’s experiences could be made more coherent to people who could lead them towards that understanding if those people had no hidden bias against autistic people. I will not carry out a sustained argument that this is the case, but there is ample evidence. The intersectionality of late diagnosis along identity lines is a convincing clue; if there were no identity injustice, the problem should not be worse for women and black people, and especially worse for black women in particular. Below, I will describe at length several mechanisms that one should expect to create epistemic injustice all else being equal, which have been acting unimpeded. I hold it to be abundantly clear that neurodivergent people are treated unjustly as knowers, and whether or not this is connected to the rise in adult diagnoses—or to frame it another way, the lack of diagnoses prior to now—it is something I would like to understand better and fight against.

My goal here, then, will be to articulate in as much detail as possible some of the mechanisms by which neurodivergent people are unjustly disadvantaged as knowers, to the end

of combating it. In particular, my approach will use the methodology of formal epistemology, which includes mathematical and computational tools. This follows a tradition championed by Cailin O'Connor and James Owen Weatherall to use formal tools to guide philosophy of social issues, including successful and influential projects like their book *The Misinformation Age* [7]. In particular, I will be modeling the impact of a few possible causes of this injustice, first on ability to learn in a social network using a network of probabilist learner agents (also called a Bayesian network learning model), and second on formation of communication norms in a less structured social environment using a type of formal game called a signal game (specifically, a Lewis-Skyrms signal game).

Ultimately, my models will show a few things in an abstract environment. First, that being isolated is bad for divergent learning agents, but being misunderstood as the only divergent agent in a community is worse, causing them to conform to the majority to their detriment. Second, that cultural-evolutionary pressures towards norms that disadvantage divergent agents can emerge from a variety of factors including mere difference in communication style. Absent any reason to think that these effects would be mitigated in the real world, this establishes minimum baselines for effects that push our norms towards ableism, that are therefore worth resisting. Strategies for anti-ableist action informed by these results should aim to integrate as many neurodivergent people into the mainstream as possible as a first step, and then create countervailing pressures to effect systemic change instead of trying to win over the hearts and minds of specific individuals. The specific strategies we use to combat ableism may be improved by study of my models themselves, which provide additional information about how these mechanisms function, at least in abstract, too detailed to include in an introduction.

Chapter 1

Injustice in Epistemic Networks

In this chapter I will motivate my use of Bayesian epistemic network models to explore the effects of certain types of social conditions on group learning in a formalized setting, informed by and engaging with the literature on epistemic injustice. In particular, I will draw on disability theory and my own experiences to describe some real life examples around disability and epistemic injustice, and defend the use of formal representations of elements of these examples in the models I will explore in the following chapter, to a specific limited extent.

Network models of this kind are already popular in the literature. I will specifically be using a variation on the model initially developed by Venkatesh Bala and Sanjeev Goyal in their landmark 1998 paper “Learning from Neighbours” [8], and recently advanced by Cailin O’Connor and James Weatherall in a number of places, most notably their 2018 paper “Scientific Polarization” [9]. The strength of these models is that they give a powerful and straightforward analogy to actual communities of individuals earnestly trying to learn the best course of action between alternatives, and in some cases demonstrate that it takes very little to upset the end resultant beliefs for such a community. If even these small perturbations from best-case-scenarios can disrupt good learning, we have a high minimum level of disruption that is likely to occur in real-world examples where human beings may

not be fully and exclusively interested in finding truth, may not be doing their own research, and may not express the same level of skill in learning the right lessons from evidence in the right degrees that idealized bayesian agents do. As well, we can look with laser focus on where these breakdowns occur, and extrapolate that similar breakdowns may occur in the real world, and try to combat the problems they raise. There are clear weaknesses as well, which I will touch on in detail later, but can mostly be summed up by noting that the models make so many unrealistic assumptions that their relevance to real world situations is limited to the very conservative claims above, if those are not too strong already.

The benefit of this kind of model is that the information they give is systemic in nature. In the epistemic injustice literature, it has been noted that too much of what has been suggested is at the individual level—especially Miranda Fricker’s individual virtue ethics—despite epistemic injustice being an inherently systemic issue [10]. There is a place for work that is in a better position to give systemic descriptions and recommendations around this topic, and network models can fit in that place nicely. There have also been calls for more discussion around epistemic injustice and disability [11], which my work will help meet by focusing on epistemic injustice towards cognitively disabled individuals. The specific models will look at a running example of a community trying something new to see how it works for each member, with the example of coffee when specificity is needed, and I will draw conclusions from this example extrapolating to other situations with similar dynamics¹.

To explain and introduce this new work, I will first go over the epistemic injustice literature and my examples. Then, I will justify my use of network learning models and give a brief explanation of work done so far in this area. Next, I will explain one key difference between my analysis and most work done so far, namely why I am basing my conclusions partially on statistical analysis of results at various stages of completion of partially randomly generated models, rather than only on direct analysis of particular mathematical

¹The effect of coffee is not particularly important, but it provides a clean example with *some* importance that easily maps onto other drugs and some situations that do not involve substances.

facts about particular models, tailored to the problem at hand and analyzed holistically. My methodology is closer to random sampling done in social science, and I argue gives a finer view into how the models function in some ways, without losing any fidelity thanks to the possibility of joining the two methods. This will give the background needed to describe the model in detail and explore its results in chapter 2.

1.1 Model Motivation: Epistemic Injustice

1.1.1 Fricker and Epistemic Injustice

The primary motivation for this project is the book *Epistemic Injustice* by Miranda Fricker [6]. In this book, Fricker describes a specific kind of injustice that harms someone in their capacity as a knower, termed *epistemic injustice*. This concept is to be differentiated from injustices that simply have an epistemic bent to them, which might include unfair distribution of epistemic resources like attention and trust, insofar as epistemic resources can be justly distributed, which would therefore be unjust for the same reasons any other negative outcome could be an injustice. It is also to be differentiated from injustices which harm people in other ways, such as by depriving them of resources, as the mechanics and harms of epistemic injustice are different from other injustices.

Fricker specifically outlines two kinds of epistemic injustice. The first, and focus of the bulk of her book, is *testimonial injustice*. Testimonial injustice occurs when there is a deficit of trust put in the testimony of someone, due to systemic identity-based prejudice. Fricker defends the conditions that it must be structural, identity-based, and prejudicial, on the grounds that to be a kind of injustice rather than just a generic bad thing each condition must be met. The central example comes from the novel *To Kill a Mockingbird*, in which the legal testimony of the character Tom Robinson is not believed by an entirely white jury because he is a black man. This case is prejudicial because the jury is judging Robinson before hearing his testimony, identity-based because it is his race that causes this prejudice,

and structural because it is not due to a coincidence that enough people who prejudge black men got into a jury, but rather due to the trial existing within an entire system of racism. The majority of the book is then devoted to giving a virtue epistemological account of testimonial injustice, the epistemic virtue that could help fight it, and finally the character of the harm of testimonial injustice.

The second kind of epistemic injustice, which will be my focus in later chapters, is *hermeneutical injustice*. Fricker gives a specific definition of this on page 155, in which she says hermeneutical injustice is

the injustice of having some significant area of one's social experience obscured from collective understanding owing to a structural identity prejudice in the collective hermeneutical resource [6].

Again, the conditions of being structural, identity-based, and prejudicial are present in order to differentiate mundane hermeneutical misfortune from actual hermeneutical injustice. Here, "collective understanding" can either be taken to be the larger cultural understanding and lexicon, or less broadly situations where understanding of a concept is allowed to spread between individuals at all. One major example given here is workplace sexual harassment. Fricker quotes a story from Susan Brownmiller's memoir, in which she explains that before workplace sexual harassment was given a name, incidents were often understood to be isolated and personal, and because they were considered embarrassing, not often discussed. However, in women's groups, a space was created to discuss these incidents, and the term was created to describe a structural and common issue, and proclaim it as such. Once this term existed and was popularized through speak-outs, women were much more able to protect themselves against it, and it became possible for legislature to attack it. For Fricker, because women were marginalized in workplaces and otherwise², it was made more difficult for them to create this collective understanding, and therefore this manifest lack of a term, rather than an unfortunate lack of understanding that contributed

²"Were" here because it is relevant that it was true at this time, not because it is no longer true.

to a separate injustice, was an injustice in itself, until rectified by feminist movements.

While Fricker does not claim these two are exhaustive, they are both thought-provoking concepts. Hermeneutical injustice is especially thought-provoking because it shows that the mere way society is arranged can be unjust to us in a uniquely epistemic way, without any specific bad actors—or individuals acting out of ignorance for that matter—who are to blame. Anyone who follows feminist social epistemology will be intrigued by this idea, and find it immediately plausible. My purpose here will be to investigate certain ways we can investigate structural epistemic injustice, focusing first on testimonial injustice, and in later chapters with a focus on hermeneutical injustice. To do this, I will finish motivating my methodology by sketching how a specific kind of model might help with a problem of this kind, then describing situations that can plausibly be modeled in this way. Once motivated, I will construct and run these models, and give a discussion about what kinds of conclusions can be drawn from their results, including a discussion of how formal epistemic models are often used in the literature.

1.1.2 Other Comments on Epistemic Injustice

As I wish to draw on the epistemic injustice literature in general, a small amount of additional background is needed. I will focus on points that will be relevant, rather than giving a full review of the epistemic injustice literature.

In ‘Hermeneutical Injustice and Polyphonic Contextualism: Social Silences and Shared Hermeneutical Responsibilities,’ José Medina argues that epistemic injustices can only be discovered when looking at temporally and socially extended contexts, rather than the individual interactions Fricker focuses on. I agree with this position, which further motivates the use of networks in models of epistemic injustice, as network models are better able to look at overall social structures rather than individual interactions. Connected to this is what Medina calls “communicative pluralism,” the claim that examples of what is called epistemic injustice go beyond the epistemic in their execution, and further cover communi-

cation more broadly. When lacunae of conceptual understanding are created via systemic identity prejudice, they therefore are not a result of practices only around knowledge, they involve larger patterns of how communication is organized and carried out. For example, Medina notes that people can be *preemptively silenced* by being excluded from all conversations on a topic, which could be considered outside of the realm of epistemic considerations [12]. This will be more of a motivation in chapters 3 and 4, but will also end up squaring with my analysis in the end of chapter 2.

Elizabeth Anderson, in “Epistemic Justice as a Virtue of Social Institutions,” argues that combating epistemic injustice must be done at the structural level, against Fricker’s individualistic recommendations. She connects Fricker’s virtue ethics, which struggles to tackle ethical problems beyond individual virtues and vices, to the lack of systemic responses Fricker gives. Developing this theme, Anderson says—unsurprisingly given the article’s title—that if epistemic justice is a virtue, it is a virtue of systems, not people, and that therefore we should work to rectify our epistemic systems. As a suggestion for how structural change can combat epistemic injustice, Anderson offers radical desegregation; if many of the causes of epistemic injustice is not involving marginalized people in certain conversations or people being isolated from the viewpoints of marginalized people, including other marginalized people, then increasing communication between and among different groups of people can mitigate those causes. This kind of change does not require any individual actors to change behaviour, but by making it much easier to practice the virtue of epistemic justice—or harder to practice the vice of epistemic injustice—more good can be done than by simply suggesting how people can be more virtuous [10]. Once again I agree. Anderson both motivates the use of network models and squares with latter analysis.

It has been noted by several authors that disability is a common basis for identity-based prejudice in epistemic contexts, more so than has been acknowledged by the core epistemic injustice literature. In “Feminism and Disability,” Joel Michael Reynolds and Anita Silvers

note, among a broader discussion of disability in feminist philosophy, that the concept of epistemic injustice maps well onto previous work in feminist disability theory, which describes cognitively disabled people as disadvantaged as knowers, not due to their disability, but due to social conditions³ [14]. Shelley Tremain goes further, charging both Fricker and Medina with missing the disability axis in the central example of Tom Robinson in *To Kill a Mockingbird*, in which the same disability that should have proven his innocence became a basis to ignore his testimony [11]. Overall, it is clear that epistemic injustice is highly applicable to issues around disability, as will be explored below.

Finally, a mechanical explanation of epistemic injustice like mine is going to naturally run up against Fricker's criterion of being rooted in prejudice. She contrasts epistemic injustice with what she terms "epistemic bad luck," situations in which good epistemic practices disadvantage people. She uses the example of poor eye contact, saying that it is generally good epistemic practice to associate poor eye contact with shiftiness or unreliability, and someone who has poor eye contact despite being trustworthy is not experiencing genuine prejudice when not trusted, but rather experiencing epistemic bad luck. She also uses an example of someone's medical testimony being discredited due to contingent historical lack of understanding of that disease [6]. The danger, then, is that by locating harms as being caused by automatic processes rather than intentional wrongdoing, my account may naturalize actual injustices, and place them out of the bounds of Fricker's definition of epistemic injustice. I take this concern seriously as it evokes a process within the mechanics of epistemic injustice that Sally Haslanger has outlined, which she refers to as "epistemic objectification." Epistemic objectification occurs when present epistemic deficits are taken

³Incidentally, it is a common refrain that epistemic injustice maps well onto ideas that were previously popular in specific niches. For example, both Gail Pohlhaus Jr. and José Medina have drawn out that what early black feminists called "epistemic violence" and "practices of silencing" leave out almost nothing that Fricker discussed more recently [13] [12]. It has been argued that she herself perpetuates epistemic injustice by carrying out this work without reference to these authors. At any rate, writing after the publishing of *Epistemic Injustice*, I take the term as at least a useful unifying concept that is drawing attention to important issues that have not had enough uptake outside of the subdisciplines in which they have been discussed, and Fricker's analysis to be powerful and distinct enough to merit significant discussion, but do not deny that there is merit to accounts that have problematized Fricker's role overall.

to be natural, rather than the result of past injustice, in order to excuse further injustices as resulting from a natural state of affairs [15].

I do not think my account will advance epistemic objectification for two reasons. First, epistemic bad luck is a problematic concept in the first place. Kristie Dotson has challenged epistemic bad luck on the ground that the progression of history is not purely accidental. To meet Fricker's medical example, social values drive social and scientific progress, and so it often is a matter of structural injustice that certain medical conditions end up poorly understood [16]. I would add that eye contact being considered a good indicator of character is also rooted in existing ableism, to address her other example. The range of identity-based conceptual lacunae that fall under epistemic injustice should therefore be taken to be expanded to cover many examples that Fricker would call "epistemic bad luck." An account that describes cultural evolutionary forces creating imbalances may still be taken as describing those forces creating unjust outcomes. Second, my results in later chapters will not support the thesis that innocent actions cause a significant portion of the described epistemic harm, as the most relevant causes of epistemic injustice I will model are generally normatively charged, with mere-difference models acting as a contrast.

1.1.3 How to Model Epistemic Injustice

Fricker is a feminist analytic philosopher, and her methodology reflects that. She draws on lived experiences, rather than theorizing from first principles. Then, she outlines a framework, in her case virtue epistemology, and uses it to dissect those lived experiences in order to find the components of harm, so that those components can be classified and then brought back together to form a whole picture. This is evidently a good methodology, but my hope is that it can be improved upon.

As Anderson notes, Fricker's recommendations for dealing with both kinds of epistemic injustice are each for the reader to espouse a specific epistemic virtue in contrast to the vice that begets the described injustices [10]. Fricker seems to hope that readers will do so, and

that each will act as, as she says, drops in the ocean of structural identity-based prejudice, and have a non-zero effect [6]. It can be hoped, however, that philosophy can have an effect beyond making the individuals who read it—most of whom are fellow academics—better people. However, describing a structural issue and prescribing an individual response is not likely to achieve this. Work beyond the foundation Fricker has lain should strive to ultimately prescribe systemic responses. If Fricker's virtue epistemology is the core issue, I could take on a different epistemological framework, but I would like to draw from additional sources. I would suggest that the social sciences employ methodologies well-suited to study how philosophers' concepts affect real people, what kind of changes in an environment lead to aggravation or mitigation of effects we have described, to what degree, and, most importantly, what can be done on a systemic level.

Unfortunately, I share Fricker's limitation in that I am an analytic philosopher, not a social scientist. One step I can take jumping off from her analysis is to consider what a social scientist *might* do, in ideal circumstance, and see if I can do it anyway. Now, the goal is to improve understanding of the structure of epistemic injustice, in order to find what kinds of structural changes can mitigate it. A study to this effect might start by seeking out proposed structural causes of epistemic injustice, and studying situations in which those structural causes differ from each other. For example, a cause of testimonial injustice may be racial bias, lack of understanding between race groups, or incentive among one racial group to willfully neglect to engage ideas from a more marginalized racial group where possible. So, a study could look at several population groups, one without these effects, and one each for each possible combination of them, for eight total population groups. Better, it could separate effect levels of each effect, so that rather than just looking at racial bias, it looks at weak, moderate, and strong racial bias, for a total of sixty four population groups (e.g. one group would have weak racial bias, moderate lack of understanding, and no incentives, while another would have strong racial bias, lack of understanding, and incentives). Now, each of these population groups should be numerous enough to yield

statistically significant results, the researchers should have clear ways of measuring each effect size and their epistemic results, and it should be carried out over a long period of time to allow for the dynamics, and not merely the results, of these effects to come out.

I will give the social scientists a moment to carry out this study.

Now, the study described above is patently impractical. It would be tremendously expensive, we would not have results for some time, and it is likely that the described populations do not exactly exist, or if they do it is unlikely we would get sizeable chunks of them for long enough without some intercommunication or change in effect sizes (one hopes). Nonetheless, we do not have to give up our hopes of using an experimental methodology to give informed structural suggestions. My hope is that one can get *something* of epistemic value from building a model that in some sense aims to predict how systems of people are likely to behave, the most that can be done as a philosopher. That *something* will not be the exact same knowledge we could have gleaned from a successful long-term human study, but it could play a useful role. Similar to what Cailin O'Connor said of her own contributions in *The Origins of Unfairness*, such a model could contribute things like bare minimum effect sizes. If people are acting in a generally predictable, self-interested, rational, and non-malicious way, we can know from a formal model that a certain effect would likely still come about at some minimum level. I would argue that another role formal models could play would be to give alternative setups, and measure the difference in result between them, in order to guide our expectations about which kinds of setups in the real world will give which sorts of results [17].

As a simple example, if I introduce a mathematical analogue of implicit identity bias in one model and not another, and, staying very abstract at this stage, one model in some sense goes better for a marginalized identity group than the other model, we can make predictions about whether identity bias is more likely to improve or degrade conditions for a marginalized identity group⁴. Further, close subjective analysis of the character of the

⁴It will degrade conditions. You should not need math or a footnote to tell you this. Nonetheless, some effects will be less obvious, or mechanically complex, and then models may help.

effects caused by various factors like bias could yield more interesting recommendations than “bias is bad, let us do less bias.” In the proceeding, for example, I will analyse *why* in formal models a split of 80% of agents using one language and 20% using another has a harsher measured effect on truth-seeking than a split down the middle, and discover that the real effect in either case is caused by isolating individual agents from being able to share their results well. This has wildly different implications than are merely suggested by the numbers—on first blush the recommendation seems to be that the existence of minority language groups at all is more harmful than mere language splitting, but in reality it is isolation in general that is the culprit. So, in my analysis, I will where necessary limit myself to the type of analysis O’Connor concerns herself with, pointing out minimum effect sizes in whatever direction caused by various factors, but I will also inform as much of my discussion as possible with this detailed analysis of what the effects look like and how they behave in a clean and isolated environment, in hopes of discovering higher-fidelity minimum effects and hidden factors.

The general outline of my models will hopefully be similar enough to a social-scientific study to gain *some* of the benefits one could give an analysis, with certain advantages beyond ease of running due to being computer models. The models will have mechanical *agents*, standing in for people, who have numbers attached to them representing degrees of belief in certain hypotheses. By no means do I think real people walk around with numbers in their heads describing how much they believe certain hypotheses, whether or not they are aware of it; this is an abstraction that will be immensely helpful in describing abstract change in opinion over groups of people over time, and I expect it to work relatively well at that systemic population level discussed above. These agents will be trying to improve their knowledge by collecting evidence, and will share evidence with particular other agents to whom they are connected in a structured way. Different models will structure these relationships differently, and through mostly changes in this last detail they will explore different structural situations in order to compare and contrast communities’ and individual

agents' abilities to determine which of two hypotheses most accurately describes the actual environment. The amount that a model fitting this description abstracts from the real world is hard to overstate, and I will have to take repeated care to avoid making claims they do not entitle me to make. On the other hand, being able to actually see what a formal agent "believes" at each stage, and look at the math that resulted in this, will allow much deeper analysis of the non-real situations than would be possible for real situations.

To quickly recap my methodology, I am doing conceptual analysis like any other analytic philosopher, with the addition of computer models in order to have more to base my analysis on. The analysis I will give has all the same shortcomings of Fricker's philosophical analysis, with the sole advantage that while virtue epistemology is equipped to give only individualistic suggestions, formal epistemology is well-equipped to give structural ones.

1.1.4 Examples of Epistemic Injustice to Model

Before getting specific, I want to establish a learning problem to add a few complications to. As mentioned above, I will be using a standard one in the formal epistemology literature, and here I will attempt to justify the use of that problem. Suppose a community of agents are considering drinking coffee in order to be more productive. I will measure the productivity by saying that at each step of the model, agents not drinking coffee have a 50% chance of success, or, on average agents gain 0.5 points of progress per step. We can think of a step of the model as a work week, and then on weekends agents reflect on their weeks and discuss them with connected agents. The difficulty is that agents are trying to decide if coffee is working for them, as they don't know whether it increases or decreases productivity. To let us use Bayes's theorem, we need hypotheses, so the hypotheses will be that it either changes that success chance to 54% or 46%. In reality many people find coffee makes working a little easier, so the general case will be that for all agents the actual success rate will be that higher 54% chance (I will consider a case where it is not the

same for everyone below). These numbers are taken from Bala and Goyal's more abstract learning problem [8].

The choice of coffee comes from the fact that people with ADHD have different reactions to stimulants of all kinds. ADHD has been linked to increased use of caffeine [18], and ADHD and caffeine use together have been linked to worse overall well-being [19]. As well, the ADHD community discusses differences in how stimulants affect them in general, with some joking that stimulants prescribed for ADHD being controlled for their addictive properties does not stop them from forgetting to take them, which is incidentally supported by the finding that people with ADHD develop substance abuse disorders less frequently when they were previously prescribed stimulants [20]. If my choice of benign example seems unfitting, replace "coffee" in the proceeding with the stimulant of your choice (your choice for discussion, that is).

One consequence of this numerical setup is that if an agent has access to information only from agents trying option 1, they will never change opinion in either direction, as option 1 has the same likelihood of working no matter what is true about option 2. If you already know that option 1 has a 50% chance of success, then neither succeeding or failing with it makes you more likely to think that option 2 has a 54% chance of success as opposed to a 46% chance. So it is possible for a community to become stuck on option 1, and often agents will only be getting helpful information from some sources, possibly including themselves. However, community convergence to the more successful action 2 is still overall more likely than a state of permanent inability to escape the incorrect answer, that is, absent any other considerations (ie in epistemically just situations where nobody is in any way stopped from learning just as well as the others). This makes it a good baseline to add new features, so let us look at a few.

This is not the only possible numerical setup. Network models of this kind have been created that have two unknown options, as in Zollman's paper "the epistemic benefit of transient diversity" [21]. The reason Zollman uses this type of setup is that he is specifically

modeling a situation in which getting an entire community to agree is of specific difficulty. This setup is necessary for him to show that temporary diversity of opinion is important for eventual correct agreement. This paper was crucial for the development of a “cognitive diversity” literature in formal epistemology, which investigates the effects on science of having a diverse range of thinkers, meaning scientists who are more or less prone to risk, exploration, starting in different places, and so on. I will discuss below why this literature is less relevant to my work than it may appear at first blush.

This said, the present problem is still better represented by the version of the problem with one definite option. As Jingyu Wu explains, this version of the problem is typically used for problems like introducing a new drug, which is similar to the example I have chosen. Wu is also using a bias model in her paper “epistemic advantage on the margin: a network standpoint epistemology,” which provides another reason to stick to the version she uses; I am avoiding an unmotivated divergence from the closest previous work, and I will be better able to compare and contrast results with the existing literature.

To start with, there are a few obvious ways in which epistemic injustice can occur, and I wish to model these to give more detail on how they occur, and give a clear baseline for less common examples. Firstly, if a person or group of people have a bias against an identity, clearly those people are likely to not lend proper credence to the testimony of others with the biased-against identity. If there is a belief, spoken or not, that neurodivergent people are less capable knowers either in general or on specific topics like keeping up with work, then whether or not this belief is well-founded as a generalization, the result will be that neurodivergent people are believed less often. This can have a number of negative outcomes, but I want to focus on modeling the results this can have on learning for neurodivergent people and for communities that include this kind of bias in general. In our running coffee example, suppose agent 1 is talking to agents 2 and 3. Agent 3 has the “neurodivergent” marker. The idea is that if agent 1 thinks coffee is good, they are likely to listen to agent 2 no matter what they say, but if agent 3 says coffee is bad, they are less interested in

listening. The interesting thing to see will be whether and how the existence of this kind of bias affects both how communities converge overall, and in particular how agents with the “neurodivergent” marker change how they understand their own coffee use when not listened to by dissenting agents.

As Medina noted, another possible source of epistemic injustice is simply excluding people from discussions before they occur [12]. Real people are, more or less, free to associate with whomever they please, or at least completely ignore whomever they please. Contrary to a normal network learning model, people can form new connections, and lose connections, often in a motivated way. Being socially ostracized is a very common experience for neurodivergent people, and will naturally lead to a form of marginalization that lessens the presence of their opinions in a larger social context. Modeling how this runs involves giving agents a new mechanism for changing the shape of their own networks on the basis of belief. It is then possible to look at the effect of a dynamic network on how views change.

It has also been noted that neurodivergent people can struggle to communicate with neurotypical people (or, equivalently, that neurotypical people can struggle to communicate with neurodivergent people) [22]. The mere fact that misunderstandings are more common across neurotypes may also be a driver of epistemic injustice⁵. For this kind of case, suppose agents 1 and 2 have different *language* tags, representing different styles of communication. Then, when communicating about coffee, there will be some percent chance of evidence in one direction being taken as evidence in the opposite direction. Again, then, it will be interesting to see how much more of a disadvantage this miscommunication will be among individuals with the less common language.

Finally, most germane to neurodivergence in particular are the bald differences in how people actually experience the world. The topic is coffee—or, recall, “coffee”—and peo-

⁵To address the potential argument that this should not be called injustice as it is accidental, I argue that this kind of misunderstanding results more from structural ableism than from bare differences in preferred communication style, and that in a just world more patience would be practiced in cases of initial misunderstanding anyway.

ple with ADHD can have wildly different reactions to stimulants than neurotypical people. One element that could affect each of the above situations is that individuals with the identity marker “neurodivergent” may not have the same reactions to coffee as others. So, in a model, the expected productivity might be different across the two options. However, this can appear in a couple of different ways. In particular, how someone comes to understand these differences could vary wildly depending on whether or not they are aware that they might have a different neurotype from the people around them at all. Someone who knows they have ADHD *and* that this can affect how coffee works for them will investigate coffee differently from someone who knows nothing about ADHD at all, and believes that their experience with coffee should mirror others’. So, a model exploring difference among agents would employ differences in actual expected productivity between actions, and could also have a different level of baseline productivity⁶. More importantly though, in a Bayesian model, there is also a choice to make about who has what hypotheses available, which models different environments of general understanding of and attitudes towards ADHD. If everyone has the hypothesis that coffee will be as effective for them as it is for some ADHD people, this would be a high-information environment with a lot of room for people to accept they may have ADHD, and if nobody does there may be very low information or very low acceptance.

As mentioned before, Fricker’s position is that while mere differences can lead to epistemically unfortunate outcomes, unless someone is being discriminated against on an axis of identity unfairly, it is not a situation of epistemic injustice. I sided with Dotson that assumptions that seem accidental or necessary may often turn out to be systemic, pointed, and avoidable [16]. In the case of ADHD and caffeine, if the agents lack the hypothesis that coffee will affect them how it affects someone with ADHD, this can be a social failing. At

⁶I again note that under contemporary disability theory, saying that disabled people are less productive in their lives absent interventions does not require that they are inherently less useful people—it may be that this difference in productivity comes from an ableist society not well set up to allow for the conditions in which these people are most productive. On the other hand, disabilities that simply make people less productive also exist, and still do not devalue the people who live with them as people. Productivity in a work setting is not everything.

any rate, whether or not differences are alone enough for injustice, the structure of a model that accounts for difference can be applied to any of the above models as well, in order to investigate how difference exasperates those forms of injustice.

1.2 Network Learning Models

Cailin O'Connor and James Owen Weatherall, in several journal articles and books both together and separately, most notably O'Connor's book *The Origins of Unfairness* [17], and together the book *The Misinformation Age*, are using formal epistemology to tackle social issues [7]. The two draw heavily on the paper "Learning from Neighbours" by Venkatesh Bala and Sanjeev Goyal, which establishes a probabilist formal methodology for studying situations of learning in a social network [8].

My guiding question of how the social situation of knowledge leads to epistemic injustice against neurodivergent people is in some ways similar and in some ways dissimilar from the types of questions O'Connor and Weatherall tend to ask. For example, in their paper "Scientific Polarization," they explore how bias can lead to a community becoming split into factions with distinct beliefs even though each member of the community is earnestly trying to find truth and doing and sharing their own good research [9]. They can be said to be asking the question, "how can bias impede learning when nothing else is?" I will ask a similar question below about identity-based bias, but I will remove the "nothing else." It is prima facie plausible that bias acts differently when different factors are in play, and this will turn out to be reflected in my models. In this way, by combining different factors that change information flow in a network, I already differ from the existing literature, which mostly dissects particular issues in isolation. The other difference is that I am not looking at the concept of "bias" as a clean, abstract notion. I am looking at specifically identity-based bias, and when combined with other features, centered around a certain kind of identity marker, neurodivergence. This grounding in a specific kind of experience

narrows the application of my results in some ways, but it also gives useful guidance and potentially makes it far more powerful to apply to that situation than previous work has been in general.

In the next section I will explain the methodology of applying probabilist formal epistemology to social issues as it has been practiced, as well as why and how I intend to iterate on it for this project.

1.2.1 Bayesian Epistemic Network Models

Say I want to model a community of learners.

First, I must model a learner. I will call this learner an *agent*. My agent is given some kind of information. More on this later, but this information will only be evidence about the state of the world, what is called a *pure learning experience*. The agent is trying to differentiate between different states the world could possibly be in. There will have to be a finite list of possible ways it could be—while this list could be arbitrarily long, here I will simplify things considerably by only specifying state *a* or state *b* as possibilities. When running the model below, it will always be the case that the world is in state *b*, but the theory can cover situations where the world changes states between steps. At any rate, the agent does not start knowing the truth about what state the world is in, or in the case that the world states are determined probabilistically each step, what the objective probabilities of each state are. The information the agent receives is going to be, again, one at a time out of a list of possible pieces of information. For example if world state *a* is that there is a coin being flipped that is biased strongly toward “heads” and world state *b* is that a fair coin is being flipped, the two kinds of information it can receive are “the coin landed heads” and “the coin landed tails.” Because each state has a specific probability it predicts for heads and tails—stipulating that world state *a* refers to a specific bias and therefore a specific probability of landing “heads”—each piece of information will give the agent a clue as to

which world state obtained⁷. The question is how should the agent change its beliefs based on these clues.

Probabilist methodology answers by representing the relative degrees of belief the agent has in each world state as subjective probabilities from 0 to 1, and using Bayes's theorem for conditional probability, given as such:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where $P(A|B)$ is the probability of the thing we are trying to determine, A , on the condition that B occurs or holds true; $P(B|A)$ is the probability that B would occur or hold true on the condition that A occurs or holds true, and $P(A)$ and $P(B)$ are the *prior* probabilities of A and B , or how subjectively probable the agent held them to be beforehand. In other words, the probability that a hypothesis is true after observing some evidence is however probable that hypothesis was before, multiplied by the ratio of how probable that evidence was according to the hypothesis to how probable that evidence was subjectively held to be before. There may be some initial unease that both sides of the equation have this new idea of conditional probability, but recall that the world states give individual probabilities for each piece of information to occur, which is what $P(B|A)$ stands in for. This does further limit applicability, since if the learning problem being modeled does not allow the possibility of naming probabilities of each possible piece of information, then the theorem cannot be used. However, the theorem only requires that some numbers be chosen, as long as they successfully characterize which information points to which hypotheses and roughly how strongly, the theorem will still do roughly the work required, so situations where it does not apply are fairly narrow, requiring some kind of fundamental indeterminacy.

⁷I put to the side issues of whether we can be sure that the evidence we observe is real; there is a place later for discussing disbelief that certain observations being reported to the agent have actually occurred, but the model is greatly simplified if we take for granted that observations are themselves veridical. To do otherwise would both be difficult—how much credence should be placed in the idea that an observation is untrustworthy?—and pointless—all hypotheses should be equally hurt by this consideration, so end results should not change relative to each other, at worst being eaten into equally by some extra hypothesis in favour of solipsism.

There are a number of motivations for using probabilities and Bayes's theorem. There is a large body of work describing various optimality results for usage of the theorem, for example. It can be argued that constraints that an agent be *rational* in a strong sense require use of Bayes's theorem, owing to this optimality. This kind of motivation is somewhat controversial, with others giving metrics by which another learning methodology performs better. To accept the methodology presented here it is not important to come down on one side of this debate, so I will not do so here⁸. Instead, the motivation I will most rest on is that it is a standout among learning algorithms in being both very effective *and* very simple. It is at least powerful enough that agents using only this learning rule will converge to the correct result the majority of the time in an incredibly wide field of situations. It is at least simple enough to implement in a computer program in relatively little time. It is also robust enough that it can survive being meddled with at various stages, which will allow me to implement additional features to describe the conditions for epistemic injustice. As well, network learning models used to model real social phenomena have already seen success using probabilist learning, as in the cases of O'Connor, Weatherall, Bala, and Goyal [7] [8]. If I were to concede a different learning algorithm was more powerful for my context, it would not follow that this would be a better algorithm to use, for this algorithm may be more time-consuming to implement than its increase in fidelity is worth, it may not be as easy to modify, and it would not easily lend itself to comparison to previous work for the sake of evaluating results.

Before moving on from this part of the model, it is worth going back to what was meant by "pure learning experiences." It is possible for a real person to have their beliefs transformed in many ways by information and experiences other than by learning. In a probabilist framework, the currently available hypotheses must exhaust probability space, that is, the total probabilities for all hypotheses should sum to 1. Therefore, if someone introduces a new hypothesis, a probabilist giving non-zero credence to this new hypothesis

⁸Unsurprisingly, however, my view is that the probabilist side is generally better supported.

will have to decrease their credences in other hypotheses to make room for the new one. This is a change in probabilities for hypotheses, but does not involve learning of new evidence, and therefore would not be termed a pure learning experience. Another example would be reasoning about evidence—an imperfect probabilist might notice that some evidence confirms a hypothesis more strongly than they initially realized, and be forced to update to make up for this earlier error, which again would not constitute new evidence. The reason I find it satisfactory to look only at pure learning experiences is that, as we will see, my models are trying to work at the population level, not the individual level, and are trying to capture a snapshot of how dynamics play out in the present. A picture of a situation that includes only pure learning experiences should be able to accurately model situations that are stable at the population level but include some individual transformation, as long as the ideas people are considering stay roughly the same. At any rate, if this is not true, we can simply limit our analysis to situations in which there is no other kind of belief-changing experience, and remember that we are doing this below.

There is one more step for defining the methodology. In these models, the agent will be capable of taking *actions*. What these actions are will be defined at the same time as the world states are defined. If we continue to use coin flips, the actions can be gambling. The agent either places a bet that the coin will land heads, or that it will land tails. For simplicity, the agent will not have to calculate the best amount to gamble or anything like that, this is just our arbitrary way of defining two possible actions. If the coin lands heads, agents who bet on “heads” will receive a *payout* of 1, and those who bet on “tails” will receive 0, and vice versa for tails. This payout does not do anything for the agent, except give it a reason to take one action or another; although it doesn’t improve in any way for collecting payouts, it will always take the action with the highest subjective expected payout. Bala and Goyal describe this as maximizing the one-step expected utility for an agent [8], and describe this in an equation I will give shortly, but which requires describing the model to understand.

In a model, a set of states the world can take is specified. The set of possible states will be denoted Θ . Any given state in this set will be referred to with a lowercase θ . Without loss of generality, we can have a global probability function over Θ that describes a probability of each state occurring each *step* of the model. For example, an even coin gives a measure $P := P(\theta_H) = .5; P(\theta_T) = .5$, where θ_H is the state of the coin being in heads, and θ_T is tails; but we could also model the question, “is there a coin?” with possible hypotheses θ_a that there is a coin and θ_b that there is not, with $P := P(\theta_a) = 1; P(\theta_b) = 0$, or even adding possibilities like “it is indeterminate whether or not there is a coin” as θ_c also with a 0 probability, and so on.

There will also be a defined set of actions each agent is allowed to take, denoted X , and a set of outcomes Y , with individual elements named x and y respectively. Since Y is a descriptive list, we need a function $r(x, y)$ to map outcomes to values of rewards, and since it is sometimes desirable for the framework to handle uncertain outcomes⁹, there is a measure $\phi(y; x, \theta)$ giving probabilities of each outcome conditional on each state-action pair. Again, these can be limited to 1 or 0 on particular models if the world should not be chancy.

I can finally give the equation for determining how an agent will act on this methodology. Let $u(x, \mu)$ be the expected utility of action x conditional on belief measure μ . Then,

$$u(x, \mu) = \sum_{\theta \in \Theta} \mu(\theta) r(x, y) \phi(y; x, \theta).$$

In words, the expected utility of an action given a set of beliefs is the sum of the expected utility of that action in each possible state, weighted by the subjective probability of each state. To be more plain, at each step, agents will take the action that appears best for them just for that one step, and not take actions that are less appealing in the short term in order to gain information.

⁹Indeed, it is usually desirable; if a world state has probability 0 of a specific piece of information, it becomes very easy to rule that state out, by observing that information. Therefore, learning problems usually have an agent trying to decide between chancy world states, in order to be non-trivial.

Now that we have agents and an environment, we create a network. Each agent is given a unique number from 0 to $n - 1$, where there are n agents, both to identify it and determine the order in which they act¹⁰. Each agent will have a list of other agents associated with it. If agent a is on agent b 's list, then so too will agent b be on agent a 's. This can be defined in any way, but I will do so randomly in order to generate a large number of networks quickly. This list will be called the agent's *neighbourhood*. This is the same as the agents being nodes on a graph, connected by vertices. For example, see figure 1.1 for the initial graph of connections generated for one of my models. I will switch between the two methods of talking about neighbourhoods, but prefer to think of it visually, as a graph.

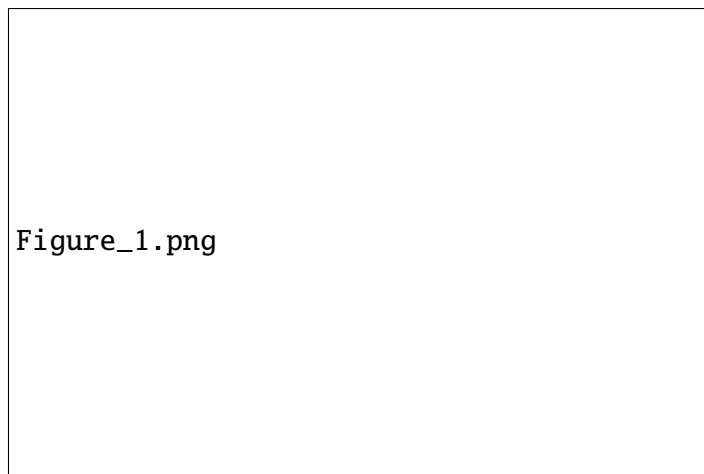


Figure 1.1: A network of ten agents labeled 0-9 connected randomly to each other, coloured depending on their random starting beliefs

Once everything has been defined for a model, it can proceed in steps. For each step, a state θ is decided upon for the world based on the probability measure ϕ . Then agents take turn going in order going through a short list of instructions, as follows. First, the agent takes the action it deems optimal for this step. Then it observes the reward determined based on the state of the world. It then uses Bayes's theorem to update its beliefs based on what reward it received. Finally, it finds out what the most recent observation seen by each

¹⁰Numbering starts at 0 due to that being standard in programming languages, including Python, which I will be using. In papers that do not use computer models, numbering usually goes 1 to n . This does not make a difference.

of its neighbours was, and updates its beliefs for each of these the same way. Once agent $n - 1$ has acted, the next step begins. This continues for as many steps as desired (or, if you like, continues forever, and I will look only at a large finite subsection of steps).

1.2.2 Why These Models

The groundwork for this kind of project was laid, again, by Bala and Goyal in “Learning from neighbours.” The paper itself is quite sparse on conclusions, only really demonstrating that in a situation with limited connections, limit of probability of convergence to the optimal action as agents increase is one, and that the existence of a central agent or set of agents that all other agents observe is capable of stalling a network that otherwise would have converged to the right answer. The analogy to real life is not made at all; while the central agents are called the “royal family,” the role they play does not seem to represent any actual royal family’s role in any epistemic process. Beyond this, the paper describes itself as developing a framework for studying how structure of social networks impacts learning. The simple fact that a limited network generally converges but that changing the structure to be *more* connected can change this is an interesting and worthwhile feature. Anyone using this framework can model a variety of situations to study what affects convergence in a community, and in what way [8].

It is worth knowing that Venkatesh Bala and Sanjeev Goyal are economists, and that their immediate uptake was in economics. It is Kevin Zollman who brought their framework to philosophy with the paper “The Communication Structure of Epistemic Networks” [23]. Zollman’s innovation is to look at the speed of convergence, not only probability; while more connectivity makes convergence less likely, it also makes networks that do converge do so much faster. Beyond this modest result this paper also made the framework more visible to philosophers, like Bennett Holman and Justin Bruner, who, in “The Problem of Intransigently Biased Agents,” introduced a biased agent to the epistemic community, who desired to convince the community of one hypothesis, whether true or not. They showed

that while communities of earnest learners are more likely to converge to the truth when there are fewer learners and connections, this reverses when exactly one agent is trying to convince the community of a falsehood. What's more, when allowed to choose their connections, other agents were able to sniff out the biased agent quite frequently, and the destructive effect of the biased agent was greatly mitigated. They concluded that scientists should be vigilant of biased sources like studies run by pharmaceutical companies [24].

Bruner has also worked with O'Connor and Weatherall, the most direct influences of the present work. Together they developed the propaganda network model present in *The Misinformation Age* and "How to Beat Science and Influence People," which has a similar goal to Bruner's earlier work, but introduces more robust networks of agents using a different format for influencing the community at large; *propagandists* do not outright lie, but filter results so that only certain ones are shared, and greatly impact convergence this way [7] [25].

One important example from the cognitive diversity literature in formal epistemology is a recent paper that, like this dissertation, is also inspired by Fricker, "Epistemic advantage on the margin: a network standpoint epistemology" by Jingyi Wu [26]. Wu looks at three models, one of which is formally identical to one of my own¹¹. The first considers situations in which marginalized agents are completely ignored by powerful ones, the second changes complete ignoring to a kind of bias modeled by Jeffrey conditionalizing, as my model does, and the third biases network formation to initially mostly include only agents that share group membership. She ultimately finds that being biased against is generally good for the biased-against agents, a finding that supports standpoint epistemology. However, Wu calls out her lack of modeling the lack of shared realities among groups of people as a weakness; my model will therefore advance the literature beyond this similar paper by attaching one of its models to my difference model.

¹¹The move to take O'Connor and Weatherall's use of Jeffrey conditionalizing to represent bias and apply it only to agents bearing an identity tag is a very natural one, so it is not surprising that two people independently made this move unaware of each other. This does not impact the novelty of my larger model, which incorporates bias alongside other features.

There are a number of other papers that are part of this literature using probabilist epistemic networks, but none are as directly relevant. In particular, Kummerfield and Zollman have a model arguing that funding bodies can promote diverse research by funding more exploratory work [27], Rosenstock, Bruner, and O'Connor argue that decreases in communication in order to promote diversity of research are not worth the inefficiencies they cause [28], and Holman and Bruner show that enough diversity of research can be abused by industry interests [29]. These examples should illustrate a few things about the cognitive diversity literature. Importantly, they are discussing only transient diversity within science. While there is a live debate as to whether disability should be taken to be transient¹², this type of transient diversity works in the inverse way to what is being discussed here. Transient diversity in formal epistemology refers to communities temporarily diversifying their opinions on a matter, then coming to agreement; if an analogy were drawn to disability, it would have to be that some people are becoming disabled, then ceasing to be disabled as the community comes to understand their disabled standpoint. However, in disability studies, transience of ability refers to the capacity for anyone to become disabled, and not the ability for anyone to cease to be disabled, so there is no clear applicability. The type of diversity in the cognitive diversity literature is fundamentally mutable, and therefore the concept does not align with discussion of neurodivergence as a matter of course. As well, the literature is particularly focused on science, which makes it harder to pull conclusions into non-academic spaces.

A similar paper is “diversity, trust, and conformity: a simulation study” by Sina Fazelpour and Daniel Steel. Fazelpour and Steel are specifically looking at social diversity *absent cognitive diversity*, which makes it more germane to my topic; while I am not using the social model to describe all disability, I maintain that some disability is social, and *prima facie* social diversity without cognitive diversity would be a way to model these kinds. The pa-

¹²See Bill Hughes’ critical discussion of the term “temporarily able-bodied” by other disability theorists [30]. It is outside the scope of this project to argue for either side of this debate; I recommend Hughes as a starting place for the completeness of his summary of the debate.

per makes a distinction between informational group influences, which are influences that come from a need to be able to share the same kind of information with group members, and normative group influences, which are influences that come from a need to adhere to group norms. The concept is that social diversity should counteract both influences, and uses group-based weighting, assigning groups based on initial priors. The paper ends up showing that the Zollman effect applies here, and some diversity of this kind is good [31]. While group tags are something I will use, basing them on priors is not something that appears to require further investigation.

Overall, Bala and Goyal's framework has yielded fruitful research on social epistemology with a formal methodology. It is, of course, not the only way to join these ideas, however, and a full justification should look at rivals.

In "A Bayesian Simulation Model of Group Deliberation and Polarization," Erik Olsson discusses the *Laputa*¹³ computer modeling framework. This framework is very similar to Bala and Goyal's, but is in some ways more powerful. Connections can be one-way, agents have thresholds for how strong their belief must be before communicating, and likewise for how trustworthy they find others before listening, and finally a percent chance of conducting inquiry each step. Olsson is interested in modeling argumentation in social epistemology, which he says is too-often ignored, so the framework fits perfectly. He finds that certain social parameters *polarize* the community towards a certain outcome regardless of the truth, causing them to be far more likely to come to a specific outcome. Note the difference in usage here from how O'Connor and Weatherall use "polarization." For Olsson, a community is said to be "polarized" when a community is especially likely to converge to one thing for reasons other than bias towards truth, but for O'Connor and Weatherall, a community is said to be "polarized" when both poles are represented in the community [32]. While Laputa is an interesting and powerful tool, it neither offers any clear advantage to the specific situations I will discuss below, nor does it allow the imple-

¹³Apologies to any Spanish speakers; the name appears to be a reference to Gulliver's Travels, not carrying the vulgarity Swift may or may not have intended.

mentation of several considerations I want to look at.

Consider also work by O'Connor not using the present framework. In *Origins of Unfairness* and “The Cultural Red King Effect,” she uses formal tools to argue for a *Red King effect*, the inverse of a Red Queen effect. The more a minority group is outnumbered, the more it will have to interact with the majority group, which can be a detrimental feature in bargaining. In a hawk-dove game¹⁴ designed to represent bargaining over a resource split, for example, as long as agents can tell which group an opponent belongs to, it is easier for a majority group to enforce an equilibrium whereby majority members play “hawk” against minority members, who play “dove,” which is favourable to the majority. Each individual majority member has an easy time sometimes losing big in matchups against minority group members, compared to the minority group members struggling to lose big time and time again, so “hawk” is a much safer play for majority group members, and this becomes the equilibrium the majority of the time [33] [17]. This is a worthwhile result that does not use the framework I intend to use.

Another paper not using the framework is Mayo-Wilson et al., “The Independence Thesis: When Individual and Social Epistemology Diverge.” Here the authors establish five independence theorems, exploring the connection, or lack thereof, between individual and group rationality in learning problems [34]. This is an important result for individual recommendations of research methodology, since it can be justified to recommend individually irrational action, like testing theories one does not believe in.

One type of model that is also explored a lot in the cognitive diversity literature is the landscape model. The original model comes from Weisberg and Muldoon. Their model

¹⁴In a hawk-dove game, there are two moves, *hawk*, and *dove*. If both players play hawk, they have a harshly negative outcome. If both play dove, they have a slightly positive outcome. If one plays dove and the other hawk, the dove player has a slightly negative outcome, and the hawk player has a greatly positive outcome. An example is traffic laws. Dove is following the law—it works well enough normally, and others slow you down a bit when taking advantage of you. Hawk is breaking the law—as long as everyone else follows the law you can get home earlier, but as soon as another driver does the same thing, you risk getting into an accident. In a world of doves, there is no selfish reason not to run a red light when others going the speed limit will be able to brake in time, but if there are sufficiently many other hawks, even a selfish driver should get defensive.

has a height function over a 2-dimensional landscape where height represents viability, and 2-dimensional location represents a research program, so that movement on one dimension models changing the program in one abstract way; obviously there are more than two dimensions defining research programs, but this is a worthwhile simplification. The argument in their paper is that a diverse population of researchers finds the highest/most viable programs best, though later work has mostly agreed that the model fails to show this, and changes need to be made to actually show this [35]. For example, Thoma's types of researchers are "explorers" who roam around and "extractors" who prefer to stay in the highest places they know, which are actually shown to be better in diverse communities [36]. Pöyhönen, on the other hand, focuses on the social aspect of exploration, varying willingness to follow others directly among agents, and also finds that diversity of social predilections is helpful [37]. The strongest counterpoint seems to be Alexander et al., who show that a behaviour of "swarming" can be defined that leads to individual agents behaving differently based on context, but which is optimal despite not representing diversity of strategy; this paper, however, significantly tempers its conclusion by noting that it is unrealistic to expect real people to follow such a strategy [38].

In summary, landscape models show interesting things about exploration of conceptual space, but as before, their focus on scientific context makes it hard to apply any previous work to epistemic injustice, and the type of diversity explored is not relevant to disability studies. This does not mean future work using landscapes could not touch on epistemic injustice towards neurodivergent people. A type of landscape model that defined different height functions for different agents and considered different kinds of communication could model conceptual space in a way that connects to hermeneutic injustice, but there would be a massive disconnect between such a model and the landscape model literature that presently exists. At any rate, while this could be a fruitful area of research to consider in the future, it does not capture the type of dynamics my models are interested in, and it will not be possible to draw inspiration from them for the present project.

Overall, it seems that Bala and Goyal’s framework has been useful in some of the most impressive results in the intersection of these fields. More importantly, out of available methods, it best fits the kind of question I am asking here. Results without the framework have either been more formal, or less epistemic, in general. O’Connor’s bargaining game is an important result in structural causes of injustice, but not of structural causes of *epistemic* injustice. Likewise, the independence result is worth knowing for normative rationality claims about groups of researchers, but not for normative claims about just *distribution* of epistemic resources. The track record of various frameworks proves that Bala and Goyal’s is best-suited to answering this kind of question, and that no other framework that works well seems well-suited to answering this kind of question.

1.2.3 Using Computer Models

The methodology of setting up a specific kind of formal model and exploring how certain results change as parameters change is powerful. For example, it can show that the probability that a network will converge to acting correctly changes based on level of connectivity. However, all of the power that can be captured by giving proofs of probabilities in the limit and so on can also be captured by computerized statistical models. Following the example, if the probability of convergence actually changes substantially, then over a thousand networks, the number of networks that actually converge in a certain number of steps should also change. While statistical coincidences are logically possible, risk of misleading results is negligible to the point of virtual nonexistence in a dataset of the size that a computer can generate automatically¹⁵. On the other hand, there may be some advan-

¹⁵To quantify, suppose the actual probability of convergence by a specific step changes as little as from 90% to 80% between two kinds of models. The chance that out of 1000 runs each, 850 or more of the 80% networks converge successfully is $\sum_{x=850}^{1000} \binom{1000}{x} 0.8^x 0.2^{1000-x} = 4.49 \times 10^{-7}$ and the chance that 850 or less of the 90% ones do is $1 - \sum_{x=150}^{1000} \binom{1000}{x} 0.1^x 0.9^{1000-x} = 2.77 \times 10^{-7}$. By contrast, the 80% model will give the exact correct proportion, 800 converging networks, with probability $\binom{1000}{800} 0.8^{800} 0.2^{200} = 0.0315$, and the 90% with probability $\binom{1000}{900} 0.9^{900} 0.1^{100} = 0.0420$. In words, it is ten billion times more likely that a pair of batches give the *exact* expectation value for each model than that a pair of batches makes the less likely appear more likely. Many of the differences observed in the data are far larger than this.

tages to a statistical approach. For example, one could collect multiple statistics at once, and measure covariance. This kind of computer model, then, lives up to the task described above of finding something to play the role of a study in social science, and allows the same kind of statistical analysis, with a different set of qualifications on applying the results.

My methodology, which will use a computer program to generate statistics, is therefore not unlike the Laputa program utilized by Olsson. While Laputa can be used to hand-craft a specific network for study, it also has a batch function that Olsson describes as its most powerful function. One can set certain parameters to vary in a particular way across a large batch of model runs, and receive statistical information out of it. It is this kind of statistical information that he used in his paper [32]. As mentioned above, however, I will not be using the actual Laputa program, and instead will write a program in Python using the Mesa and Networkx libraries which have been built to make agent-based models like these quick and simple to design and run in the language. While this means I cannot boast a project output of an easy-to-use program complete with UI and ready to take on any number of projects easily, like Laputa, I at least have a code appendix that will allow reproduction, and can be, with some work, reworked to fit a number of other projects on the scale of mine. As well, this methodology allows me to test a number of features that Laputa cannot presently test.

The approach of using a computer model in place of another kind of experiment is not entirely novel in philosophy. In addition to some of the above mentioned papers in this subdiscipline, Mayo-Wilson and Zollman have argued together in “The Computational Philosophy: Simulation as a Core Philosophical Method” that computational models are an excellent tool that can in some cases play a similar role in formal arguments that thought experiments play in other philosophical settings [39]. I would tend to agree. Most philosophical arguments are greatly enhanced by walking through a specific example for demonstration. In a lot of settings, the way these can pump one’s intuition is enough to make a compelling case, and no contact with reality outside of one’s head is necessary. However, in a setting where the problems being discussed are sociological, where there is both great

complexity and great need to get it right as it impacts a lot of real people, there is need to go beyond thought experiments. Computer simulations, like empirical experiments, have the advantage of being able to correct faulty intuitions. Attention should be paid to making sure these simulations have what Fazelpour and Steel call “empirically sensitive robustness.” That is, their results are robust where we have empirical reason to expect them, and not elsewhere [31]. This can be achieved by taking special care to draw on empirical evidence in model design, as I have done, and undertaking empirical investigation of any novel results, which is beyond the scope of this project.

Generating random networks also has the benefit of generating example networks. In *The Misinformation Age*, the authors include examples walking through how specific networks evolve over a number of steps [7]. This provides a level of depth that is missed by purely numerical argument—one may be persuaded to agree with O’Connor and Weatherall’s conclusions by being shown formal results, but this does not confer understanding of *why* the formal results are what they are. Conferring this understanding is a better outcome than merely changing opinions; it is one thing to have the correct view on how propagandists change opinions, and quite another to have a deep enough knowledge to enact efficacious solutions. The shorter papers referenced above generally do not do this, contented with reporting formal results. I view it as a strength of the framework overall to be able to give a coherent narrative of what is occurring, and this is a strength I wish to lean into.

My specific methodology will make this easier and more powerful. Rather than generating a random one thousand networks anew for each model, I will generate a thousand random starting situations that are run through each model, with the same random seeds for the random number generators in each model. As a result, the pseudorandom results will be the same between two models, just with sometimes different outcomes. Step 137 of network 481 will always have agent 3 generate a 0.52 on the outcome randomizer, so that it will see a failure if it takes action 1 (50% chance of success) and a success if it takes action

2 (54% chance of success). The reason different models will have different results is that what happens after this number is generated will differ. Depending on the model, agent 3 might be giving that information to a different set of neighbours, those neighbours may misunderstand agent 3, and so on. Therefore, by tracking which networks are changing between kinds of models, the computer program will point us directly to which networks to look at to see examples of why things turned out differently. If network 481 converges in one model type and not another, we can look at the steps in detail to see why the same starting configuration and random seed played out differently. This speeds up the process of contriving examples, and, also importantly, assures that the examples given to the reader accurately reflect what actual differences end up making a difference. It is possible for an author to be wrong about why a result changes, so getting the example directly from the dataset is desirable.

Chapter 2

Network Learning: Modeling

Testimonial Injustice

Now that the foundation has been laid, this chapter will dive into the mechanics of my network learning model tailored for epistemic injustice, especially testimonial injustice against neurodivergent people.

First, I will present a mathematical learning problem—one that is standard in the literature—that I have modeled with a number of different complications in order to represent the real life concerns described in the previous chapter. In particular, the learning dynamics will be affected by differences in best actions and communication style, biases between individuals, and association with like-minded others.

Most of these are mathematically novel in the literature, introducing elements that have not been studied formally in the past, at least in this way¹. The major exception is that I will be using O'Connor and Weatherall's method for modeling bias in a network learning model, the differences being that I will approach it with my statistical methodology, and

¹Naturally, people working in formal philosophy will be drawn to important issues, so anyone working on important issues is likely to be working on something that has been modeled in some way. As has been discussed extensively, details of a model have a massive impact on the results it appears to suggest; thus, when dealing with a problem in a specific domain, it is often necessary to tweak existing models or create new ones from whole cloth. What I mean, then, is that the use of Jeffrey conditionalizing to model bias is the only network model type here I would characterize as tweaking something that exists in the literature.

that I will be limiting to bias to identity tags instead of having all agents biased against all other agents². The two reasons for including this kind of model are, firstly, to serve as a control, since there is existing research on how it impacts a network, and secondly, to see how bias *combines* with other features. It is also worth noting that the concept behind the mingle model is only somewhat novel; previous work in probabilist networks have looked at static networks that are initially built in a way that biases results in one direction [7][26], and dynamic networks have been used in the very different context of opinion models [40]. However, the model I will describe below is entirely novel in how its dynamics work, and using the dynamics for a changing learning network.

After explaining the details of the model, I will present the results. The standard learning problem is itself fairly easy for the formal agents, but this changes rapidly as breakdowns in communication or understanding are introduced. In particular, the results will suggest that structural problems are more important than individual bias, that minority group members are helped by expanding their visibility to each other and others, and that individual bias causes a one-way epistemic homogenizing effect on the biased-against party. The specific way in which minority group members are helped involves those members being more coherent to each other than to the majority group, which can result in a solidarity that both helps the individuals maintain their own views to themselves and makes those views more visible to the majority. The homogenizing effect results from biased-against divergent agents not being exposed to views like their own as much as the majority view due to a lack of dissemination of the latter, which creates a sort of screen allowing only one kind of information to flow. From this I will emphasize the importance of creating communities that integrate disabled people with others in a meaningful way, to the benefit of all. I will then conclude the chapter with a note on the power of this kind of model.

²While it was published too late to be an inspiration for this project, see Wu (2022) for a very similar move [26]

2.1 Methodology and Results

2.1.1 The Code

I used the Python library Mesa to program a series of computer models based on the general model described by Bala and Goyal in “Learning from Neighbours” [8] and ran each on the same one thousand different randomized networks for ten thousand steps per network, logging data at 1, 10, 100, 1 000, and 10 000 steps³. Because I have made the code available in its entirety in appendix A, I will not describe it here, and will instead focus on the program in its theoretical form, followed by a general description of how the data were generated.

To reiterate, the general structure of this kind of model creates a network of agents randomly connected to each other, with random starting beliefs between two different possible world states, one of which actually always occurs, while the other does not. For this specific learning problem, in either world state action 1 succeeds 50% of the time, but action 2 is successful 54% of the time in the actual world state, and 46% of the time in the non-actual one. Therefore, an agent trying action 1 never changes beliefs, but trying action 2 increases belief in the actual world state when it succeeds, and increases belief in the non-actual world state when it fails. Each step of the model, the agents, which are numbered 0-9, take turns in numerical order taking the action that maximizes the short-term expected utility, conditionalizing their beliefs on the result of that action, and then also conditionalizing their beliefs on the results of their neighbours’ most recent actions if applicable.

Here is how I have implemented this in the code. First, I define a number of objects and functions. The objects are the agents and model. Some functions are within the objects, such as both agents and models have a “step” function that gets called each step, and others are external, such as the function for parsing the data.

³I attempted to prepare this and chapter 4’s model for supercomputer computation to get larger population sizes, but the code did not scale for multiple processors. Future work could make better use of computation resources by running model types single-threaded but as separate programs instead, and check whether the results below are similar at much higher populations.

Every time something in the model has an element of chance, I use the ubiquitous mathematical library NumPy to create a random number generator with a unique random seed that counts uniquely to that random process, in order to maximize the reproducibility of my results⁴. More plainly, random results are generated using a *Mersenne twister*, which chaotically cycles through a very large number of values even when starting points are very close, and each process is given a different very close starting place to ensure distinct results. In this case, when going through the 1000 models, each agent is given a starting place based on how many agents there have been in total, and each model is given a starting place based on how many models there have been in total. The effect this has is that even when changing variables like whether the agents have bias, every random process will yield the same values as previous times the program has been run. So not only can any reader get the exact same results as I did by running the same code, I can also track specific networks that are highly typical or atypical across multiple kinds of models, and see why they have the results they do. This is like being able to create the same community over and over again with slightly different situations to see how those differences affect things. In addition to this, we can be more sure that changes like adding bias are actually changing the outcomes, rather than taking the admittedly vanishingly small risk that a specific batch of one thousand runs through the model is not representative of how the model usually changes.

In one section of the code, intended to be edited each time it is used, there are a number of variables that affect how the models will be constructed and run, such as number of models to run, number of agents, probability of connections being formed between agents when constructing the network, and so on. In particular, the *type* of the model can be changed in between, and this is how the models are made to conform to the above specific

⁴The developers of NumPy no longer recommend the previous best practice, which was to set a singular random seed at the beginning of a program, and to use a single random number generator for all processes [41]. When employing the older method, processes from other libraries or parallel processes could cause unpredictable changes in how the random number generator is accessed, making results not entirely reproducible. Creating separate random number generators for each random process gives substantially more control.

examples. A model can be of a base type, working identical to Bala and Goyal's original learning problem, or any combination of a number of additional factors can be added. A model with type "b," for example, introduces bias against some or all agents in a manner described in the next section, depending on the number placed next to the "b" when giving the model type. The other type letters are "l" for "language," "m" for "mingle," and "d" for "difference." See section 1.2.2 for the justifications for these specific types of model.

Each model type introduces numerical parameters that affect how the models are run, such as strength of bias. Since there are a limited number of hours in the day, but an unlimited number of rational numbers that could be used for these values, I chose, with intention, one or two specific values for each of these, as described in the following subsection, and ran models using these values. When noting these values, I will justify my selections, and discuss the usefulness of exploring additional values.

Once the code runs through one thousand network models for ten thousand steps each, the code returns a massive database formatted according to the Python data science library Pandas. Included in the code is a function that extracts information about the agents in each model at certain steps, and outputs this parsed information in a spreadsheet for easy reading. In particular, the function tells me what percentage of networks had complete convergence, how many of those converged to the superior action 2, the average percentage of agreement in best action among networks, and average percentage of preference for the superior action 2 among networks. It also lists which runs out of the first hundred have converged to either action 1 or 2, and which have not converged. This gives us a general quantitative picture of how much of an effect each model type has on the network's ability to converge to the correct action, as well as helping us zoom in on specific networks worth analyzing for a qualitative look at how the networks learn differently under different model types. Finally, when running difference models, I also include a line that has the program report specifically on what percentage of agents with different payout matrices prefer action 2. The raw chart output of my code when run using many select model type combinations

is given in appendix A, and a summary of the important details will be given in section 2.2.1.

2.1.2 Model Types

The first type of model to discuss is the “b” type, or bias models. These can in turn be split into general bias and identity bias models. General bias models serve as a sort of control—they show us how much of a difference is made if every agent is biased against every other agent, so that we can compare this to the effect identity bias models have. When the code tries to conditionalize an agent’s beliefs on information gleaned from another agent, and it detects that agent is biased against the other, it switches from Bayes’s theorem to Jeffrey’s theorem. O’Connor and Weatherall use Jeffrey’s theorem when exploring bias in their paper *Scientific Polarization*, where they show that implementing it how I do in my general bias models introduces a considerable chance of communities failing to converge because the network fractures into two communities who refuse to listen to each other [9]. I will briefly explain Jeffrey’s rule before moving on.

Recall that Bayes’s theorem is written

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where $P(A|B)$ is the probability of A conditional on B . Jeffrey’s rule, on the other hand, is written

$$P_f(A) = P_i(A|B)P_f(B) + P_i(A|\sim B)P_f(\sim B)$$

where $P_f(A)$ is the final probability of A and $P_i(A)$ is the initial probability of A . According to this rule, the final probability of A depends on the final probability the agent has that B actually occurred. In other words, I am implementing bias by saying that if agent 1 is biased against agent 2, agent 1 will take any information agent 2 communicates to them with a lower credence. Rather than accept automatically that B occurred, as in other models,

agents will have a degree of scepticism based on who is telling them they experienced B . Of course the agents will all be trustworthy—situations in which agents have an incentive to lie are philosophically interesting but beyond the scope of this project—but some will at times be treated as untrustworthy. Now, the above rule, unlike Bayes’s theorem, does not give sufficient information to write a model, as a decision process is still needed to determine what $P_f(B)$ will be in any given case. O’Connor and Weatherall used the formula,

$$P_f(B)(d) = \text{MAX}(1 - dm(1 - P_i(B)), 0)$$

where final probability in B is taken as a function of d , the difference between the conditionalizing and reporting agents’ initial probabilities for B , and there is a coefficient m describing how sceptical the conditionalizing agent is in general. When d is 0, meaning the agents have the same initial probabilities, the agent defaults to simply believing that B occurred, and as difference increases, the final probability for B decreases as a function of how likely the agent initially judged B to be. The maximum function is required here to avoid negative probabilities, which would otherwise be possible depending on choice of m (notice that for any difference above 0 and initial probability of B below 1, an m can be chosen that would result in a negative probability).

In the case of general bias models, I do not differ from O’Connor and Weatherall. However, it is both the case that I am combining general bias models and models with other complications to create novel models, and that models using this kind of bias calculation alongside agent identity markers is novel. In the latter case, where a non-zero number is given beside “b,” that number of agents with the lowest ID numbers are given a marker, and agents lacking that marker use Jeffrey’s rule when conditionalizing on information given by marked agents, but not other unmarked agents. This means that members of the marked group are biased against to a lesser degree the more likely they are to agree with the biased agent. This seems to track with tokenizing that occurs in reality. I can recall conversations

where I was told it was okay for a straight person to say a homophobic slur because another queer person told them so, even though I, another queer person, said otherwise. In these cases someone against whom the listener has a glaringly obvious bias is still taken as a giver of worthwhile testimony, because the testimony is known to agree with the listener. My testimony that they were doing harm by using the slur was devalued compared to another queer person's testimony that they were not, because the person saying slurs already believed it was okay to do so⁵. An implementation like this where identity-tracking bias also tracks disagreement between the communicating parties is therefore desirable.

Finally, while running this kind of model, a choice of m must be made. I ran 1000 general bias models for 10 000 steps using 0.5, 0.75, 1, and 2 as values for m , and found that these had dramatically different effects. $m = 0.5$ did almost nothing, while $m = 2$ made it very difficult to converge at all. $m = 0.75$ and $m = 1$ still represented a dramatic difference in effect size, but both made it much easier to track the actual differences when other changes were made. I did runs with both $m = 0.75$ and $m = 1$, and $m = 0.75$ allowed other effects to be measurable to a much greater degree, so most of my analysis will use $m = 0.75$, though it will in places be worth noting the effect of increasing m . As before, and as with all other parameters, future work using more computation time testing other values for m may be valuable.

I will now discuss the language model type. When the "I" parameter is entered, the model gives a number of agents with the lowest ID numbers one language marker, and the rest another. The reason both language and bias choose from the lowest ID numbers is that in models in which multiple types are combined, I want to be measuring just one thing, even if that one thing is complex. In this example, what happens if the same agents are both biased against, and poorly understood, by the general population. I will do the same thing when implementing difference, for the same reason. As well, in general, when

⁵I acknowledge the possibility that in none of the scenarios I experienced did the cited queer person actually exist, but this kind of claim is common enough and there is enough disagreement within any given group that events with the same structure as the example have surely occurred.

running models using all of these factors, I will set the same number of agents to have all of these features, with one exception I will discuss below.

Like the bias model type, language models check each time one agent is conditionalizing on information gleaned from another agent to see if they bear different markers. In cases of bias, it made sense to scale credence in the veracity of testimony based on how different the two agents were and how unlikely the testimony seemed on its own. In cases of differences in communication, these factors should not affect learning, and instead I implemented a set probability of misunderstanding. When a misunderstanding occurs, the agent conditionalizes as if the opposite information was being given. For example, if agent 1 speaks language 1 and agent 9 speaks language 2, then when agent 1 learns that agent 9 just attempted action 2 and failed, there is some percentage chance that agent 1 will misinterpret what agent 9 is communicating, and conditionalize on the information that an agent attempted action 2 and succeeded. This means a percentage must be chosen. I tested language alone with probabilities of misinterpreting 0.2, 0.33, 0.4, 0.5, 0.66, and 0.8. Surprisingly, the results within the range of 0.2-0.5 were quite close most of the time, though the effect was notably easier to see for 0.5 than the lower values. On the other hand, from 0.66 and up, it became extremely difficult for communities to converge, and the effect began to dominate. This alone is interesting and worth discussing, but for our present purposes it suggests that when mixing types of models the value 0.5 seemed most prudent, as otherwise misinterpreting would dominate other considerations, or it would be difficult to see a difference.

Next is what I will call the mingling model type, or “m” in the code. When agents are allowed to mingle, they are not tied to their original network configuration, and can either sever old connections or forge new ones. This is done by adding a process to each step of the model, where agents first check if their neighbours have changed at all since their previous step (that is, if they have lost or gained a connection), and if not, there is some probability of adding and losing a connection, checked separately. To add a connection,

the agent first creates a list of all agents that are neighbours of its neighbours, excluding those to which it is already connected and those who have already had a change since their last step. It then weights those neighbours as a linear function of proximity of belief, and randomly chooses one to connect to based on those weights. This simulates agents making connections only to those that are relatively close to their neighbourhood anyway, and also being more likely to forge new connections when getting on well with that new connection. To lose a connection, the agent more simply makes a list of all of its neighbours, weights them as a linear function of distance of belief, and randomly chooses one to sever based on those weights. This simulates agents losing contact with or ceasing to pay attention to contacts with which they disagree most, with random variation according to other unaccounted-for variables. Any connection can be lost for many reasons—even a fellow traveler may one day develop unbearable hygiene, indeed depending on your beliefs they may often do so. Finally, if an agent has no connections at all at the end of this process, it will forge a new connection to whichever agent is most similar to it in belief.

This model type has two parameters, the probability of a connection being lost and the probability of one being gained. It is generally best for these to be the same probability, however, or else the network will tend towards a web of all connecting to all, or towards naught but bonded pairs, neither of which is a particularly interesting network. After experimenting with several values for this parameter on its own, probabilities of 0.05, 0.1, 0.2, 0.4, 0.8, and 1, the probability 0.4 was closest in effect size to other choices made above. There is not a clear real-world analogue to probability of mingling, so I ended up choosing 0.4 for all mixed models.

Finally, there are difference models. For the scope of this project, I ended up implementing only the version of difference models in which agents all conditionalize over the same set of hypotheses not including the hypothesis that is true for agents with the “divergent” tag. This situation represents a general lack of awareness of people having differences at all, and not just the mere presence of people who diverge from the most common

experiences, which is a particularly pertinent subset, though future work may look at other situations. This is fairly simple to implement; agents with the “differing” tag were given a set of probabilities of success for each kind of action, with no other changes.

What those probabilities are is patently a very important parameter, and can give differences not only in degree, but of kind of effect modeled. In particular, whether action 1 or 2 is the correct action to converge to depends on which value is higher, and whether the probability assigned to action 2 is closer to 54% or 46% will affect how likely they are to actually converge to action 2. So a situation in which the hidden actual probabilities are 0.8 and 0.51 will mean it is much better for differing agents to choose action 1, but if they are getting information from others with the same spread⁶ they will be more likely to end up choosing action 2, since results with action 1 do not change any agent’s beliefs, and successes in action 2, more likely than failures in it, shift probabilities in favour of action 2. After that, however, the actual numbers do not make a difference. We need there to be a gulf between expectations and reality so that the effect will be noticeable, and we need none of the probabilities to be too close to 1 or 0 to allow for enough variance in the data for other effects to happen as well in mixed models. Therefore, I somewhat arbitrarily tested both [0.5, 0.33] and [0.4, 0.66] to model, on the one hand, what happens when the differing group is particularly unlikely to succeed given the riskier option, and on the other hand, when the supposedly safe option is unsafe, but the supposedly risky option is highly effective for the differing group. Changing the 0.5 to 0.4 in option 1 does not actually have an effect on how learning occurs since, again, neither result of the first kind of action changes an agent’s probabilities, but this change does help illustrate situations where unmedicated people with ADHD can struggle to keep up when using conventional methods. In other words, there are two kinds of difference model here: one where agents with the “differing”

⁶Notice that for any probability of success of action 2 less than 1, assuming an even distribution of starting priors between favouring 1 and 2, some percentage of agents acting completely alone will stay with 1 immediately because they initially favour it and will never try 2, and of the other 50%, a non-zero percentage will eventually choose to try 1 and never stop—indeed, with infinite time, all of them will eventually have such a poor run of luck they will get stuck on 1. This learning problem does not work well outside of communities of learners, and indeed it is impressive that convergence to 2 is possible and so likely given this.

tag do not succeed with action 2 even though the majority do, and one where agents with that tag succeed even more with action 2 than the other agents do.

2.1.3 Results

The code outputs information in graphical, textual, and tabular form, and I have used these data to explore the effects of the above additions to a basic network learning model. To begin, I will describe how the basic version runs, using one of the thousand network starting setups generated for the project, Network 1⁷.

See Figure 2.1 for the starting configuration of Network 1 at base. Of particular note is that Network 1 has a body of highly interconnected agents with agents 0 and 1 connected to each other, which is relevant because these are the two agents that will be singled out in several models⁸. As well, agent 8 is unique in only having a single connection to this interconnected body, that connection being to agent 7, which does not connect directly to either 0 or 1.

The story of Network 1 with base model type is this: within 30 steps, all but agent 7 are reduced to incorrectly believing that action 1—going without coffee in our running example—is superior, but agent 7 remains steadfast due to itself getting fortunate results. It then takes about 20 steps to convince agent 8, and then most of its neighbours in another 10, that action 2 really is better. The result of this insistence is that by step 110 the entire network is taking action 2, with agent 2 going back to yellow for 30 steps after that point, and then no other perturbations. I would explain the general thrust of this by noting that agent 7 is connected to a large number of other agents who start believing in action 2, but

⁷Incidentally, Network 1 is the second of the thousand networks due to the data structure of Python, which uses 0 for the index of the first item, as do most programming languages. I use Network 1 throughout because it is somewhat less typical than Network 0 in a way that makes it more susceptible to the changes made to the base model, making it a better model for demonstrating those effects. I also did not want the language “Network 0” to mislead readers less familiar with computer programming into thinking there was something special about the network being discussed. Remember that just because Network 1 is behaving in a certain way does not mean *most* networks will do so, nor does it even mean the first randomly chosen network did so.

⁸Or doubled out, if you prefer.

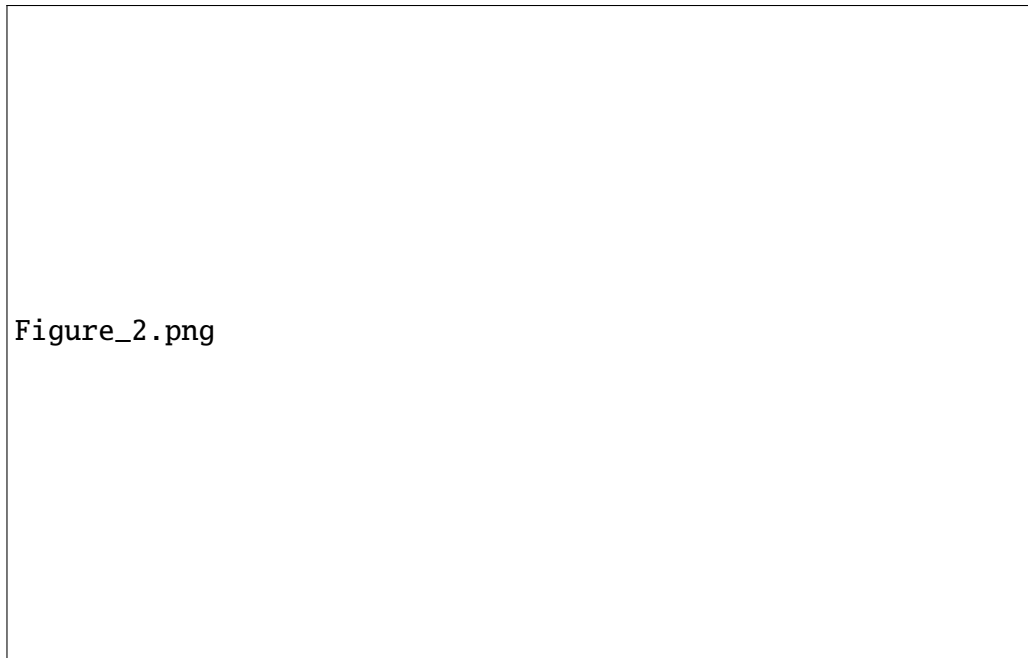


Figure 2.1: Network 1's starting arrangement and prior probability that action 2 is better. The agent's ID number is above its probability for action 2.

not to agent 0, which has a rather unfortunate run of getting the less likely failure result from its actions quite often early on, which is enough to turn itself and its neighbour 4. Since 3 and 6 start with barely favourable opinions of action 2, they falter easily. And since agent 8 only gets information from 7, not even generating information itself, it will follow what 7 does.

In this case this is a fairly standard story. See table 2.1 for a summary of relevant statistics for some of the simpler model types. Some networks do not have a flash of almost going to prefer action 1, and some do not have an agent acting like 7 to hold back this flash. But most often, something like some agents getting less lucky but other agents getting more typical results will occur, and the typical results will ultimately beat out the other results. Statistically, 86.4% of base networks converged to preferring action 2, drinking coffee, with 91.82% of agents overall preferring action 2. When nothing impedes learning, network learning models are extremely effective! On the other hand, 7.4% of networks converged to full adoption of action 1, not drinking coffee, which is a fortunately small

amount, leaving 6.2% failing to converge after ten thousand steps. It is noteworthy that networks that do not uniformly take the correct action are more likely to uniformly take the *incorrect* action than to stay mixed. This is not because of some mysterious force of convergence no matter what, rather it is better explained by noting that so long as some agent, like agent 7 in our example, avoids falling into the wrong action, it is extremely likely to turn every single other agent to action 2 eventually, but the opposite is not true of holdouts for action 1, which does not generate data. The mere existence of a network and a correct answer is enough to make the likelihood of convergence to that correct answer very high all else being equal.

	% Con	% 1-Con	% 2-Con	% Agree	% 2-Pref
Base	93.8	7.4	86.4	99.3	91.82
Bias	91.5	7.1	84.4	99	91.82
Language 2	80.7	3.8	76.9	97.67	93.71
Language 5	83.0	2.6	80.4	98.02	98.02
Difference	24.4	22.6	1.8	78.15	52.91
Mingle	47.1	7.5	39.6	84.74	72.86

Table 2.1: Some statistics for different models with one or zero additional features, from their 10 000th steps. In order: “% Con” refers to the percentage of networks that converged, broken into “% 1-Con” and “% 2-Con” for whether the network converged to action 1 or 2, respectively, “% Agree” is the percentage of agents across each network that agrees with that network’s most common preference, and “% 2-Pref” is the percentage of agents across all networks that had a preference for action 2.

I now want to look at one of our types of models in isolation, in this case the difference model, or “d2,” named for the first two agents, 0 and 1, having different payouts than the others. In particular, Network 1 has a very similar story with this type of model, with a couple differences. First, agent 0’s results cannot really be described as unfortunate anymore, as coffee, or action 2, really is bad for it. It gets even more 1-favouring results at the start, which leads to much the same starting scenario. 7 is still able to remain strong, still flips 8, and still pushes through the body of the network. However, 0 and 1 are never permanently convinced. Whenever agent 0 ends up preferring action 2, it fails quite quickly, pushing it back below 50% probability in action 2 being better. As a result, it ends up with incredibly

low probability in action 2, which will make it exceptionally hard to get back above 50% ever again. Agent 1 is more moderate, going up and down often and sometimes staying over 50% probability in action 2 being better, but ultimately being brought back down time and again. So the network never converges. See figure 2.2 for step 1000 of this model, showing a typical state of this ever-fluctuating model.

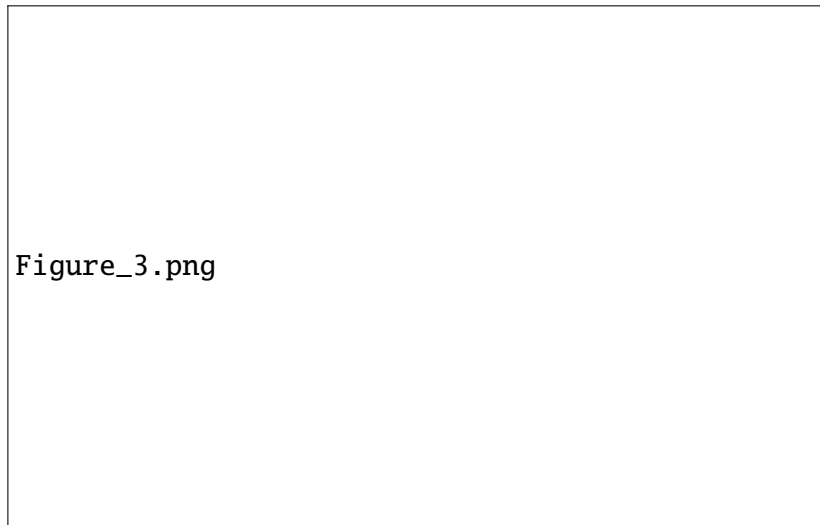


Figure 2.2: Network 1 with the difference model type at step 1000. Because agents 0 and 1 do better with action 1 while the others do better with action 2, beliefs end up differing greatly. Probabilities are rounded, no agent can have an actual probability of 1 or 0.

Once again this story is typical of difference-type models. Only 1.8% of networks converge to action 2 when agents 0 and 1 actually do better with action 1, although 22.6% of networks end up converging to action 1. Remember that when a network converges to action 1 it can never move in beliefs because action 1 generates useless data, so it is not surprising that it becomes more common in this situation, despite it being against the interest of the majority of agents. Networks frequently will have a moment of entirely favouring action 2, but because this situation allows for agents to change their mind, these moments do not result in long-term convergence. Note, however, that the mean percentage of agents who agree with the majority opinion in any given network at step ten thousand is 78.15%, which indicates that, as seen in Network 1, the majority all converging to action 2 while the two divergent agents both prefer action 1 is a very common outcome. In other

words, the network is continuing to work well! Agents mostly converge to the answer that is correct *for that agent*, even when nearby agents are giving them information that runs counter to their own experiences.

To summarize so far, whether or not agents have the same interests, information sharing networks facilitated realization of agents' interests for themselves a solid majority of the time, with the nature of the learning problem itself leading to the minority being favoured. Mere difference between agents is hardly a confounding factor. I now turn to confounding factors.

There were two types of language models. In one type, "15," half of the network spoke one language and the other half another. In the other, "12," agents 0 and 1 had their own language and the rest of the network another. Although the total number of misunderstandings would be expected to be higher in 15 models⁹, they converged more often. In particular, 76.9% of 12 networks converged to action 2, compared to 80.4% of 15.

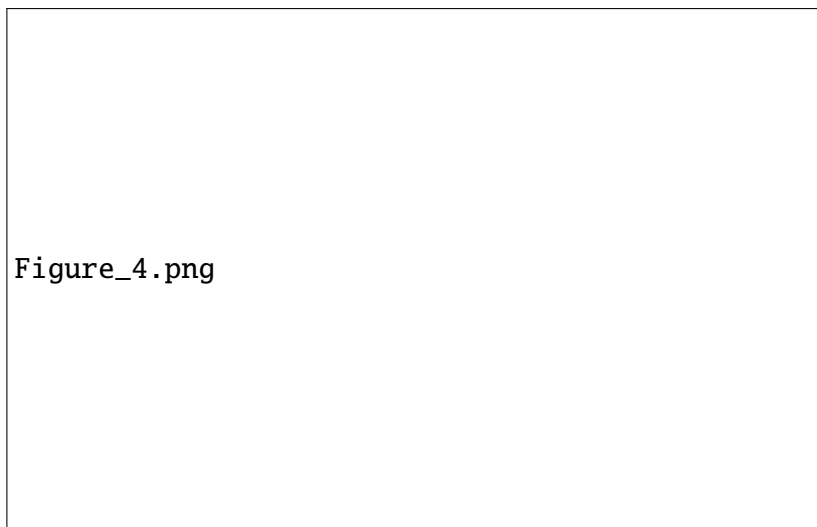


Figure 2.3: Network 9 at start. Network 9 is interesting because it is the first network that converges to action 2 when split 50/50 across languages, but not when split 80/20.

To see why, consider Network 9, which converged in the 15 model but not 12. The net-

⁹Connection probability is even across agents, so in 15 each agent has a 5/9 chance of any particular connection being with an agent with another language, for an average of 55.56% chance of a connection being cross-language. In 12, eight agents have a 2/9 chance per connection, and two have an 8/9 chance, for an average probability of 35.56%.

work is illustrated at start in figure 2.3. The structure of the network itself is unremarkable, and no structural reason jumps out as to why it would act differently depending on whether there is a 50/50 split or 80/20 split on the two languages, but the difference becomes clear when looking at the history of the network. See Figure 2.4 to follow along visually.

At step 30, Network 9 entirely favours action 2, the actual best action for all agents, regardless of language split. However, also in both models, some agents get unfavourable results, and begin to favour action 1 again. In the l5 model these are agents 0, 1, and 7, and in the l2 model these are agents 1 and 7. In l5, both of these agents have a 50/50 split between neighbours having the same language and different languages (in l5 agents 5 and below speak one language and agents 6 and above the other). In the l2 model however, 0 and 1 have neighbours speaking only the other language, and 6 has neighbours that speak only its language. Since language is the only difference it has to explain why these agents, seeing the same results, turned differently. The sum effect of misunderstandings from its neighbour agent 4 resulted in agent 6 being much more moderate in its probabilities by step 30, with 87.19% probability in action 2 being better in the l2 model where it does not misunderstand its neighbours, and 68.90% probability in action 2 being better in l5 where it can misunderstand agent 4. Agent 0 is more surprising, however. It, too, had drastically different probabilities at step 30 of the two models, but opposite what one would expect. In the l2 model, where all of its neighbours spoke a different language from it, it had 92.00% probability in action 2 being better, but in l5, where all but one spoke its language, it had 58.85%. This too has a clear explanation—in many models, a small number of agents will witness unlikely results that point away from the truth. Agent 0 is in a position not quite that extreme, in that it started with relatively low probability in action 2, 35.72%, and it witnessed only weakly supportive evidence for action 2 being better. The difference is that in a somewhat unlikely turn of events, the specific observations which it misunderstood in l2 were more often observations that would have led agent 0 away from action 2 normally. This explanation involves two steps that are individually unlikely for a single agent, but

remember this is one agent chosen after the fact out of ten agents, from the ninth possible network chosen after the fact. This type of occurrence happens far more than 1.11% of the time, so we should not be surprised by it in context. At any rate it going in this direction goes against the overall explanation I will pose.

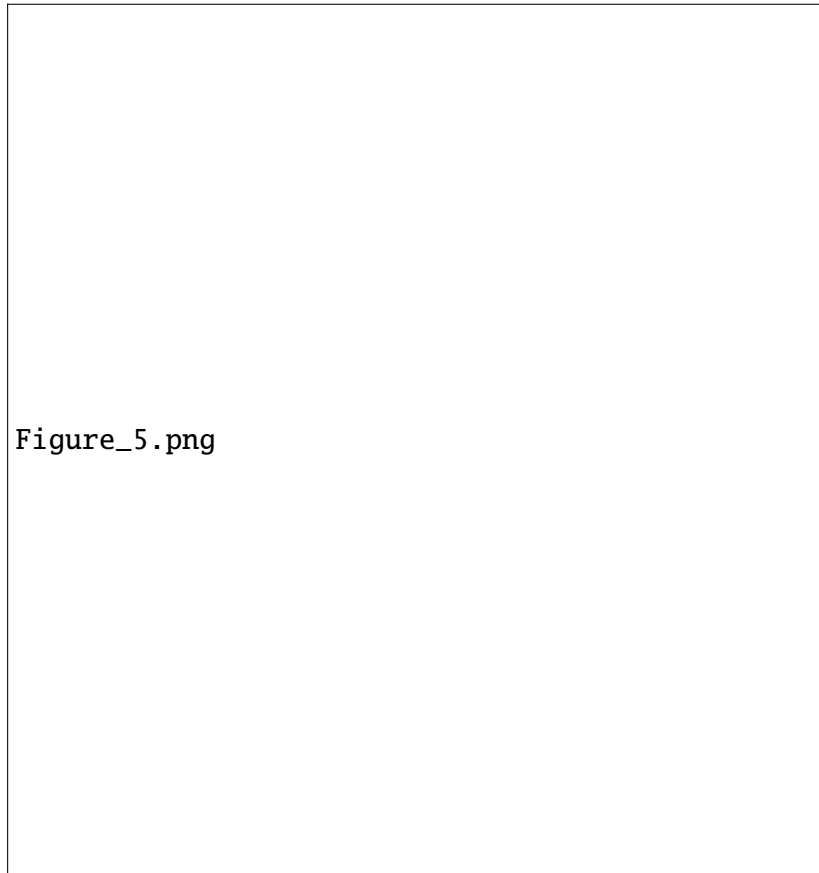


Figure 2.4: Network 9 side-by-side of 12 and 15 at steps 30, 40, and 380. Note in particular that agent 1 is able to recover in 15 but not in 12.

The more interesting thing is what happens over the next several hundred steps. In 15, each agent that flipped had some neighbours that it understood correctly, and its own trustworthy results, allowing it to eventually even out the events that turned them to action 1 from steps 30 to 40. In 12, this is true of agent 7, but not agent 1. So out of agents 0, 1, and 6 in 15, and agent 7 in 12, all were in a position where it was at least a little bit more likely for any given information they saw to hint towards action 2 than action 1. Therefore, over hundreds of steps, with other agents quite securely preferring action 2, they eventually

all turned to action 2, as was in the long run overwhelmingly likely. Agent 1, however, did not. Indeed, once preferring action 1, recall that it is not generating useful data in either direction, and that since all other information it receives passes through a filter of being exactly flipped 50% of the time, it is receiving nothing but junk information, equally likely to confirm either hypothesis. Then at any given step, over the next thousand steps, the probability that information it receives confirms action 2 more is equal to the probability that that information confirms action 1 more. By step 1000, then, it is worse than a coin flip as to whether the agent will end up in favour of action 2, since it not only has to be more confirming of action 2 than 1, it has to be more confirming to a strong enough degree to move it up enough to pass the 50% belief line. While true that once past this line it has its own information, which, no matter how much junk information it deals with, does bias it in favour of the correct hypothesis, if agent 1 never passes this line, it will never become more likely than not to increase. Thus, it is unsurprising that this network does not converge by step 1000 in l2, even though it does in l5. The difference is that in l5 no agent is entirely an island, but in l2, agents 0 and 1 effectively are, so they are very susceptible to sticking to action 1 once there. The far higher chance of this happening in l2 accounts for the lower rate of convergence in l2 models than l5.

You may find this explanation unsatisfying—I said much about things being *unsurprising in context* and *only a little bit unlikely*. Statistical explanations usually rely on things being more likely than not, or else they can hardly be said to explain much. Recall, though, that what is being explained is a 3.5 percentage point gap in rate of convergence, both models converging more than three-quarters of the time. To explain why a likely occurrence happens slightly less often in one situation than another, while staying likely, the thing to point to is that an unlikely occurrence is only slightly less likely in one situation. Network 9 was chosen because it is atypical, but it is a good example of the atypical networks that change their result.

Like language, bias had a relatively small effect. With moderate bias, coefficient 0.75,

84.4% of networks converge to action 2, down only 2 percentage points. With high bias, coefficient 1, this goes down to 80.9%, a total of 5.5 percentage points. Even with this change, percentage across all agents of preference for action 2 is unchanged to two decimal places with a 0.75 bias coefficient, and changes by only 0.62 percentage points with fairly strong bias, indicating that while in some networks bias does result failure to converge, in non-converging networks it has the opposite effect, increasing the number of agents that come to the right conclusion within those networks. Number of networks converging to action 1 also does not change much, at 7.1% with coefficient 0.75 and 6.9% with coefficient 1. In short, bias alone has little noteworthy effect on models.

It is worth noting that my results for bias do not look like Jungyi Wu's for a very similar model. Wu's model is formally identical to mine, meaning the defined question and how bias is implemented are the same. In analysis, however, our goals differed, and therefore so does the information we are reporting. Specifically, Wu is most interested in equilibrium, and is trying to argue that in the long term, the view from standpoint epistemology that marginalized agents can have epistemic advantages is borne out. For her scope, the argument works. The main difference unidirectional bias introduces at the equilibrium level is that a new possible equilibrium appears, and so the only effect it can have is that the space of models that reach the other equilibria are reduced as the space that reach the new equilibrium takes that space up. This is what Wu reports [26]. I am looking at models at various snapshot steps in order to capture behaviour on the way to equilibrium and for networks that do not reach an equilibrium for some set of parameters. At this level, bias seems to confound the process of reaching equilibrium in general, but only a small amount. Our results are not in conflict, and reflect different facts about the model, and likely different opinions about what is important about a model of this kind.

What is interesting is mixing language and bias. See table 2.2 for statistics involving these models. In "b2l2" models, in which agents 0 and 1 are both biased against by and speak a different language than the other agents, convergence to action 2 occurred only

	% Con	% 1-Con	% 2-Con	% 2-Pref	% 2-Dif
l2	80.7	3.8	76.9	93.71	n/a
b2l2	73	4.1	68.9	90.77	n/a
l2d2	9	9	0	71.35	1
b2l2d2	9.5	9.5	0	70.63	1

Table 2.2: Some statistics comparing the model types with language but not mingling, from their 10 000th steps. In order: “% Con” refers to the percentage of networks that converged, broken into “% 1-Con” and “% 2-Con” for whether the network converged to action 1 or 2, respectively, “% 2-Pref” is the percentage of agents across all networks that had a preference for action 2, and “% 2-Dif” is percentage of divergent agents that had a preference for action 2. The row titles are shorthand for model types, again in order, Language, Bias and Language, Language and Difference, and all three. The “2” in the shorthand refers to the number of tagged agents.

68.9% of the time, which is a larger than additive decrease compared to just language and just bias. Comparing Network 1 between l2 and b2l2 paints an interesting picture. The l2 model for Network 1 is very similar to the base model, though instead of agent 7 holding the line and converting the others, it is agent 0 doing so. This is a bit surprising given that agent 0 has only a 50% chance of giving favourable information to neighbours other than 1, but it still makes sense. Network 1 has a wave fairly early on of unlikely results across many agents, but agent 0 is not one such agent. Because agent 0 is insulated by random noise from most of the unlikely results, it does not begin to favour action 1, and likewise, because agent 4 is insulated from agent 0’s more likely results, it favours action 1 more quickly, which causes agent 7 to miss out on favourable results that kept it from beginning to favour action 2. The effect of having agent 0 give essentially random noise to neighbours is that its neighbours are receiving noise instead of nothing, and if the noise randomly favours action 2 enough, it can move those neighbours into preferring action 2 again and producing good communication of good data. In this case, at step 75 agent 0’s random noise accumulates into having agent 4 favour action 2 again, and it takes only five more steps from there for two more agents to join as a result. See figure 2.5 for an illustration of this. While this is going on, agent 0 is slowly working on agent 1 as well, since the two are connected without any bias or language barrier, and so by step 420 the

network entirely converges permanently back to favouring action 2.

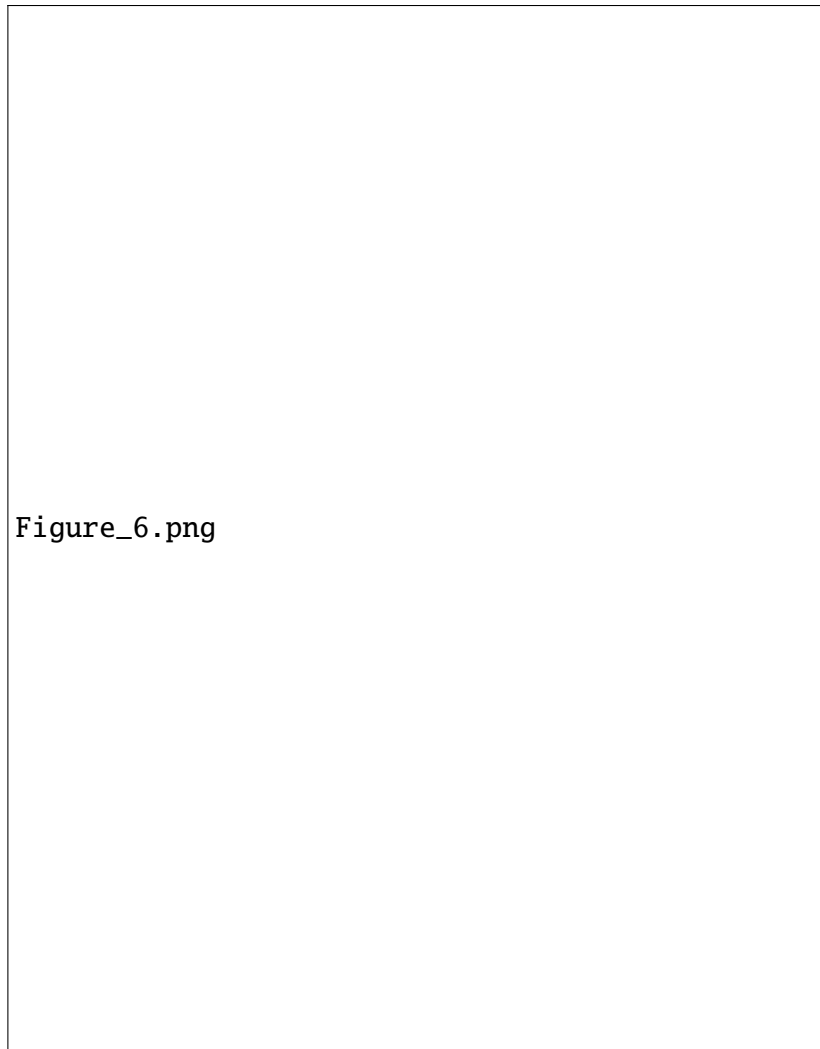


Figure 2.5: Network 1 at steps 75 and 80 of the I2 model. Step 75 has agent 0 surprisingly convert agent 4, which converts two more in five more steps.

The difference between the above story with model type I2 and model type b2I2 is that the communication from agent 0 to agent 4 is far more strained. Jeffrey's rule decreases the effect of the results randomly perceived to be in favour of action 2 significantly, and agent 4 never gets to that point, step 75 in I2, where it gets slightly over 50% in favour of action 2 again. As a direct result of this bias, the network is unable to converge.

When increasing the bias coefficient to 1, convergence goes down a small amount, from 68.9% to 67.1%. Given the above story, this shouldn't be incredibly surprising, as it is the

mere presence of bias that significantly decreases the possibility for events like those seen in Network 1 to occur. Once the two effects are co-occurring, their relative strengths seem to matter less.

Furthermore, bias seems to matter less the more other features get added. The model type combining languages and divergent agents, “l2d2,” never converges to action 2, with 9.0% of networks converging to action 1, 71.35% of agents in general preferring action 2, and only 1% of divergent agents preferring action 2. That 71.35% number is most interesting for models with divergent agents, the closer to 80% the closer to ideal. Introducing bias to this type of model, or looking at “b2l2d2” models, changes the numbers fairly little. 9.5% of models converge, none of them to action 2, with 70.63% of agents preferring action 2 and only 1% of divergent agents doing so. Stronger bias brings this to 10.1% converging to action 1 and 70.23% preferring action 2 overall. These are small changes compared to the combined effect of language and divergence. Comparatively, b2d2’s numbers—3.7% converging to action 2, 18.9% to action 1, 60.07% preference of 2, and 24% of divergent agents preferring 2—are quite different. So are b2l2’s numbers, 68.9% converging to 2, 4.1% to 1, 90.77% preferring 2, and no divergent agents to measure. The numbers indicate that bias is the least important factor once the three factors are combined, and a run through Network 1 makes this look accurate—I found it hard to tell what difference bias was making if at all, down to both versions converging at the same step.

Finally, there were models in which agents in a network were able to change their connections based on nearby agents with similar probabilities, an activity I refer to as “mingling.” Statistics for different model types with mingling are given in table 2.3. Adding mingling generally either introduces the possibility of disjoint networks forming that make convergence far less likely, or the possibility of breaking up choke points and homogenizing the network as it stays mostly together. For most combinations of factors both possibilities existed, but in different relevant amounts depending on the factors. I will give examples of both happening, after discussing the statistical results of adding mingling.

	% Con	% 1-Con	% 2-Con	% Agree	% 2-Pref
m	47.1	7.5	39.6	84.74	72.86
mb2	45.9	6.4	39.5	84.23	71.97
ml2	23.4	4.5	18.9	76.92	65.3
md2	31.2	31.2	0	77.72	30.5
mb2l2d2	10.9	10.9	0	73.26	54.98

Table 2.3: Some statistics comparing model types with mingling, from their 10 000th steps. In order: “% Con” refers to the percentage of networks that converged, broken into “% 1-Con” and “% 2-Con” for whether the network converged to action 1 or 2, respectively, “% 2-Pref” is the percentage of agents across all networks that had a preference for action 2, and “% 2-Dif” is percentage of divergent agents that had a preference for action 2. The row titles are shorthand for model types, again in order, Mingle, Mingle and Bias, Mingle and Language, Mingle and Difference, and all four. The “2” in the shorthand refers to the number of tagged agents.

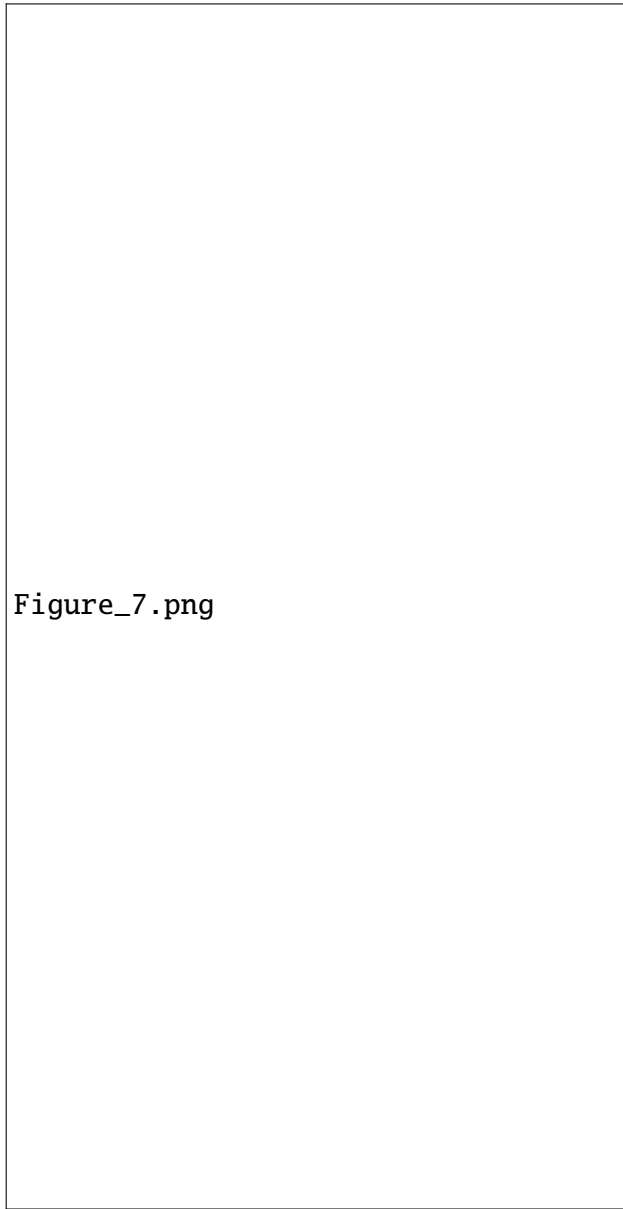
A mingling-only model, or model type “m,” since it does not need a parameter for which agents are affected, converges to action 2 only 39.6% of the time, and to action 1 7.5%. 72.86% of agents overall in these models prefer action 2 at step 10000. In this way mingling is more like divergence than language or bias, in that it is a change to the model itself so drastic it alters what numbers to expect and how to interpret them, instead of just causing some kind of impediment. Adding bias, “mb2,” does not result in very different statistics, with 39.5% converging to action 2, 6.4% to action 1, and 71.97% of agents preferring action 2. Increasing bias coefficient changes these a small amount, to 38.5%, 7.5%, and 71.54%, respectively, all negligible differences. Bias appears to matter fairly little once mingling is in effect, in other words. Language on the other hand, in “ml2,” moves the numbers down to 18.9% converging to action 2, 4.5% to action 1, and 65.3% preferring action 2. Finally, mingling with divergent agents, “md2,” leads to 0.0% converging to action 2, a whopping 31.2% converging to action 1, and 30.5% overall preferring action 2, 0% of which are divergent agents¹⁰. This latter is perhaps the most interesting—without combining mingling and divergence, the highest rate of convergence to action 1 is divergence-only,

¹⁰I use the language “0.0%” and “0%” because of the significant figures my experiment gives me. To say “none” would sound like it becomes logically impossible for convergence to occur, which is untrue. I do not have a guess as to how many networks one would have to randomly run through this model to get a divergent agent to prefer action 2 at 10000 steps, let alone for a model to converge to action 2, but there does exist a combination of random number generator results that would give these results, however unlikely.

with 22.6%. The model, then, is still acting largely like a divergence model, but in a subjective way, even more like it. The divergent agents seem to have even more of an effect. I will explore this after the more common case.

Network 1 gives a good example of what mingling does when all factors are present, or in model type shorthand, “mb2l2d2.” See Figure 2.6 to follow along. At step 47, all but agents 4 and 7 are fairly convinced that action 1 is best, while the two dissenters hold just as strong a conviction the other way. As a result, agent 4 has lost all but one connection, as other agents near to it in the network prefer not to communicate with it. In the move to step 48, agent 4 loses its only remaining connection, causing it to connect to the agent closest to it in probability (to avoid an end result of 10 disconnected agents, since connections can be made only to neighbours-of-neighbours). This of course links agents 7 and 4. At the same time, agent 7 loses one of the two connections it had previously. Then, from step 48 to step 49, agent 7’s other connection, to agent 2, is severed, but because it is connected to agent 4, it does not try to make another connection. The result is two permanently disjoint networks. 4 and 7 will always have very close probabilities since they see the exact same information, and even if they come to prefer action 1, they will not shift far below the 50% line, and since the other network are all unchanging already, this means they will always be closer to each other than anything in the other network, and vice versa. Incidentally, the model does create a second disjoint community with agents 5 and 6, who are also closer to each other than any other agent, that will stay this way, but by this point both agents are stagnating, so this does not particularly matter. The important thing is that because agents 4 and 7 are quite unlikely to both prefer action 1, and no other agent can ever change probabilities, the network as a whole is very likely to never converge. This is a common occurrence when a lot of factors are present.

Next we will look at Network 6 with divergent agents, both with and without mingling, to see why the combination has an exacerbating effect. See Figure 2.7 for an illustration of Network 6 without mingling in its equilibrium state. As often happens, a somewhat isolated



Figure_7.png

Figure 2.6: Network 1 with all factors from steps 47-49. Mingling causes 4 to become alone and then latch to 7 from step 47 to 48, and then for 7 to become detached at step 49, isolating the two dissenters.

agent does not begin to prefer action 1 as the others do, and starts converting nearby agents over time by being more likely to succeed than fail. In this case that agent is agent 7, which is connected only to agent 5. Once agent 5 turns, it begins convincing its other neighbours agents 2 and 0. It is able to convince both, and both connected to agent 9. However, since 2 has no other connections, this causes a problem. Agent 9 is now getting information from agents 2 and 0 and nobody else, which means it is getting conflicting information. Because agent 0 gives unfavourable results at a higher rate than agent 2 gives favourable results as long as agent 0 is taking action 2, agent 9 is getting deeply conflicting information. But agent 5 has three positive sources for action 2 against agent 0's negative source, so it stays preferring action 5, and as a result is able to keep flipping agent 0 back to preferring action 2 after it convinces itself to stop. The result is that agent 0 acts as a dam, holding back the spread of preference for action 2, since all of its neighbours see an agent that keeps being convinced to try action 2, and then regretting it. Agent 0's neighbours then have their probability for action 2 driven very far down, and agent 2 alone is not able to break through this stream of information. This sticks the entire network at a non-converging equilibrium, resulting from its unique structure.

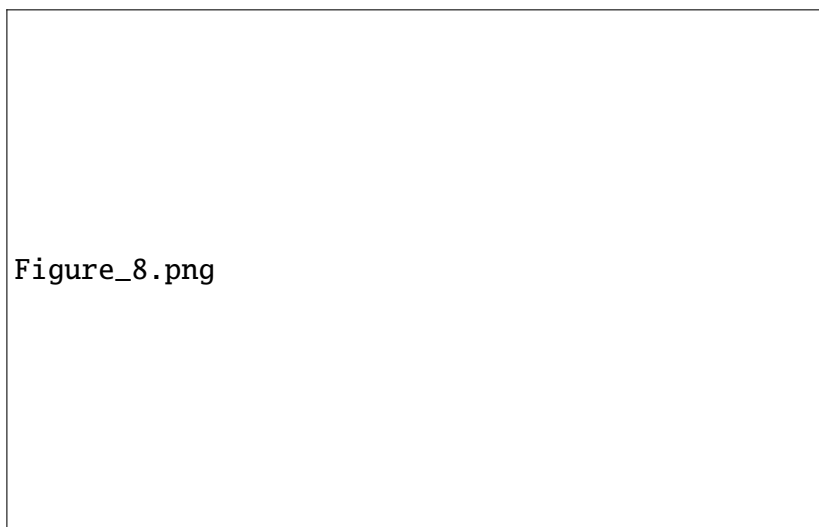
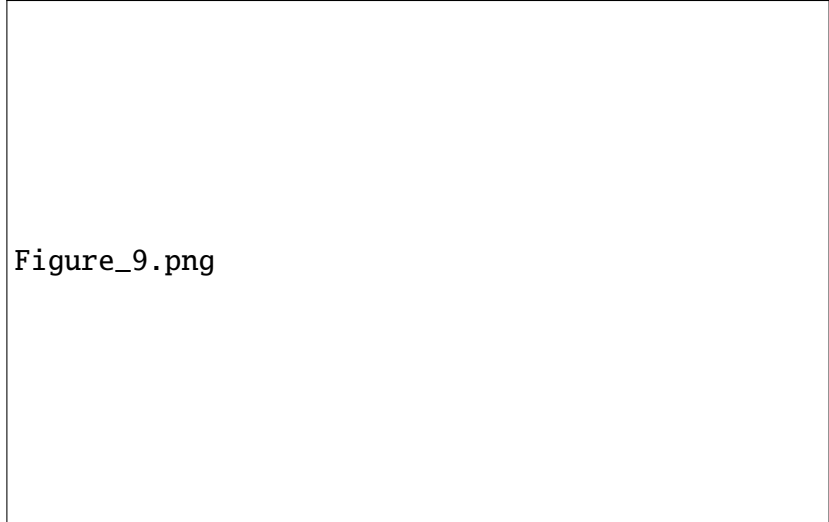


Figure 2.7: Network 6 divergent model type at step 640. Agent 0 blocks progress because as soon as it turns to action 2 again it dissuades others from trying it.

An astute reader may have a guess already as to why introducing mingling would have a different result. Without agent 0 acting as a dam, action 2 does not have its specific impediment, and if agent 0 is commuting all over the place, it should not be able to act as a dam in the same way. Indeed, with mingling, the network does not create a dam effect, but it does not converge to action 2 either. In Network 6, agent 0 always starts with a very high prior probability for preferring action 2, 92.28%, the highest out of the network. As a result, when mingling, high-probability agents seek out agent 0 as like-minded, and low-probability agents leave it. Agent 0 starts with agents 4, 5, and 9 as neighbours, but by step 20 has 4, 6, 7, and 9 (see Figure 2.8 for step 20 of this model). Agents 4, 6, and 7, now its neighbours, were the three next-highest-prior agents, and agent 5 was the lowest, while agent 9 is around the middle. In this arrangement, the other agents are unlikely to mingle away from agent 0, with whom they are in fairly close agreement, and so they are receiving a much steadier stream of information disconfirming the efficacy of action 2. As a result, agent 0 ends up the last agent with any probability in favour of action 2, while last time it was agent 7, which previously was not privy to agent 0's negative information. At this point agent 0 does what it is most likely to do, single-handedly convince itself to stop trying action 2, and the entire network has converged to action 1.

In summary, agent 0 has a profound effect in both versions of Network 6 when it has diverging interests from the rest of the network. When its location is static, agent 0 is able to stop preference for action 2 from moving from one part of the network to another. When moving, agent 0 is able to quickly collect the biggest supporters of action 2, and slowly convince them to stop supporting it. Even though the network never splits, mingling still has a destructive effect on the network at large, allowing agent 0 to influence it even more than it already was. Of course the opposite is also possible, and mingling can homogenize a network in the direction of action 2, as well—mingling is a mixed bag.



Figure_9.png

Figure 2.8: Network 6 mingling divergent model at step 20. A high-prior agent 0 is connected to the most fervent supporters of action 2, and is dissuading them from continuing due to its individual low likelihood of success with action 2.

2.2 Appraising Network Models of Epistemic Injustice

The network models above tell us a lot about how networks of computational agents are able to learn with certain additional factors added. More must be said to clarify what they tell us about epistemic injustice, however.

2.2.1 Discussing the Results

I will first summarize the results above. Then, before moving on to discussing what recommendations or other takeaways we might glean from these results, I will briefly discuss the importance of keeping in mind the divide between these models and reality.

Mingling has the biggest effect out of all features on how a model can play out. In many of the example networks, a particular feature of how the agents connect to each other is what leads to the eventual equilibrium. When agents can change their networks, this changes, and primarily in two specific ways. Either communities become totally and permanently disjoint, with dissenters being completely separated from the majority, or some agents are able to become even more influential and change the views of far more agents over time.

This homogenizing effect primarily occurs when there are divergent agents but not other factors.

Divergent agents were not only usually able to come to their correct conclusion without any help, adding language barriers and mingling made them more likely to do so by isolating them more. While biasing the other agents against divergent agents did reduce how likely the divergent agents were to come to their correct conclusion as expected, this effect size was comparatively small.

Finally, both bias and splitting into languages to isolate two agents had an exacerbating effect on other features, including each other, but were less important on their own. Between the two, language consistently was the more relevant feature.

Just from this there are some conclusions one may want to draw, but defer that for a moment. On what grounds would one be asserting that any of the above holds true in the real world? In a model, once other things are already happening, biasing agents against each other has a surprisingly small effect. This does not mean that in the far more complex real world, bias is even less likely to be important. Recall also that bias interacted in interesting ways with different factors individually. By adding more factors to get closer to reality, bias may find something else to exacerbate.

Perhaps more importantly, real people are not machines trying to solve learning problems. They lead lives coloured by emotion, sensation, and thought. As Fricker writes, the most cutting harm of epistemic injustice is not in not contributing to knowledge or knowing less in general, it is the decreased capacity for self-actualization that comes with exclusion from seemingly primarily epistemic activities [6]. If the model shows something like “identity-based bias does not hinder learning significantly,” it would be an unwarranted leap to even say, conservatively, “if all the mechanics of these models map perfectly onto the real world, bias is not a very harmful thing.” Systematic identity bias can be harmful in ways language barriers, absent bias, would not be, beyond purely epistemic harms. In short, any suggestions about what is best for information flow are not suggestions about

what is best simpliciter.

There is still a point to this kind of experiment. If nothing else, a feature like bias has *at least* as much effect as the experiment indicates. If nothing else, other features have *the potential* to individually be exacerbated by or dwarf a feature like bias. And, if you are keen on probabilistic reasoning, that communities of learners may behave in a way close to what is described in the model is *more subjectively likely* having seen it in a model than it was before. Just as you would not declare, “I have flipped this coin thrice and it has landed heads each time, surely it cannot land tails!” do not take the experimentation here as complete. As a few results of “heads” does not warrant such strong statements, but may warrant future investigation into the coin, these results do not prove anything about human community, but may warrant new lines of in-person study.

So, keeping this in mind, here is what little *is* indicated.

For starters, from the summary above, we should start paying attention to overall structures rather than individual biases. If someone is not in a conversation, it does not matter what their communication style is, or how well-received they are. As mingling is so impactful, we might expect that removal of people from conversations is likewise particularly impactful. While having the lived experience of microaggressions from those biased against you is hurtful personally, your words were still spoken, and there was a possibility for minimal impact. When not present at all, you cannot have an impact, and the conversation as a whole suffers even more. As long as our goal is effecting cultural change through conversation, structuring that conversation to centre marginalized viewpoints is easiest to recommend. This must be said with awareness that there are other important goals, of course. My purpose is not to say what values should be our priority, only what our priorities should be given certain values.

Next, recall the effects of languages. First, an 80/20 language split was more of a problem in the model than a 50/50 split. Second, whenever adding an 80/20 language split interacted with other factors, that interaction came from an increased isolation of the two

agents with the less common language. For example, bias mattered more when two of the agents are ill-understood, because the two ways of stopping agents from communicating exacerbated each other. The conclusion is that throughout the models, the strongest effect language differences had was in sometimes isolating agents significantly more than they otherwise would have been. The suggestion to real life, then, is that whatever other effects neurotypical inability to understand typical autistic communication styles has, among those effects *at least* is an isolating of an autistic person's ideas to their own head if they do not have someone close to them who shares their communication style. This seems likely. Even if a neurotypical person is not ableist, and holds in their heart a willingness to hear out neurodivergent people, if there is significant room for error, then it is likely that, especially when concepts become more complex, they will not fully understand what the neurodivergent person is saying.

This has a number of implications. For one thing, it suggests looking into neurotypical research into divergent neurotypes. Do the researchers actually understand the communication styles of their subjects, knowing that those styles often differ from their own? It may be the case that they do, but if not, no matter how well-meaning and well-researched, any conclusions requiring communication from the subjects may be suffering from misunderstandings. This is one more reason to support the disability rights slogan, "nothing about us without us," in the context of science. On the other hand, it suggests a strategy for improving the lives of neurodivergent people. We need *some* people close to us to be fluent in our native communication styles. Having one ally who is thoroughly versed in differences between neurotypical and neurodivergent styles of communication should allow for freer passage of ideas between neurotypes. We have seen that it is the complete lack of connection that causes the problems, as when an agent has a mixed neighbourhood, information flows much more freely, and the problems are significantly lessened. At least one part of the difficulty caused by difference in communication style, then, can be addressed by even minimal instruction in communication styles.

Both implications can be addressed to a large extent simply by placing neurodivergent people in more prominent positions in communities related to neurodivergence. Neurodivergent spokespeople who are capable of speaking in a way neurotypical people will be more capable of understanding, or even just communities that include larger proportions of neurodivergent people to allow communication between them, can lead to more coherent and forceful communication with neurotypical people. Placing them on research teams about neurodivergence would also go a long way to eliminate the problem, among other benefits. While suggestions need to be tempered with the understanding that these models have minimal applicability, when the suggestions pay attention to that by noting only that the models give us a minimum plausible effect, and mirror suggestions already made by people in the affected communities, they can be made with some confidence nonetheless.

I next want to address Wu's conclusions from her very similar bias model. From facts about the equilibria of her models, Wu concludes that being biased against is in some sense an epistemic advantage. The facts I present about the speed and likelihood of convergence for my model seem to indicate that it is in some sense a small epistemic disadvantage. Both are true. Bias slightly disadvantages an agent by making it unable to influence its environment when it has useful information, and also creates a possibility for an advantageous position of being uniquely positioned to hold certain information. Which of these is more important will depend on a number of factors. If there is something about information that makes it more available to one group of people, lack of flow of information out of that group does increase the likelihood that primarily members of the group will learn that information. Wu's example of racist hiring practices seems to be one such example [26]. Zooming out, this is not an advantage in real terms; it is better for marginalized racial groups if everyone understands that hiring practices are racist than if only people harmed by it are aware of the fact. On the other hand, other kinds of information are more likely to be affected by the effect I observe, that slowing rates of communication slows learning in general, albeit to a very small degree. I find Wu's result more interesting when looking

at bias in a vacuum, since it is counterintuitive but has much greater empirically sensitive robustness, and since it resists the possibility of epistemic objectification by showing that natural processes will work *in favour* of marginalized groups, and therefore observed epistemic injustices require a more intentional explanation.

That said, when looking at other aspects of my model, there is a result I find even more compelling. While bias against divergent agents made them less likely to discover the best action for themselves, a language split made them more likely to. The reason for this is that, essentially, when totally isolated, it is more likely that a divergent agent will move towards action 1 than 2, and of course, when favouring action 1, there is no chance of going back. An 80/20 language split gets divergent agents closer to isolation, so it makes them more likely to favour action 1. On the other hand, bias against divergent agents does nothing to the flow of information from the majority of agents to the divergent ones, so that isolating effect is not present. What is present is a lack of flow in the other direction, causing the agents *around* a divergent agent to be less likely to be influenced by them into preferring action 1, which would have made them cease to push them more in the direction of action 2. In this way we actually see the agents in the model can reinforce the effect on themselves of their own information by putting it out into the world. An agent whose own evidence points to action 2 moves their neighbourhood toward action 2 by sharing it, and the result is a neighbourhood giving information back that comes from action 2. Information-sharing, even in this formal context, has a self-reflective property, causing agents to solidify their own internal beliefs for themselves by changing the external world. However, when there is bias against an agent, that agent impacts the world less, and sees its own evidence reflected less, thus becoming less solid in it, even worse than if they had simply been isolated. It is worth noting here that this is something Fricker theorized in an informal context as well. Although she did not compare to a counterfactual where one is isolated from outside influence at all, she did stress that implicit bias can take away an important element of self-actualizing, where one hears back their own ideas from someone they have

just shared them with [6]. This fact, it turns out, is not only psychological—the mind’s internalizing process involves receiving our ideas from a seemingly outside perspective—it is also mathematical—the fact of influencing neighbours increases the influence an agent has on itself. This result, too, has empirically sensitive robustness.

Beyond this, the fact that isolation does not have the same effect is particularly interesting. It suggests that at least one element of bias disrupting self-actualization is more active than as described by Fricker. For Fricker, bias disrupts self-actualization by removing a step in the self-actualization process of the biased-against individual [6]. If the effect in my model carries over at all to the real world, however, at least one more thing is happening. When that step was simply removed, agents prospered. This does not happen in the real world, because real people are more complex than mathematical agents, and the psychological fact stated above obtains. However, when the step was removed, while maintaining some contact in the other direction, agents suffered. The suggestion, then, is that the epistemic suffering measured by the model is not born of the mere removal of that step—it is the fact that from there, ideas from the mainstream are still finding their way to the agent where the agent’s own ideas are supposed to be reflected. The suggestion, then, is that whatever else is happening, there is at least some additional effect caused by the existence of a one-way communication stream itself.

Drawing on real experiences of autistic masking and self-isolation, I will explain what this evokes for me. Like many autistic people, I have experienced periods of time where I suffered significantly from not self-actualizing. I was doing an activity called “masking,” where the way I behaved was the way I believed those around me expected me to behave, instead of how I authentically desired to behave. Autistic masking does not only hinder self-actualization by skipping a reflective step in the process, it hinders it by exposing the autistic self more intensely and more often to the neurotypical mainstream, and forces the self to take it seriously and take it upon itself. Via exposure to the neurotypical mainstream, the autistic self becomes, inauthentically, more like it, even as it perceives itself.

Identity-based bias becomes a hegemonic force, allowing mainstream ideas to colonize marginalized minds, invading through pathways intended for the mind to receive a reflection of itself through receptive others. Self-isolation, then, is a coherent countermeasure. The exhaustion of masking aside, when an autistic person isolates themselves from the neurotypical mainstream world, they also insulate themselves from this hegemony. When I have been least publicly myself, I have been most compelled to withdraw into my special interests, devoting thousands of hours to a single video game at a time, for example. And at these times of complete focus on my special interests, I was completely myself. An autistic special interest can be a kind of reflection. When I view my profile in the video game Dota 2 and see my statistics, I see my identity reflected—I see my favourite role, favourite character, favourite outfit for that character, and also I see statistical evidence of my playstyle, focusing more on the team and support factors than is normal for my role. I can see who I am on the screen. The autistic mind, when self-isolating, can construct new ways to self-actualize. It cannot, when immersed in a culture of hegemonic neurotypical identity, defend itself from being consumed by that identity.

Continuing on bias more generally, one interesting observation is that the way in which bias adds complications to models with other features does not depend very much on the strength of bias. As long as agents are being biased against, that fact will interact with the other features of a model in a specific way, and add one more thing blocking their communication, which has the same effect regardless of whether this is on the weaker or stronger end of blocking communication. This may have an analogy to microaggressions and implicit bias. There is a notable minimum effect of adding even implicit bias to a relationship, and this minimum does not notably increase as bias increases. Since there is no maximum given to the effect based on strength of bias, it would be too hasty to claim that this means small amounts of bias are generally just as bad as large amounts in real life. What it instead makes sense to claim is that even small amounts of bias nonetheless matter a lot.

This complicating factor may help communicate the harm of microaggressions. While it is usually obvious to marginalized people that small problematic interactions can have a long-lasting and cumulative grating effect, while also instantly switching one's stance towards someone to a more defensive one, it can nonetheless often be lost on more privileged people why this is worth discussing. This is well-trod ground, so I do not have much to lend towards the discussion of microaggressions and how they function. What I do have is mathematical demonstration of this effect that may hopefully do some work on explaining the importance of the concept. Even in simple game theory, small amounts of demonstrated bias have a large effect—something that looks like a small thing on the surface can turn out to have a very large overall effect.

Handily, there is an overall thrust to all of these conclusions. Epistemic injustice, according to the above analysis, can be addressed to some extent by placing people from different groups in positions where they will be able to communicate with many others, crucially, including other people from that same group. Neurodivergent liberation, then, involves bringing neurodivergent people into positions where they can reach both neurotypical and other neurodivergent people. While there is danger in bringing a neurodivergent person into a situation where they, acting as token neurodivergent person, end up biased against, erased, and even dominated by neurotypical people, creating social networks that involve multiple neurodivergent people, capable of supporting each other, all but eliminates this risk. While it seems likely from my armchair that more exposure to neurodivergent people will reduce bias and stigma, this is actually immaterial—at least a significant portion of the harm of that bias can be greatly mitigated just by having those other alike people to fall back on, and at least a significant amount of personal advocacy can make it through that bias, if implicit. So, unsurprisingly, nothing about us without us, more involving us either way, and emphatically, network-building over tokenizing.

2.2.2 Shortcomings of Network Models

Before we get to work implementing the above advice on the output of a computer program, it is important to consider any disconnects the models have from reality.

Something that briefly came up earlier was the inability for action 1 to generate any information. There is not a clear parallel here for the running example of substances like caffeine, or for any other examples that come to mind. If not drinking coffee works well, that is evidence in favour of not drinking coffee in the same way that succeeding after drinking coffee is evidence in favour of doing so. Unfortunately, this feature is necessary for the model to function at all. If information continues to be generated, so long as the strongest bias is in favour of the truth, it is very hard for a Bayesian agent not to converge on the truth, which is itself not a realistic assumption about human beings. None of the features added to the models above would create biases that overpower the bias to the truth, at best neutralizing some, but not all, sources of bias towards the truth. Therefore, in order to have differentiation, a stable fail state must be installed at the outset, in the form of an alternative that does not help along its own falsification. This feature is also necessary for actions or beliefs of neighbours to matter to individual agents. If action 1 and 2 both yield worthwhile information in favour of action 2, then an agent surrounded by others taking action 1 is in a functionally identical situation to one surrounded by others taking action 2. This would eliminate perhaps the most important feature of the models, and no longer track what they are intended to track. In short, a network learning model *must* have this feature in order to be maximally useful and applicable. This can be taken to be a shortcoming of network learning models.

The damage that this causes is hard to track. This inability to change was an important feature in some of the preceding analysis, but it could reasonably be argued that analysis around this feature is particularly unlikely to directly apply to the real world. This concern is most relevant to the difference between language models with 50/50 and 80/20 language splits. If agents individually will trend towards the truth even when starting with false

belief, then it is less deadly for an agent to be entirely cut off from worthwhile outside information. Then the isolating nature of rare communication styles does not necessarily matter as much as the models imply, in a similar fashion to this isolation becoming a boon for divergent agents. However, this conclusion is still applicable to any situation in which convergence to an incorrect or undesirable belief is stable. If neurodivergent people, or people in general, are unlikely to question false but either especially flattering or unflattering things about themselves, for example, then when they have rare communication styles, the model describes an effect that makes them more likely to end up stuck believing these things. So this is a limitation that should be paid attention to in specific situations, but not altogether damning.

Perhaps the greatest shortcoming, though, is the fact that the models focus only on information exchange. Epistemic injustice is injustice that harms one in their capacity as knower, which is not limited to harming only one's knowledge. As Fricker argues, this includes harming one's sense of self, among other things [6]. The above models do not have a metric for sense of self or any other social, emotional, or well-being metrics, nor are they well-suited to accommodating such a metric. I have two comments on this fact. The first is to note that the information dynamics may describe other kinds of dynamics as well. For example, a divergent agent's tendency towards the more stable belief could also describe a tendency towards an internally stable way of being, while the other agents have a tendency towards a way of being that is stable for them in that they tend further towards it when left alone, but unstable for the divergent agents for the same reason. This divorces the mathematical dynamics from epistemology entirely, and views it as a model of personality change, in line with the discussion of masking above. Why change in personality would follow Bayes's theorem is totally unclear to me, and I would not defend this reading of the models, but it bears bringing out that the models are just numerical, and do not *need* to be interpreted epistemically. It is not a foregone conclusion that other features we might be interested in would *not* follow the same patterns found in these models' dynamics merely

from the fact that the dynamics create the feature that taking action 2 causes others to be more likely to take action 2, unless the agent doing so is a divergent one, which might be true of many activities that are not epistemic in nature. The other comment I would make is that regardless of the preceding, this critique is exactly right. A pure focus on how agents learn *does* cause us to miss a lot more of what is going on. If I were actually interested in creating a mathematical dynamics to describe some interesting changes in a person's life other than their knowledge, I would plainly not have used Bayes's theorem. So even if it is plausible that some features will follow knowledge change, I would discourage conclusions based on this plausibility. While there is reason to hope that acting on these models would have positive outcomes for other aspects of people's lives, if we want to discuss other specific elements of human activity, a different kind of model would go much further. More on this later.

Also of note is that in these models each agent is interested only in what action is best for them to take, and has no choices to make in sharing information. Injustice is not entirely a naturalistic phenomenon—indeed, construed properly, it could be argued that it is entirely unnatural, “naturalistic” here meaning “arising purely from elements out of conscious human control.” To naturalize injustice is to claim that it is a non-malicious result of background conditions, downplaying intentional actions taken by malicious agents in reality. To put it bluntly, atrocities like slavery are not accidents coming from neutral environmental factors. Further, it could be taken that a proper definition of “injustice” precludes purely naturally unfortunate situations, and describes only things that naturally would be one way but instead unduly harm certain groups. I think this disagrees with some legitimate uses of the term, but not with Fricker's stipulation that epistemic injustice must be identity-based [6]. To move away from the terminological, someone's beliefs are more naturalistic about injustice the more they think is explained by happenstance and less naturalistic about it the more they think is explained by blameworthy actions. The problem then is that the features of the models cannot be blamed on the agents. They do not arise

from the few choices the agents make. A more antinaturalistic study of injustice, or at least one more agnostic to naturalism about injustice, would endow agents with the ability to take actions that can lead to advantages or disadvantages for other agents. Without possible incentives and capacities for deception, conflicts of interest, and so on, a key element of epistemic injustice is missing.

This point is interesting in that it might lead to future more powerful models, but for now I do not think it presents a particularly serious problem for these models. Fricker and O'Connor both provide justification. Fricker's examples do often involve only implicit bias, and where other authors disagree with her, they have so far typically moved further in this direction, implicating actions that Fricker had excused as merely unfortunate¹¹. The phenomenon of epistemic injustice, as it is currently understood, is understood to be at least partially caused by features of social dynamics not directly caused by conscious choices made by individuals. Likewise my agents are not blameworthy in the sense of deceiving or acting against other agents, but may map onto individuals who are blameworthy for leaving their biases unchecked. O'Connor also gives the justification that a formal model does not need to purport to fully explain a phenomenon, and indeed generally *should* not [17]. By giving a minimum effect that comes up without active intervention, some element is at least explained. So, at worst, the above models do have explanatory power about background conditions that play a role in epistemic injustice, which means they can be helpful in combating epistemic injustice. Even if epistemic injustice is constructed entirely out of the malicious intentional actions of oppressors, an understanding of which background conditions make those malicious actions more effective or less effective is instrumental in fighting back.

Finally, there is of course also a distinction between real-world factors that lead to epistemic injustice, such as bias and communication failure, and the factors of the above models intended to emulate these real factors. Implicit bias is a philosophically problem-

¹¹See Dotson 2012 and the discussion of it in chapter 1

atic concept; Jules Holroyd, Robin Scaife, and Tom Stafford have described a number of philosophical issues engaged when trying to circumscribe it [42]. Whatever implicit bias is, it is not a specific percentage chance of disregarding information. Therefore, if I say “the bias models revealed a certain dynamic, so we should consider this dynamic when fighting bias in the real world,” I am using the same word to mean two different things, and risk equivocating. I have made an effort to not make this mistake, and distinguished in discussion of actual on-the-ground strategy between real-world bias and bias models. Conclusions are still possible that may be helpful to combating bias because we can talk about effects that arise from situations that fit a general description, and that description may cover some situations of real bias. My response, then, is that I have been aware of this throughout and it should not mitigate any findings as they have been presented, but it certainly does bear repeating that just because a factor I have named after something real has a certain effect in the model, that does not mean that the factor’s namesake behaves in the same way in general or ever.

To conclude, there are some problems with this kind of model, and in sum this may raise some worry about directly applying findings that come from them. The discussion here has been far from useless, and I stand by the overall recommendations I have made on the basis of points of agreement between my models and either prior experience or the literature. Nonetheless, there is a hope that more can be said and done. If some of the biggest problems for this kind of model are due to ignoring certain factors, then the next step may be to then try to account for these factors in a different kind of model. It is in theory possible to get results beyond boilerplate “put neurodivergent people into the mainstream” by studying epistemic injustice and neurodivergence with an entirely different kind of model that *does not* use Bayes’s theorem.

Chapter 3

Injustice in Norms of Communication

In this chapter I will motivate the following chapter's models of hermeneutical injustice based on signal games. I will outline signal games and game theory more broadly, as well as explain my somewhat novel signal models. I will also revisit the literature on specifically hermeneutical injustice in more detail, and while I primarily draw on the same experiences described in chapter 1, here I will focus on the hermeneutical aspects of those experiences to help motivate the move to a different formal tool.

Unlike simple network learning models, Lewis-Skyrms signal models are not popular among the subsection of the literature that is especially concerned with real-world problems. David Lewis introduced signal games in *Convention* to demonstrate that linguistic meaning can arise naturalistically out of situations initially devoid of meaningful communication [43]. Skyrms, in *Evolution of the Social Contract*, adds a powerful temporal aspect by adding probabilistic dynamics to the game, to demonstrate that meaning can *evolve* out of such situations [44]. However, while Skyrms makes some social commentary, he does not apply these or any other models to anything related to epistemic injustice. Significant modification of the Lewis-Skyrms version is needed to make the models suit this particular purpose, though the essence of them will remain the same.

As before, this type of model will give extensive information on how the systems it

can describe develop. Much good work has been done to understand how marginalized people come to be excluded from cultural formation of meaning, and many of the recommendations of this work are well-suited for implementation, and even similar to those that arose from the mathematical analysis in the previous chapter (for example quite simply just putting marginalized people in the spaces where conversations relevant to them happen and allowing them to participate). However, there is still a niche to be explored for tools that are capable of getting into the details of the results of specific kinds of exclusion in different situations, in order to more finely guide our response. The goal of this chapter, then, is to motivate signal models as a way to fill that niche, so that the following chapter can use one to do so.

3.1 Hermeneutical Injustice

In the previous chapters, I discussed the concept of epistemic injustice, injustices that harm someone specifically in their capacity as knower. I also briefly explained specifically hermeneutical injustice, epistemic injustices that specifically draw harm from or cause harm to someone's available concepts. Here I will explore this latter kind of injustice in more detail, and give examples that will inform my models.

3.1.1 Fricker and Hermeneutical Injustice

As discussed in the previous chapters, when defining hermeneutical injustice in particular, Fricker describes it as "the injustice of having some significant area of one's social experience obscured from collective understanding owing to a structural identity prejudice in the collective hermeneutical resource." The core idea is that once articulated into a larger cultural conversation, concepts can make cogent to all the struggles of a few individuals. When the structure of that larger cultural conversation is such that the concepts that become formed within it are not suited to explaining the experiences of a group, that group expe-

riences a powerful harm, their marginalization being worsened by their social lives being made incomprehensible. In some cases, Fricker poses a worry that there is a sense in which this can lead a group to have struggles that are incomprehensible even to members of that group. This can be seen in her canonical example of workplace sexual harassment; it is not that individuals do not know that they are being harassed in a sexual manner, understand how that works, why it is bad, why their harassers can get away with it, and so on. It is that prior to speak-outs popularizing the concept, they do not recognize it as a singular struggle that others are going through with them by having this fact conceptually obscured by their harassers and those who would protect them [6].

As before, Fricker gives specific conditions that must obtain for a situation to truly be hermeneutically unjust. As before, these conditions amount to requiring that the injustice come from conditions that are themselves unjust, and that marginalization is happening, distinct from a merely unfortunate lacuna in understanding. To use one of Fricker's examples, Edmund White discusses in his memoir that as a child his only conception of a gay man is a sick and sinister vampire-like creature, and so he was unable to conceptualize himself as gay. If, as obviously is the case, it is because of identity prejudice and not some innocent mix-up that the cultural conceptual resource allows only negative concepts of homosexuality, then he does experience hermeneutical injustice. Likewise for any particular group [6].

I think it is obvious that neurodivergent people suffer from hermeneutical injustice's primary harm of not being understood. Autistic people are frequently represented as child-like and unable to make choices for themselves, and people with ADHD are often regarded as capable but, in a sense that is intentional and morally vicious, lazy. That these perceptions are identity-based is patent, and that they come from ableism seems overwhelmingly likely from the presence of ableism throughout society, and from the clarity with which anyone who is sufficiently unclouded by ableism and knows adult neurodivergent people can tell the inaccuracy of these portrayals.

When neurotypical people regard autistic people as infantile, it can be very easy for them to be paternalistic towards them, to their direct harm. One controversial example around autistic people is applied behaviour analysis (ABA) therapy, a type of childhood intervention in which practitioners use negative and positive reinforcement to eliminate behaviours that it labels disruptive in the life of an autistic child, most notably stimming, and encourages masking. “Stimming” refers to actions many neurodivergent people take to stimulate themselves and relieve distress, which include a wide variety of actions including rubbing fingertips together, repeating sounds, flapping hands, or biting oneself. Stimming, however, is often a rational behaviour in the context of autistic sensory phenomenology, and erasing it in all forms can amount to a sensory nightmare with no available coping mechanisms. Chapter 1 discussed problems associated with masking, or hiding one’s autism. Many autistic adults report deep resentment at having gone through ABA, and some researchers have argued it constitutes abuse, but parents who have not gone through it, and obviously have the best interests of their children at heart, subject autistic children to it out of lack of understanding [45] [46] [47] [48]. For balance, it is worth admitting that some stimming behaviours can be dangerous and ABA is effective at eliminating such behaviours, but a strategy that does not replace dangerous behaviours with safe alternatives is an incomplete one at best, and it is unclear what the benefit is supposed to be for eliminating the vast majority which are safe, even if ABA does so efficaciously. While this is a fairly extreme example, it should not be controversial that neurodivergent people are at least to some degree, some of the time, treated worse than they could be due to misunderstanding.

Neurodivergent people also experience what Fricker gives as a secondary harm of hermeneutical injustice, that individuals experiencing hermeneutical injustice may be less able to understand themselves. Many neurodivergent people discover only in adulthood that they are neurodivergent [1] [2]. A very common thought is that using the concepts that come with neurodivergence can lead to far deeper self-understanding and self-acceptance, as we become able to articulate to ourselves why we do certain things we do, and come

to understand our actions as normal within the context we belong to. Prior to finding this understanding, there is a sense of confusion not unlike what Fricker describes. It is clear to neurodivergent people that we are different from neurotypical people, but not necessarily how or why. When we have to articulate our actions to others and to ourselves using the conceptual resource designed by and for neurotypical people, there is a lot that is lost. By way of example, I recall vividly a time I expressed my confusion and frustration to a friend that I really wanted to do something with my afternoon, and found myself lying in bed, not even scrolling social media or daydreaming, just wanting to get up and not. I could not understand my own actions, as I was working under a theory of action where desire and will beget like action without fail. I could not understand my experience of seeming to be willing something, but believing I must not *really* be willing it, as it is not happening. My friend told me I was describing executive dysfunction. Since this experience, I have become much more articulate to myself, and much more accepting of my own difficulties in task-changing as neurological in nature, not stemming from a mysterious personal vice.

Neurodivergent people suffer a lot when not given the conceptual resources of neurodivergence. If it is true that neurodivergent people are being deprived access to this conceptual resource due in part to structural identity prejudice¹, then we are suffering a substantial amount from hermeneutical injustice.

Fricker's response to hermeneutical injustice is not particularly different to her response to testimonial injustice, and has the same problems. Epistemic justice as a virtue of individuals is touted as a way to combat its corresponding vice in individuals, and as Anderson has noted, this is an odd response to something that is systemic by definition [10]. Alongside Anderson rather than Fricker once again, I would like to use my computer models to explore systemic solutions to systemic problems, in this case the problem of neurodivergent people being hermeneutically marginalized to the point of being incomprehensible to themselves and others.

¹It is.

Before moving on I want to note one addendum to Fricker's theory from José Medina that will be instrumental in motivating the use of signal models to tackle this issue. In "Hermeneutical Injustice and Polyphonic Contextualism," Medina notes that hermeneutical injustice arises from failures in "communicative and interpretive responsibilities." He stresses that while hermeneutical and testimonial injustice are far more connected than Fricker describes, hermeneutical failures come from communicative failures [12]. For this reason, modeling specifically hermeneutical injustice requires modelling communication in detail, rather than learning. The conceptual resource available to a group is shaped by the history of communication within and around that group, so while the previous chapter's models focused on describing a history of learning, that history of communication should be the focus in chapter 4's models.

3.1.2 Examples of Hermeneutical Injustice to Model

Some situations that can rightly be called hermeneutical injustice are already covered by the models in the previous chapter, although those models did not necessarily get at their specifically hermeneutical aspects. I will briefly outline these situations and what was missing, then go over a few more that I will model in chapter 4.

Bias has been dealt with in the literature, and focusing on identity-based bias has been dealt with in the previous chapter. In these previous implementations, bias is modeled by having agents, all of whom are trying to learn, be less likely to believe the testimony of agents whose views diverge significantly from their own. What is measured by these models is epistemic, and not hermeneutical, in nature, since the models involve only learning and information sharing. Identity-based bias is also central in hermeneutical injustice. In Fricker's central case, it is identity-based bias that causes the epistemically vicious actor to fail to mirror the biased-against actor's thoughts back to them as expected, causing a disruption in their self-actualizing activities [6]. The phenomenon to model, then, is when a community has some degree of bias against a group with a specific identity, and therefore

does not give them the same uptake on how they communicate.

I would here like to expand this to include success-based bias. Success in the real world is not always due to merit, and paying attention only to measurable success may be one way ableism is systematized. By having agents pay specific attention to which other agents have had the most success communicating, it will be interesting to see whether these agents cause further hermeneutical injustice to develop. This may be of particular interest in advancing understanding of epistemic injustice because one issue I would raise to Fricker's initial treatment of the topic is her usage of eye contact as a legitimate indicator of trustworthiness. For Fricker, there are some instances of epistemic unfairness that are merely unfortunate but not unjust, as when the individual being communicated with has good reason not to regard the communicating individual's testimony as particularly credible [6]. Her example of distrusting a genuinely trustworthy individual because of a lack of eye contact seems very strange to me, given that, as is well known, autistic people struggle with making the expected amount of eye contact. It strikes me as unjust for people to be prejudged due to eye contact, and it strikes Fricker as unfortunate but not a matter of justice. It is possible that results of modeling success-focused metrics for hermeneutical uptake may be helpful in arguing one side or the other.

Mere difference can also play a role in hermeneutical injustice. This was dealt with to some extent in the divergent agent model and the communication styles model. When agents are modeled as communicating differently, some aspect of hermeneutical injustice is being addressed. We do not see concept formation in the previous models, but epistemic injustice brought about as a result of lack of concept uptake is still arguably in overlap with hermeneutical injustice, as the cause of the injustice is still hermeneutical in nature. Nonetheless, existing difference in communication styles could certainly play a role in unjust failures for communication of certain kinds to develop. For example, if it is already difficult for a neurotypical individual to understand the experiences of a neurodivergent one, an attitude that sometimes causes the neurodivergent person to be ignored may result

in a more complete ignorance, lowering the number of chances for connection despite that attitude. Therefore, in some of the below models, model types with agents that differ from each other should continue to be explored.

I previously explored the role of agents having free association with mostly like-minded others. Isolating different communities based on ideas, like the behaviour modeled in the previous chapter's mingle models, can lead not only to disruption in good learning, but can lead to substantial changes in communication on its own. If someone is not present in a conversation, they plainly cannot contribute, an effect that likely works in degrees as well. To test both the extreme and the degrees, it would make sense to implement something like mingling again, having agents choose which others to communicate with.

Finally, something that was entirely absent before was difference in goals. While in this project I do not intend to explore models with players incentivized to act against the interests of other agents, I do think it would be useful to look at how communication evolves when certain agents are completely indifferent to the interests of other agents. It seems to me that the mere attitude of indifference can cause hermeneutical injustice; hermenetucial injustice arises from lack of uptake of and attention to communication from particular identity groups, so if agents simply have no reason to lend attention to this communication, we should see it arising absent any ill intent.

3.2 Signal Games

The specific formal tool I use to describe specifically hermeneutical forms of injustice is the signal game. The basic premise is that some number of agents are trying to decide how to communicate toward a common goal, and have two different kinds of actions: actions meant to communicate, and actions meant to interact with the world. The game is to pick the best communicative actions to help the other agents act correctly when there is no prior convention on what communicative actions are intended to communicate. The hope of this

chapter is that a dynamics for choosing how to signal and act used while modeling sources of hermeneutical injustice will give some idea of how methods of communication begin to form in ways that disadvantage marginalized people.

3.2.1 A Brief Introduction to Game Theory

Game theory is a type of formalism relevant to math, economics, and philosophy, that models real-life decision making by pretending it is a game the decision-maker is trying to succeed at. The primary benefit of this is that it creates a robust way to study good decision-making around other decision-makers.

To give a simple example, suppose you are trying to decide what to wear to a party. What the best outfit is going to be will be entirely subjective. Nonetheless, in a certain view, there will be an end result that makes you happiest, even if you do not already know which end result. There may be one outfit, or a set of outfits, that makes you the most confident, gets you the most compliments, or whatever else matters to you about your outfit. So, it can be made into a one-player game. You have a number of moves available equal to the number of outfits you have available. You can get a certain number of points by playing each move—the scale of the points is arbitrary, but if you are twice as happy wearing the elegant blue dress as you would be wearing the simple khaki suit, then the elegant blue dress should give twice the points as the khaki suit, and if showing up in a white t-shirt and boxers is going to be a worse time than just not playing at all, which in this case probably means staying home, then that is usually represented as giving negative points, though formally there is no difference between doing that and just pinning the worst choice at 0 and making “refusal to play” a move with a specific point reward.

Where this begins to have utility beyond just labeling the best options as the best options, which does not require the trappings of games, is when another person’s decisions matter. Say I have a similar wardrobe to you. I might have an elegant blue dress that is similar to yours, but that I wear slightly better. If we both wear our dresses, then you will

	Blue Dress	Suit	Stay Home	T-Shirt and Boxers
Blue Dress	-1/-1	2/1	2/0	2/-1
Green Dress	2/2	2/1	2/0	2/-1
Red Dress	2/2	2/1	2/0	2/-1
Stay Home	0/2	0/1	0/0	0/-1

Table 3.1: First outfit game payout matrix. Each cell gives the payout of the row player on the left and the payout of the column player on the right

	Blue Dress	Boring Suit
Blue Dress	-1/-1	2/1
Green/Red Dress	2/2	2/1

Table 3.2: Second outfit game payout matrix, simplified from the first.

actually do very poorly, because you will constantly be self-conscious about being compared to me. Indeed, I might also be self-conscious because I erroneously believe that you are wearing the dress better, or more likely because I feel bad for ruining your night². This can be turned into a game where the wardrobe choices are moves. If only one of us is aggressive, wearing the dress despite the risks, that person wins out. If both of us do it, we will both be sorry. Now we can pin specific payouts to each combination of moves, that is, wardrobe options. From the discussion so far, and populating my own wardrobe, we can generate a *payout matrix*, a table that tells us how much each player likes each option. This game's payout matrix is table 3.1.

As can be seen from table 3.1, our decisions impact each other only if a blue dress is involved. As well, if we are worried about this clash happening, we each have a clear other option we can choose. Therefore, the table can be simplified for anyone who cares about only the best choices available to everyone, given as table 3.2.

Table 3.2 gives a fairly clear image of this game's best moves. From your perspective, it would be best if you wear your blue dress and I wear my green one, but you can avoid risk by picking the boring suit. A naive decision-maker might pick the boring suit just from noticing that a payout of 1 is better than the average of the possibilities of wearing a blue dress, -1 and 2. However, game theory gives a different result. Notice that for me, I am

²No offense to the reader. It's just, have you seen me in a dress?

guaranteed 2 points if I wear the green or red dress, but if I wear the blue dress, I am taking a risk without any possible reward of doing any better. The green dress is perfectly fine. If you know that I know about your blue dress, for example if we were together when you bought it, then you have reason to believe I will not take the unnecessary risk of wearing my blue dress. Why would I want to ruin both of our nights? So, it turns out that you are safe to wear your blue dress, as long as I am thinking the situation through as well, and there are no more dress-wearers we need to worry about. You can show up in blue, the two of us will be dazzling but not too similar, and we will be the talk of the night!

Notice that even though I said that I look better in the blue dress, it was not necessary to give me more points for wearing it. In fact, the payout matrices in tables 3.3 and 3.4 give the exact same result. It is still true regardless of the magnitudes chosen that I am better off wearing green or red, and you are better off wearing blue. This result, and not a precise representation of who is happiest, is what matters. Nobody is trying to *win* by having more points than the other players, the goal is to maximize your own points, whether or not you bring others down. For example, in the abstract game described by table 3.5, it is still in the interests of the row player to pick option 1, even though it helps the column player even more. If I have 1 point and you have 0, I am less well off than if I have 2 points and you have 100 or 50. The points are more like happiness than money, the presence of a player with substantially more than me does not hurt me in any way. At the same time, no matter how nice I am, I should not pick option 1 in the game described by table 3.6, which helps you substantially, but does not help me. If I am kind and care more about your well-being than my own, then this game can only make sense if there is a difference in opinion about what is best for you. The table describes a situation in which I genuinely believe that I prefer with option 2, including my preference for you being happy, if I have one. For example, the row player may be a parent and the column player their child. The parent may believe that forbidding the child from seeing industrial metal band Ministry is in the child's best interest to avoid any satanic influence on the child, while the child believes that seeing

the lead singer Al Jourgenson perform in the flesh would be emotionally and spiritually uplifting. As game-theory is about decision-making, in situations like this where there is a difference in opinion, the numbers will represent what each player believes, not what ends up happening. Even if the parent comes to realize their paranoid decision has stunted their relationship with their child, or even if the child's soul is dragged to hell in an industrial metal ritual sacrifice, that does not change what the best way to represent the decisions in a payout matrix was.

	Blue Dress	Boring Suit
Blue Dress	2/-1	3/1
Green/Red Dress	3/2	3/1

Table 3.3: Third outfit game payout matrix.

	Blue Dress	Boring Suit
Blue Dress	-5000/-1	-1/1
Green/Red Dress	-1/2	-1/1

Table 3.4: Fourth outfit game payout matrix. Each outfit game has the same results.

	Option 1	Option 2
Option 1	2/100	2/50
Option 2	1/0	1/0

Table 3.5: Abstract game demonstrating that giving another player more points is not a bad thing.

	Option 1	Option 2
Option 1	1/100	1/50
Option 2	2/0	2/0

Table 3.6: Abstract game demonstrating that kindness is accounted for in points.

Situations like that which arose in the outfit game, in which there is one pairing of moves that neither player has an interest in deviating from, are called “Nash equilibria.” Nash equilibria are quite common in games, and can involve probabilistic strategies. A more common example of a Nash equilibrium is in a game called a *prisoner's dilemma*.

In this dilemma, two prisoners can make the other worse off by testifying against them, reducing their own sentence. Because the sentence is reduced no matter what the other agent does, both testifying is a Nash equilibrium, as the agent making the choice does not benefit from refusing. See table 3.7 for the payout matrix of a prisoner's dilemma game. Note that this game is often misunderstood due to its common framing. It does not suggest that it is actually prudent to do what is best for you and damn the consequences for others—if two prisoners are loyal to each other, their payout matrix will not actually look like a prisoner's dilemma. It is only when there is an actual clash in overall interest that a dilemma like this can occur. A probabilistic example can be found in the common game *rock, paper, scissors*. Its matrix given as table 3.8, both for completeness and in the hopes that it helps clarify how payout matrices work. It is a Nash equilibrium to play each move exactly one third of the time. If I know you play rock more often than the others, I can gain an advantage by playing paper more often. However, even if I don't change my strategy, you still have not changed the fact that we each have a 50% chance of winning, since rock still has the same win rate as the other moves against me. Therefore, even if I do not think you know about game theory, it can be argued it is rational for me to stick to the Nash equilibrium, since I deprive you of the ability to strategize against me. The best possible play in reality is to anticipate the opponent's strategy and play only the highest-winrate move against them on a given throw³, but this is not generally a reliable strategy, as things get complicated fast with opponents able to make it look like they have one strategy in order to bait out another and so on. Therefore Nash equilibria are often taken to represent the most rational endpoints of any game in which they exist.

	Rat	Silence
Rat	-1/-1	2/-2
Silence	-2/2	1/1

Table 3.7: The Prisoner's Dilemma. For each player, ratting is a better option no matter what, so the rat/rat result is an equilibrium, but the universally preferable silence/silence result is not.

³This is something my sister seems able to do.

	Rock	Paper	Scissors
Rock	0/0	-1/1	1/-1
Paper	1/-1	0/0	-1/1
Scissors	-1/1	1/-1	0/0

Table 3.8: Rock, paper, scissors, given as a payout matrix. The only Nash equilibrium is probabilistic.

3.2.2 Lewis Signal Games

The first signal game was described by David Lewis in *Convention*. He described the game as follows. There are two agents, a signaller and a receiver. The signaller observes either one state of nature or another. The signaller and receiver have a mutual interest in the receiver acting a particular way based on this state of nature [43]. To remain on-theme, let us say that two agents are trying to determine whether to go to a party or stay in. Going out will either be an enjoyable experience or not depending on whether the party will be a good one or not. The signaling agent has received word on whether or not a party is good, and is trying to help the other agent decide whether to go. If the party is good and the receiver goes, both agents are happy, and if the the receiver stays, both are unhappy, reversing these results if the party is not good. “Good” here is left intentionally vague, but might include features like whether there are people there an agent does not know, how loud it will be, and so on, that are not, in reality, objective. For the standard case of these models, suppose it is objective whether or not a party is good. I will complicate this with differing views among agents of what makes a good party later. The signaller, then, can take one of two actions to try to communicate to the receiver which to do. We have a chain of actions: Nature is observed by signaller, and signaller takes a signal action; a signal action is observed by receiver, and receiver takes an action. The problem, though, is that there is no pretheoretical reason that one signaling action would communicate one thing and another signaling action would signal the other - there is not as yet meaning assigned to either action.

If we turn this into a game, there are two possible *Nash equilibria*, which again means situations which, once attained, will remain in equilibrium because no agent in the situation can benefit from diverging from it. If we label each state of nature state 1 and state 2, the corresponding actions likewise action 1 and action 2, and the signals signal *a* and signal *b*, then the equilibria are that signal *a* corresponds to action 1 and signal *b* to action 2, or that signal *a* corresponds to action 2 and signal *b* to action 1. Once such an equilibrium is established, the signaller will always choose the signal corresponding to the action corresponding to the observed state of nature, and once given a signal the receiver will always choose the action corresponding to that signal. These equilibria are called *signal systems*, and they describe systems in which the signals have clear meanings. See the payout matrix of this game in table 3.9, and a visual representation of the game in figure 3.1.

	Nature 1: Signal <i>a</i>	Nature 1: Signal <i>b</i>	Nature 2: Signal <i>a</i>	Nature 2: Signal <i>b</i>
Action 1	1/1	1/1	0/0	0/0
Action 2	0/0	0/0	1/1	1/1

Table 3.9: Lewis Signal Game payout matrix. While payout depends on the state of nature rather than the action chosen by the column player, there are still Nash equilibria since the row player can base its choice on the column player's choice, and the column player can base its choice on the state of nature.



Figure 3.1: A Lewis signal game. The signaller sees a state of nature, which has a corresponding correct action, and sends a signal that the receiver must use to determine the correct action. In this case both have settled on the same equilibrium and there is no reason for either to diverge.

The argument, then, is that pure convention can arise prior to meaning and create that meaning, rather than conventions requiring meaningful language prior to their being formed in that language. The arbitrary choice of which signal corresponds to which action creates

the meaning of the signals. There is a lot that can be said about this argument, but that is the motivation at any rate, and I do not need to establish how well it works.

3.2.3 Lewis-Skyrms Signal Models

Skyrms is unsatisfied with the above, desiring a mechanism by which one convention can be chosen. He notes that evolution can fill this role. In Skyrms' formalism, there are two sets of urns. The signaller has one urn for each possible state of nature, and the receiver one urn for each possible signal. When the signaller observes nature, they find the corresponding urn, and pull a ball from it. The kind of ball will decide for them what to signal. Perhaps the balls have colours corresponding to flags, but for easy analogy to the above I will say the balls have letters printed on them like bingo balls, and signal a or b will happen when a ball with an "a" or "b," respectively, is drawn. Likewise, the receiver has an urn a and urn b , with their own collections of balls labeled "1" and "2," and will take action 1 when drawing a "1" ball or action 2 when drawing a "2" ball. Where evolution comes into it is that when the two act successfully, they will place some number of balls of the same kind that they just drew into the urn from which they just drew. As a result, over time, communication can become more efficacious [44].



Figure 3.2: A Lewis-Skyrms signal model. Each agent has a pair of ball-filled urns representing the likelihood it will take each action conditional on each input. In this case the correct action was taken due to the first agent taking the slightly less likely signal action given its urns, and each will change the urns used to make this occurrence more likely.

A perfect signal system in this formalism is described by either the signaller having all "a"s in urn 1 and all "b"s in urn 2 while the receiver has all "1"s in urn a and all "2"s in urn

b , or the signaller having all “ b ”s in urn 1 and all “ a ”s in urn 2 while the receiver has all “2”s in urn a and all “1”s in urn b . It is possible to have a perfect signal system evolve if either it is the state in which the urns start, or if there is a mechanism for removal of balls. The latter can be easily done by simply not replacing balls in the urns once drawn, or by having a random ball removed each round. Models that leave the drawn ball out of the urn are simply referred to as using “drawing without replacement,” and models with random loss of balls are referred to as using “forgetting.”

	Nature 1: Signal a (90%)	Nature 1: Signal b (10%)	Nature 2: Signal a (5%)	Nature 2: Signal b (95%)
Signal a : Action 1 (90%)	1/1 (81%)	-	0/0 (4.5%)	-
Signal a : Action 2 (10%)	0/0 (9%)	-	1/1 (0.5%)	-
Signal b : Action 1 (15%)	-	1/1 (1.5%)	-	0/0 (14.25%)
Signal b : Action 2 (85%)	-	0/0 (8.5%)	-	1/1 (80.75%)

Table 3.10: A step in an example Lewis-Skyrms signal model represented with the percent chances each agent has of making each choice based on the input they see, with each possible result cell giving the percent chance that it happens conditional on that state of nature occurring. The expected payout for both agents in either state of nature is the sum of payouts weighted by their chance of occurring, which here is 0.825 in state 1 and 0.8125 in state 2.

Removal of balls does not guarantee a perfect signal system will evolve. By poor luck, it is always possible, for example, that one agent ends up with only balls of one kind in either urn, and then no signal can have any meaning, since either only one signal is possible no matter the state of action, or only one action is possible no matter the signal. For this reason I will use a safeguarded version of the model where no ball can be removed if it is the last of its kind in an urn. This can be seen as representing agents applying the reasoning principle

beloved by many probabilists that no possibility should ever be ruled out entirely.

However the specifics are set up, Skyrms finds that not only are the signal systems Nash equilibria, they are also evolutionary equilibria, meaning situations that systems with evolutionary dynamics are more likely to evolve towards and stay near than they are to evolve away from. In other words, it is far more likely that a particular model results in an approximate signal system after enough steps than that it does not do so⁴. The argument here, then, is that evolution can be a source of convention. Its chaotic processes can lead to natural signals with natural meanings even when it seems a matter or naught else but convention that those meanings would come about. The more mechanical nature of this argument, to me, makes it more compelling than Lewis'. There is much less room for interpretation here, in most conditions benefiting from collaboration, meaning is demonstrably mathematically likely to evolve from meaninglessness.

3.2.4 Simultaneous Signal Models

My signal models are essentially Lewis-Skyrms signal models, with a more complicated mechanism for selection of action urn and best action in order to implement some of the features.

The main focus is to increase the number of agents in a given model, in order to explore social dynamics. Skyrms has previously explored signal models with multiple signallers and receivers, but his signal models are different than the ones I will use. In particular, Skyrms' multi-agent models are not large signal models; instead they have populations in which individual agents come together to play single-run dyadic signal models and learn from these experiences [49]. As I wish to have a single ongoing model in which the agents can give different weight to information received from multiple sources, agents will need to be acting simultaneously as signallers and receivers, and be simultaneously receiving signals from multiple signallers, instead of playing many distinct concurrent models.

⁴Indeed, eliminating the possibilities of only one kind of ball means signal systems are the only equilibria.

Simultaneous models have been done before with two agents. It is easy enough and changes very little to have an agent act as a signaller, then immediately act based on the other agent's signal, then both update both their signal and action urns. The only interesting thing to say about this kind of simultaneous model is that unless there is a larger population of such agents, as in Skyrms' many-agent models, half of the equilibria involve the agents having opposite signal systems working at once, as if one agent can only speak in Spanish and only listen in French, and vice versa for the other [44]. However, applying the same rules as a dyadic model stops working quickly beyond two both-role agents. If an agent signals and then receives two signals, which urn should the agent choose? Either the agent will have four action urns, one for each permutation of signals, or the agent will have to have a decision process to pick one of the signals to actually listen to. The former turns out to work poorly. Not only does the number of urns grow exponentially as more signals or agents are added, to a total of s^{n-1} urns for s signals and n total agents, I have found in modeling that convergence becomes far more difficult even at just four urns. In my simultaneous models, then, agents will have a decision process for choosing a particular type of signal out of the ones they received, and will choose the action urn corresponding to that type of signal.

There is a further problem of deciding what an agent's best action actually is. In a 2-agent simultaneous game, each agent is generally described as viewing a distinct state of nature that the other agent is acting on. This verbiage is confusing, as nature can at once be in two different states for the two different agents, and the way each agent should act depends on the state seen by the other. So instead, from this point, when discussing simultaneous models, I will merely say that each agent makes an observation. The observations will be taken to be accurate, meaning there is not a probabilistic process like an urn for whether the observation matches reality. The first agent makes an accurate observation about what the best action for the second action would be, and vice versa; the game is still in the communication.

Again, however, this will not work beyond two agents. Either there will be s^{n-1} actions, where s is now the number of states of nature, and convergence becomes much more difficult, or at the model level there will be a decision process to assign each agent a best action. I settled on a decision process that works the same in each case, resulting in one best action shared by all agents, or in analogy to the running example, there is one party that all agents are trying to decide whether to go to. First, each agent receives their own observation, representing one piece of information about the party pointing unambiguously to it either being more likely to be good (state 1) or bad (state 2). The actual state of the party will remain binary, decided by summing the observations of each kind, and selecting the state matching more observations. For example, in a ten-agent system, if six agents make observation 1 and four make observation 2, it is actually going to be a good party, state 1. If the number of possible observations and number of agents line up such that ties are possible, ties will be decided by coin flip.

A very similar procedure will also occur at the agent level. The agent receives $n - 1$ signals, but also has its own observation. It would be simple to just have the agent count its own signal, so that each agent sees the same incoming signals and chooses the same action urn. However, it is possible to do slightly better. The agent knows how many balls of each kind are in each urn, and knows what its own information says about the best action. So, it can instead count its information as a signal pointing to the urn that is more likely to give the correct action. Indeed, since it knows its own information, and only has ambiguous signals from other sources, the agent could choose to inflate the weight of its own observation, or even ignore all other signals entirely. To best meet the goal of what these models are supposed to do, however, I stop short of this, since it is necessary that decisions be primarily made based on what signals are received in order to have the agents learn to signal⁵.

Figure 3.3 is an example of one step of this type of signal model. I will walk through the

⁵I will, however, use this feature for the mechanics of the difference model type described below

specific example in this paragraph. First, each agent has a randomly generated observation, in the circle in the middle of the diagram. It happens by pure chance that three agents, agents 1, 2, and 4 have “observation 1,” and the other two “ observation 2.” Following the running example, this means agents 1, 2, and 4 have reason to believe the party will be good, and the others that it will be bad. Agents 1, 3, and 5 lean towards signaling and acting the same number as their input, e.g. observation 1 goes to signal 1 and signal 1 goes to action 1, or signal 1 means “good party.” Two of these three witnessed information that ended up not reflecting the actual majority of information, and have reason to think it will be a bad party, even though it will be good. Four of the agents send signal 2, and one sends signal one, all happening to line up with what is more likely in their respective signal urns. Because each agent receives at least three matching signals, their individual counts do not end up making a difference; each is acting like there were at least three total 2-signals, and each uses their action urn labeled “signal 2.” In the end, the agents that lean towards the choice matching their input all acted incorrectly, choosing action 2, “stay home,” but the the agents that lean towards the choice differing from their input all acted correctly, choosing action 1, “go to the party.” While each step had random chance involved, there is a strong sense in which the initial condition of only matching-choice agents getting bad information caused this outcome, as this made it far more likely that the signal differing from the best action was a dominant signal, which made it far more likely that the differing-choice agents would act correctly but the matching-choice agents would act incorrectly.

3.2.5 Hermeneutical Injustice in Signal Models

Now that the previous subsection has established a framework that will allow for more complex features, I will discuss the situations that can be involved in hermeneutical injustice that I wish to model. First, there are the signal models modeling identity-based bias. While this had been done in bayesian learning network models before, applying bias to signal models is entirely novel. In order to model bias, I am once again tagging certain



Figure 3.3: An example of a step in a simultaneous Lewis-Skyrms signal model. Agents are shaded yellow if they are closer to the 1:1 and 2:2 signal system, and they are shaded blue if they are closer to the 1:2 and 2:1 signal system.

agents, those with ID numbers below a certain given parameter, with an identity tag. When receiving signals from agents other than tagged agents, agents will count those signals three times. As well, they will also count their own information three times, to avoid being biased against themselves.

I will use similar strategies to deal with success-based bias. Agents in success-based bias models will count each signal received once for each time previously the signaling agent has signaled successfully, to a minimum of one count. As with identity bias, success-based bias requires changing how an agent's own information is counted. If the agent is less than half as successful as the most successful agent, it will count its own data twice as much as it would count another agent at its success level. Otherwise, it counts its own data as much as the most successful agent.

Difference can be modeled in two different ways. First, the best action for particular agents can be flipped. This alone should not make a real difference, as the names of the different states and actions is actually arbitrary, so as before, it is necessary that these agents not know that their best actions have been flipped. In other words, when counting their own data, they will count it towards the urn that is less likely to have them take the action that is best for them. What counts as a "good party" changes for some neurodivergent people, who may, for example, be sensitive to a party being loud. If someone does not know this about themselves, it may play a role in how communication about parties develops for them.

The other way will force certain starting urns, to again simulate differences in communication style. If such a difference can be read as neurological and not merely cultural, then a good way to simulate the result of that is to have some agents' starting urns favour one configuration strongly, and the majority slightly favour the other configuration. It is possible for the end configuration to be anything, but it will be interesting to see which agents are favoured and in what ways.

It will again be useful to see agents choosing whom to communicate with. Unlike earlier, agents will not be on a network, so this requires new structure. Agents will have one

more urn in addition to a number of signal urns and action urns, which I will term the “listen urn.” It has balls labelled to correspond to each agent in the model, aside from themselves. When receiving signals, the agent will pull balls until they have pulled a predetermined number of balls, and for each pulled ball, count the corresponding agent’s signal. Then, if the agent acts successfully, in addition to reinforcing that action by placing a number of balls with that action’s number into the action urn they used based on learning speed, it will reinforce listening to all agents it listened to by placing a number of balls with those agents’ numbers into the listen urn based on that same learning speed. As well, if the model is with forgetting, the agent will forget balls from the listen urn at the same rate as other urns, again not removing a ball if it’s a particular agent’s last ball. This way it is possible for agents to start to ignore other agents by pure happenstance of not listening to them when successful, no matter how helpful their advice actually was, but one agent helping another will make it more likely that other agent will listen to it again.

Finally, I wish to look at systems where not all agents are able to cooperate. For larger number of agents, I will look at a model type where the best action for any given agent will depend only on their own observation and the observations of the two nearest-numbered agents. For example, agent 1’s best action will be the one corresponding to the state observed most often between agents 0, 1, and 2. Agent 0 and the highest-numbered agent will be considered right next to each other for this purpose, as if they were all in a circle. However, unless another model type is added and changes this, they will all still be communicating with each other, so that agent 1 has to decide what to do based not only on information from itself and agents 0 and 2, but potentially from a number of other agents giving irrelevant information as well. It will be most interesting to see how this combines with the ability to choose which other agents to listen to, though I will run it on its own to at least establish a baseline as well.

3.3 Equilibrium Analysis

Before concluding the chapter, it is worth analyzing the equilibria of the model I have outlined. This kind of simultaneous signal model is novel, so analysis is needed to explore its equilibria. Such an analysis will be more involved than usual due to the large, variable number of agents simultaneously involved in the game the model is built around, but the equilibria themselves are not overly difficult to understand. On the other hand, an analysis of the evolutionary stability of these equilibria is not possible with the same tools one would use for a 2-agent model. Computational results are therefore still necessary to characterize the likelihood of landing at any of these equilibria, and will of course be further helpful in characterizing the speed at which the models reach equilibrium. That said, this section will provide a useful background for analyzing those results.

3.3.1 Base Model Equilibrium Analysis

In this section I will show that there are three classes of equilibria and describe these equilibria, then take a stab at discussing their evolutionary stability in a similar matter to how the Lewis-Skyrms signal model can be discussed [44]. First, however, I will explain why such a discussion cannot work exactly the way it usually would, and why a best attempt at doing so is unlikely to get very far. The focus of this section is the equilibria themselves, and their evolutionary stability will have to primarily be estimated from the computational results.

The first thing to say is that the formal definition of evolutionary stability that comes from the use of replicator dynamics does not apply. A strategy σ is considered evolutionarily stable if and only if, for all strategies $\mu \neq \sigma$,

$$u(\sigma|\sigma) > u(\mu|\sigma),$$

or

$$u(\sigma|\sigma) = u(\mu|\sigma) \text{ and } u(\sigma|\mu) > u(\mu|\mu),$$

where $u(a|b)$ is the expected utility of strategy a played against strategy b . This means that adding an agent with a strategy other than σ would not lead to a population with a larger number of agents using that strategy, since using the replicator dynamics, that strategy will not be more represented in the next step [50]. However, my model is not using the replicator dynamics, and so a definition based on expected population changes is ill-suited to the task. Importantly, it is unclear what $u(a|b)$ should mean when there are three or more players. What a means is clear, but b cannot just be one strategy. If we define $u(a|b)$ as the expected utility of strategy a when all other players play one strategy b , we logically exclude the possibility of stable situations in which more than one strategy is in play (this will turn out to be a statistical possibility, so we should make sure we are not defining these situations to be formally “unstable” when they meet any reasonable definition of “stable”). Therefore b has to be the set of other strategies being played. The definition has to be adapted, then, as in some places μ has to be replaced by a set of strategies M but in other places there must still be only one strategy $\mu \neq \sigma$.

An analogous definition will, rather than delineating the situations in which the replicator dynamics ensure no other strategy can invade, aim to delineate the situations in which : a set of n strategies $\Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_n\}$ is evolutionarily stable on a model with n agents if and only if $\forall i \in \{1, 2, \dots, n\}$, letting Σ' be the set that results when removing σ_i from Σ ,

$$\forall \mu \neq \sigma_i : u(\sigma_i|\Sigma') > u(\mu|\Sigma').$$

In words, any time you change the strategy of any given agent while fixing the strategies of the others, you do not increase the expected utility for that agent. This definition does not require a disjunction as the previous one did. The disjunction in the previous definition is necessary to look at what happens when the invading strategy is played against the test

strategy *and* against itself, but this distinction does not make sense in this context, and here one equation is able to show all important aspects of what happens when any one agent's strategy shifts. However, this is only simplicity in how the definition is *expressed*; functionally far more inequalities are involved, since a check by exhaustion will require n^2 calculations per alternative strategy instead of one. This combined with the fact that, letting s be the number of possible agent strategies, there are $\binom{s}{n}$ possible combinations of strategies to check, instead of s . So exhaustion of all replacement strategies for all sets of strategies is not a computationally tractable method. In addition to this complexity, it is unclear at this stage whether this definition will even have the same properties as the definition for the replicator dynamics, and it seems very unlikely that it will. Nonetheless it is a start, and we will be able to come back to it later to discuss it as a possible property that an equilibrium might or might not have.

While we could write a program to do the very computationally demanding work of checking all arrangements by exhaustion, a better approach is to find the Nash equilibria for the non-evolutionary game version of the model, and discuss the qualities of these equilibria, since evolutionarily stable setups must be equilibria anyway⁶. This is still an intensive procedure by elimination, as with 16 strategy sets per agent, a 3-agent model requires investigating a 16x16x16 matrix, and the complexity grows exponentially from there. Instead, I will go through classes of strategy-sets together for all numbers of agents.

To do this, I will use the same notation Skyrms uses in *Evolution of the Social Contract*. The four deterministic signaling strategies will be called:

S1: Observation 1 \rightarrow Signal 1; 2 \rightarrow 2

S2: Observation 1 \rightarrow Signal 2; 2 \rightarrow 1

S3: Observation 1 \rightarrow Signal 1; 2 \rightarrow 1

⁶This has been shown for the replicator dynamics [50], and is simple to show for this model. Suppose Σ is a set of n strategies that do not form a Nash equilibrium when used by an n -agent signal model. By the definition of a Nash equilibrium, there exists a strategy μ such that some agent in the model would expect higher utility using that strategy than its current strategy. Then this μ does not satisfy the inequality in the definition of evolutionary stability, and Σ is not evolutionarily stable. Therefore all non-Nash strategy sets are not evolutionarily stable. By contraposition, all strategy sets that are evolutionarily stable are Nash equilibria.

S4: Observation 1 \rightarrow Signal 2; 2 \rightarrow 2

Likewise, the four deterministic receiving strategies will be called:

R1: Signal 1 \rightarrow Action 1; 2 \rightarrow 2

R2: Signal 1 \rightarrow Action 2; 2 \rightarrow 1

R3: Signal 1 \rightarrow Action 1; 2 \rightarrow 1

R4: Signal 1 \rightarrow Action 2; 2 \rightarrow 2

Finally, when combining these, the possible overall strategies will be called:

I1: S1, R1

I2: S2, R2

I3: S1, R2

I4: S2, R1

I5: S1, R3

I6: S2, R3

I7: S1, R4

I8: S2, R4

I9: S3, R1

I10: S3, R2

I11: S3, R3

I12: S3, R4

I13: S4, R1

I14: S4, R2

I15: S4, R3

I16: S4, R4

This completes Skyrms' notation [44]. I will be making extensive use of it, so before moving on, I will also describe the strategies less formally, in service of making it clearer what the names correspond to. S1, S2, R1, and R2 are all strategies that map one input to one output and vice versa, and S3, S4, R3, and R4 are strategies that map every input to the

same output. Remembering that S1 and R1 correspond, as do S2 and R2, is more important than remembering which of S1 and S2 is which, due to symmetry. I1 and I2 represent signal systems when all agents adopt them, the first where observation 1 leads to action 1 in both cases, and the second where 1 leads to 2—I will refer to these as “1:1” and “1:2” strategies, respectively. I3 and I4 represent signalling according to one signal convention and acting according to the other, so in accord with other places I discuss these, I will call them the “mixed strategies,” as this is more concise than something like “French-Spanish strategies,” which is also in line with how I discuss these setups. The remaining strategies all include at least one strategy of either sending only one signal or taking only one action.

Let us start with the obvious, then. Sets of all I1 or all I2, the signal systems, are of course Nash equilibria. If all other agents are using one of these strategies, the expected values of the other strategies are all lower, since they require the agent to either not give signals that are leading the others to act better or not take actions that are informed by the signals the other agents give.

Less obviously, an equilibrium occurs if all agents use one of I11 or I16, that is, either every agent is only sending signal 1 or every agent is only sending signal 2. No change in signaling can increase that agent’s utility from signaling, since the other agents do not change their actions in response. It is less obvious that no change in acting can increase that agent’s utility from acting, but this is still true. The agent has some information about the best action, in its own observation, which it counts as a vote towards the urn that is most likely to yield that action. This information ultimately does not help the agent at all. If the agent switches, for example, to mapping signal a to action 1 and signal b to action 2, it will still always take action 1, since it is receiving at least two votes towards that urn, and so its own vote does not help it. There is therefore no change a single agent can make that increases its own expected utility. For this latter reason, notice that the receiving strategies for these agents does not matter at all; as long as all agents are choosing the same signal, any receiving strategy results in always choosing the same action anyway, since

the same signal majority is always met. Therefore, all systems made entirely of strategies using either only S3 or only S4 are equilibria, meaning all combinations of only I9-12 and I13-16.

Now consider mixes of S3 and S4: there is a specific subset that allows an agent to increase its expected utility from acting. Specifically, this is the subset in which an agent can use one of two arrangements of action strategies to allow its own vote to be a deciding factor. This occurs in an n -agent system whenever, for each signal, at least $\frac{n}{2} - 1$ agents always give that signal.

Lemma 3.3.1. *Any strategy set such that, for each signal, at least $\frac{n}{2} - 1$ agents always send that signal regardless of input, there is some agent that can influence its own action urn choice in its best receiving strategy.*

Proof. Suppose Σ describes such a strategy set. Choose any agent that always gives the most-represented signal (meaning either signal in case of a tie). One of two things is true: either the agent sees an exactly equal number of each signal each time, which occurs in odd-population models because it now receives $\frac{n}{2} - 0.5$ signals of each kind, or the agent sees one more of one kind of signal than the other, which occurs in even-population models since either the total number of signals is even and the agent does not see its own, or the total number of signals of the agent's own kind is two more than the other, and it does not see its own, therefore seeing $\frac{n}{2} - 1$ of one kind and $\frac{n}{2}$ of the other. In the first possibility, the agent can switch to any strategy where one signal maps to action 1 and the other maps to action 2. The agent now always acts in accord with its own observation. This changes the agent's expected utility due to acting from 0.5 to the probability that its observation matches the best action, which is > 0.5 for all finite n since there is an even number of 50% chances of each possibility for the other agents' observations, and the agent has 1 utility whenever the result is mostly the same as its or is tied, and 0 whenever the result is mostly not the same as its. Therefore, changing to these strategies increases its expected utility. In the second possibility, the agent again can switch to any strategy where one signal maps

to action 1 and the other maps to action 2. One of two things can occur. Either the agent makes the observation that points it towards the urn corresponding to the minority $\frac{n}{2} - 1$ received signals, in which case its choice of urn is an even random chance, and therefore resultant utility is 0.5, or it makes the observation that points it to the urn corresponding to the majority $\frac{n}{2}$ received signals, in which case it definitely chooses that urn. Again this yields > 0.5 expected utility for all finite n . Recall that there are an odd number of other agents. There are equal chances that the difference in number of other agents' observations of each kind is greater than 1 in either direction, so the expected utility looking at only this subset is 0.5. Consider the two also equally likely situations in which the difference in number of other agents' observations of each kind is exactly 1. If the more-observed state is the same as the agent's observation, its utility is 1. If the more-observed state is the other possible observation, its utility is 0.5, since the model chooses a best action at random. Therefore, its average expected utility in this subset is > 0.5 , and since it is 0.5 in all other situations, the total expected utility is > 0.5 . \square

This result is useful for the corollary that all systems using only S3 and S4 strategies that do not meet the above condition are equilibria, since the dominant signal again drowns out any action-urn choice. Furthermore, any system in which at least $\frac{n}{2} + 1.5$ agents are in agreement with only one choice for signal will be in equilibrium even when one to all of the remaining agents are mapping one to the other within that minority, since these activities cannot influence action urn choice. Therefore, no choice an agent makes can impact its expected utility from this choice. Overall, call this "the $\frac{n}{2} + 1.5$ equilibrium condition."

Next, for the same reason that this is true of the Lewis-Skyrms model, no pure system containing, as Skyrms says, only either "half" of a signal system can be an equilibrium, because any agent can do better by changing the other half of their strategy to match the first half [44]. This precludes I3-10, I13, and I14. The only remaining one-strategy systems have already been shown to be equilibria.

This leaves systems with mixed strategies. Notice that we are now interested only in systems not meeting the $\frac{n}{2} + 1.5$ equilibrium condition, as we already know those that meet this condition are equilibrium systems. Notice that for all of these systems, each round, there is some chance that any given agent will have the option of influencing some agent's action urn choice depending on what signal it sends. It follows that, for all of these systems, it is necessarily better for each agent to use one of R1 or R2.

Lemma 3.3.2. *In any system not meeting the $\frac{n}{2} + 1.5$ equilibrium condition, every agent gets higher expected utility from at least one of R1 or R2 than either of R3 or R4.*

Proof. Suppose Σ is a strategy set describing such a system. By lemma 3.3.1, there is an agent such that its best choice of receiving strategy allows it to influence its own action urn choice. In the average of cases where it does not do so, R1 and R2 are never both worse than the 0.5 expected utility R3 and R4 always represent. This is because the expected utility of R1 and R2 sum to 1, since they always represent taking the opposite action from the other, which means in any given round, one strategy scores 1 and the other scores 0. So, if R1 is worse than 0.5 on average in these situations, R2 must be better than 0.5, and vice versa. Because, also according to lemma 3.3.1, the agent does better when it can influence itself, then even when both R1 and R2 have expected utility 0.5 when only considering the other agents' information, it is still better to choose one of these than R3 and R4. \square

It follows that there are no equilibria not meeting the $\frac{n}{2} + 1.5$ equilibrium condition that use R3 and R4 at all.

There is one equilibrium not meeting the $\frac{n}{2} + 1.5$ equilibrium condition. In systems with even n , suppose $\frac{n}{2}$ agents use I3 and the other half use I4, the two mixed-system strategies. Each I3 agent sees more other agents signaling according to S2 than S1, and more other agents acting according to R1 than R2, and vice versa for I4. Each of these agents therefore has its highest expected utility keeping to its current strategy, meaning it is in equilibrium.

The similar situation of all agents having a 50% chance of choosing S1/S2 and a 50%

chance of choosing R1/R2 is not an equilibrium, however. By lemma 3.3.1, each agent could do better by arbitrarily choosing between R1 and R2. Then each other agent would have an incentive to choose S1 or S2, respectively, and so on. The same is true of any other probabilistic-strategy system. For each agent, either it does not matter what its receiving strategy is, which cannot be true by the above argument, or it is better to lean towards one or the other. If there is a reason for an agent to lean one way, then it again does better by increasing the power of its own information and picking deterministically. Therefore, no probabilistic equilibria exist.

I will now show that there are no more mixed equilibria.

Lemma 3.3.3. *If it is best for any one agent to use R1 over R2 or R2 over R1, and they are not in the mixed equilibrium, then it is either a signal system or not an equilibrium.*

Proof. Suppose there is a system of n agents such that no agent uses R3 or R4, the system is not an even split of I3 and I4 agents, and that there is some agent A that has higher expected utility using R1 than R2. If any agent acts non-deterministically, the system is not in equilibrium. Suppose none does. Now suppose A uses R2. Then the system is not in equilibrium, since A can increase its expected utility by changing strategy. Suppose instead that A uses R1. Let A' be another agent in the system. Suppose A' has higher expected utility using R2 than R1. Again, if A' does not use R2, the system is not in equilibrium. Suppose, then, that A' actually uses R2. Because there is a difference in expected utility for A between R1 and R2, it follows that they cannot both have expected utility 0.5, and therefore A is choosing action urn with better than random chance. No matter how many S3 and S4 agents there are, the number of agents other than A that use S1 must be greater than the number that use S2, since the S1 and S2 agents are the ones influencing whether A chooses the best action urn, and by the definition of A these agents must on average influence A towards choosing urn 1 when the best action is 1 and towards choosing urn 2 when the best action is 2. If A' does not use S1, then the same is true of it, which contradicts the definition of A' ; therefore, A' uses S1, A uses S2, and the total number of agents using

each signal strategy is the same, resulting in A' seeing one more agent using $S2$ than $S1$. By symmetry, either the system is not an equilibrium, or each agent using $S1$ also uses $R2$, and each agent using $S2$ also uses $R1$. Suppose the latter is true. Because the system is not an even split of $I3$ and $I4$ agents by definition, either the system is not an equilibrium, or some agent uses $S3$ or $S4$. If one such agent uses $R1$, then by symmetry to A , either the system is not an equilibrium, or that agent prefers $S2$. Therefore no such agent uses $R1$ if the system is an equilibrium. If one such agent uses $R2$, then by symmetry to A' , either the system is not an equilibrium, or that agent prefers $S1$. Therefore no such agent uses $R2$ if the system is an equilibrium. Because no agent is using $R3$ or $R4$, the system is not an equilibrium.

We have now shown that if any agent has higher expected utility using $R2$ than $R1$, the system is not an equilibrium. Therefore, either the system is not an equilibrium, or all agents have higher expected utility using $R1$ than $R2$. If all agents use $R1$, then all agents have higher expected utility using $S1$ than $S2$. Therefore, the assumptions imply that either all agents are using $I1$ or the system is not an equilibrium. By symmetry, any one agent preferring $R2$ implies that the only way for it to be an equilibrium is if it is a signal system as well. □

Lemma 3.3.3 implies there are no more equilibria, because the antecedent describes the remaining unexplored strategy space, and the consequent is a disjunct between an already defined equilibrium and non-equilibrium. In conclusion, then, there are three classes of equilibria: signal systems, mixed $I3/I4$ systems, and systems where either $S3$ or $S4$ is used by at least $\frac{n}{2} + 1.5$ agents.

I will now briefly describe the formal evolutionary stability of these equilibria. Signal systems are patently evolutionarily stable. Any agent that attempts to break convention loses its power to communicate with the others, and thereby lowers its expected utility. Mixed $I3/I4$ systems are stable according to the above definition, because each agent cannot do better by changing its strategy. However, unlike the signal system (assuming $n > 3$),

notice that if one agent by chance changes its strategy to one not using S3, S4, R3, or R4, then other agents are also incentivized to change. Therefore, though it meets the first-attempt stability criterion above, there is a sense in which the equilibrium looks like it would be unstable compared to signal systems with fast enough mutation. If learning is slow enough, there might be a region of drift the system mostly stays within once it reaches this equilibrium, but at this stage it is hard to tell how slow learning would have to be, or how long it would take to drift outside of this window anyway, as nothing seems to be forcing the system to stay within the window, so informally, the equilibrium appears unstable. Note that this counterexample seems to show that the above criterion does not have the properties of the replicator dynamics version. Finally, $\frac{n}{2} + 1.5$ equilibrium condition systems are unstable, since no agent's signaling strategy matters, and therefore a large number of alternate strategies μ exist for which $u(\sigma_i|\Sigma') > u(\mu|\Sigma')$ does not hold on account of each side of the inequality being equal.

3.3.2 Altered Model Equilibrium Analyses

I will also look at altered versions of the model that introduce a number of complicating factors. The strategy here will be to look at each such factor and determine whether any of them could interrupt any of the arguments above.

None of the complications that change the learning dynamics make a difference, as the above arguments do not use learning dynamics; some of the equilibria may become informally less stable, but the details of how this would work are going to be complex enough that statistical discussion is the best way to tease out how. This eliminates the need to discuss bias, winnings, choice, and starting urn models, leaving only neighbour and difference models.

The difference model changes the model by including at least one agent that does not use its own information correctly. For this model type, we cannot use the assumption that every agent is better off with a receiving strategy that allows it to influence its own action

urn choice using its own information. However, a model in which all agents do this would not be a “difference” model, so we can still use the assumption that this is true of *some* agent. This means that the proof that all agents in mixed-strategy systems that do not meet the $\frac{n}{2} + 1.5$ equilibrium condition are better off using R1 or R2 does not run; agents with the difference tag are not better off influencing their own action urn choice, and if the number of S1 agents it sees is no more than one more than the number of S2 agents it sees, then using R1 is not actually better for it than R3 or R4 (and vice versa). If we change the proof that follows it so that the only agents that are using R3 or R4 have the dif tag instead of having no R3/R4 agents, the proof gets tripped up at the step that agents using S3 or S4 must have the same of R1 or R2. This adds a way for a system to be in equilibrium.

Before I describe this class of equilibria, however, one condition that was ignored above must be introduced, since there will now be an equilibrium using S3 and S4, where agents are able to sometimes choose their action urn. These systems are only in equilibrium if the numbers of S3 and S4 agents differ by no more than one. If they do, any of the agents in the larger group of S3/S4 can increase its expected utility from signaling as easily as switching to the other, as this increases the percentage of rounds in which the other agents will be able to impact their own choice of action urn without cost. Call this “the even-noise condition,” because S3 and S4 are essentially sending noise, and this condition imposes that the amount of noise of each kind evens out.

For difference models, if the majority agents are an even split of I3 and I4 agents, and the difference-tagged agents are all some mix of I11, I12, I15, or I16 (only using S3, S4, R3, and R4) that meets the even-noise condition, then the system is in equilibrium for the same reason that an even split of I3 and I4 agents is in equilibrium. This equilibrium does not meet the first-blush formal stability condition because each of the difference-tagged agents does just as well by changing to any other signal strategy, and is informally unstable because if such an agent switches to S1 or S2, then the agents using the opposite receiving strategy are also able to drift, and if more than one does so, will be incentivized to switch;

therefore, just from drifting, the system can start to be pulled to a different equilibrium, which is not what we usually think of as stable. Again, however, a qualitative sense of what this degree of instability looks like will have to come from the computational analysis.

The neighbour model is a more substantial change; it rewrites the expected utility function for each agent when the number of agents is $n > 3$. The new utility function only cares about the agent's neighbours, that is, the agents whose numerical ID are one higher and lower mod n . Every equilibrium described above is still an equilibrium for the neighbour model, though in one case there is an added condition. While the reasoning runs the same for the other equilibria, in the I3/I4 split, no agent can have two neighbours with the same strategy as itself, because otherwise it would improve its expected value by swapping its signaling strategy and therefore increasing the action success rate of its two neighbours at the expense of the system as a whole for that step. After this happened, there would also then be a cascade of better-choice swaps that would lead to a signal system.

Likewise, all of the pure-strategy non-equilibrium arguments and the probabilistic-strategy non-equilibrium argument run the same. The remaining argument is the mixed-strategy non-equilibrium argument. Fixing agent A 's best receiving strategy intuitively fixes less about the system as a whole; rather than tell us the balance of all other agents' signal strategies, the agent's best strategy being either R1 or R2 tells us that there is not a majority of other agents that all use one of S3 or S4, and that both its neighbours use S1 or that both use S2, respectively. The condition becomes, then, that the system is an equilibrium solution any time every agent's strategies match their neighbours'. An agent's strategy matches its neighbours if and only if two of these conditions fit: (a) either both neighbours use S1 or use S2 and the agent uses R1 or R2 respectively; (b) vice versa; (c) each neighbour uses a different one of S1 or S2, and the agent uses either R1 or R2; (d) one neighbour uses either S3 or S4, and the other neighbour uses either S1 or S2, with the agent matching R1 or R2 to the S1/S2 neighbour; (e) both neighbours use S3 or S4, and the agent uses either R1 or R2; (f) each neighbour uses one of R1 or R2, in which case the

agent can use any signal. Note that there is no option where an agent's neighbour uses R3 or R4, because then that neighbour would preclude the system from being in equilibrium. This describes a very wide range of new equilibria.

Lemma 3.3.4. *For any set Σ of n receiving strategies only including R1 and R2, in any order, some number of sets of signal strategies for those same agents exist that is an equilibrium.*

Proof. Let R be a set of n receiving strategies that are each other R1 or R2, describing the strategies actually used by a system of n agents in a circle using the neighbour model, such that R contains at least one R1 and at least one R2. If the system is in equilibrium, each agent with two neighbours of the same receiving strategy has the corresponding signal strategy. Therefore, let there be a set S of n signal strategies such that each agent that has two neighbours with the same receiving strategy has the corresponding signal strategy. There is at least one agent the neighbours of which use different receiving strategies. Let A be one such agent. We must be able to arbitrarily choose S1 or S2 for this agent, since each of A 's neighbours either already have another neighbour that has a signal strategy that justifies their receiving strategy—since if that neighbour has a signal strategy fixed already, it is a strategy matching that agent's receiving strategy as it was fixed by the neighbour being surrounded by that strategy—or has another neighbour without a strategy fixed. Therefore the neighbour will either meet condition (a) or (c). If both neighbours-of-neighbours already have a strategy fixed, this was a valid choice of signal strategy. If not, let A' be one of its neighbour-of-neighbours without a signal strategy. Either A 's signal strategy and the intermediate agent's receiving strategy fixes what A' 's signal strategy must be, or one can be chosen arbitrarily for the same reason. Choose a valid signal strategy for A' . Repeat the process for the next neighbour-of-neighbour over, until all neighbours-of-neighbours in either direction have signal strategies fixed. Then choose another agent without a signal strategy and repeat until all agents are assigned a valid signal strategy, or an agent is found which does not have a valid choice. In the former case, the resultant system is an

equilibrium, because each agent's strategy was chosen to maximize utility. In the latter case, call the agent B . One of B 's neighbours must fail all of the conditions regardless of choice of $S1$ or $S2$. By the definitions of the conditions, they must be an $R1$ agent whose other neighbour uses $S2$, and the other must be an $R2$ agent whose other neighbour uses $S1$, since any other configuration has a valid choice between $R1$ and $R2$. Neither of B 's neighbours-of-neighbours had their signal strategy fixed by being surrounded by agents of the same receiving strategy, since their signal strategies do not correspond to one of their neighbours' receiving strategies. Therefore, both neighbours-of-neighbours had their signal strategies chosen due to an arbitrary choice made by the above process. It is therefore possible to backtrack to the last arbitrary choice the process made and reverse it. Then continue the process until it is back to B . If the signal strategy of the chosen neighbour of B 's neighbour is the same after doing this, another arbitrary choice was made after the previous one was changed. Repeat changing this choice until that agent's signal strategy is different. It must be possible to do so, because the only way to run out of arbitrary choices is to return to the first one, which cannot have been fixed by its neighbours, and therefore cannot meet the requirement met by B . This means both of B 's neighbours-of-neighbours now have the same signal strategy, and it is no longer impossible to choose a valid signal strategy for B . This process is now robust enough to guarantee a set S of signal strategies that results in an equilibrium when combined with R .

If R consisted of only $R1$ or only $R2$, then the set S of only $S1$ or only $S2$ respectively gives an equilibrium because it describes a signal system. Therefore, a response S must exist for all R such that the combination Σ describes an equilibrium. \square

The set of all combinations of (not exclusively) $R1$ and $R2$ and all possible resultant sets S of signal strategies from the above process does not exhaust remaining equilibria. An equilibrium found from the above process may result in some agents with signal strategies that are not forced by its neighbours. That is, some agent A such that both of its neighbours-of-neighbours have signal strategies that are the same value as the agent they share as

neighbours with A . In addition to the other equilibrium the above process could have found by making the other arbitrary choice, replacing this strategy with $S3$ or $S4$ would also result in equilibria. Therefore, another process going from equilibrium to equilibrium could exist, arbitrarily choosing an agent with free choice of signal strategy and arbitrarily choosing one of $S3$ or $S4$ such that the even-noise condition is still met. This process reminds me of testing blocks in a Jenga tower to see if they are loose, since that process also sometimes changes which other blocks are loose, so call these the Jenga equilibria. Since this extends to the above process not involving $S3/S4$ agents, I will call the above process “setting the Jenga tower” for the sake of having a name that clearly covers its entire class.

I do not have a proof that the space of Jenga equilibria exhausts the remaining space of equilibria; it is *prima facie* unlikely that an equilibrium exists that for some reason could not be reached via this method, but I do not see a clear way to show this definitively. At any rate, such an equilibrium would still have to look similar to the results of the above processes, so I count them as part of the same class.

The Jenga equilibria are a new class of equilibria, so we must ask if they are evolutionarily stable. They are not. If no agent exists which can freely change its signal strategy, then no choices were arbitrary, and the system is already either a signal system or $I3/I4$ split, and therefore not part of the Jenga equilibria class. If one does exist, then the stability inequality is not satisfied.

Finally, there is the combination neighbour-difference model. The only difference here is that the process of setting the Jenga tower cannot assume that no agents use $R1$ or $R2$. The fix is very simple⁷. Only a specific subset of agents have the difference tag. For each of these agents, repeat the process fixing them such that they are using an even combination of $R3$ and $R4$, and again with each agent that is able to, also choosing like a Jenga tower, moving the agent back to $R1$ and once again $R2$, while also changing the neighbouring agent’s signal strategies to both $S1$ or both $S2$, respectively. These agents will therefore not

⁷Thank God for small miracles.

be thwarting the equilibrium, and the rest of the process runs the same as written above. So the combined model still just has the Jenga equilibrium class, it is just expanded to include some R3 and R4 agents sometimes.

Chapter 4

Signal Models of Hermeneutical Injustice

The previous chapter described signal models as a formal tool for modeling communication, and motivated their use for modeling hermeneutical injustice. This chapter explores a computer model designed for this use, and the conclusions one should come to based on the results of that model.

As signal models come with a built-in learning problem, I will not need to define a new one here. Instead, I will need to devote additional space to describing the results of the model as modified for the present purposes even before introducing specific possible causes of injustice, after I explain the mechanics of the model. I will then explore results that suggest, alongside the results of chapter 2, cultural-evolutionary pressure can cause systemic ableism in communication norms, but also represents a possible solution to that systemic ableism. I will argue that this pressure is best created through harmony of means and ends, where activists advocate for a future that uplifts disabled people in particular, but ideally not at the expense of majority groups, whose minds and hearts can best be won by improving their conditions as well and via the same action. That said, the results also provide guidance for those who may have more faith in more direct negative pressure.

4.1 Methodology and Results

4.1.1 The Code

Once again I used the Python library Mesa, this time to program a series of computer models based on the simultaneous signal model I described in the previous chapter. I ran it on one thousand networks for ten thousand steps per network, logging data at 1, 10, 100, 1 000, and 10 000 steps, and I have made this code available as well in appendix A. I will now give an abstract description of how it works.

To reiterate, my signal models involve some number of agents attempting to aid each other in taking the correct action using signals. Each agent sees one piece of information, either a 0 or 1, which in the running example corresponds to evidence an upcoming party will be bad or good, respectively. Each agent is also to choose between action 0 or action 1, corresponding to staying home or going to the party. Action 0 is best if most agents saw information 0, and action 1 is best if most agents saw information 1, with a coin flip determining ties. In order to help each other act best, each agent will also take a signal action, either action *a* or *b*. An agent decides which to do by keeping two urns full of balls labeled “a” and “b,” with each urn corresponding to one piece of information. The agent goes to the urn for the information it saw, pulls a ball at random, and does the signal action corresponding to what is written on the ball. Likewise, agents receive each other’s signals, and pick from a second set of urns based on their own information and which signal they received the most of, again settling ties at random. Finally, they act based on the ball pulled from the urn, and add balls to that urn corresponding to that action if it was the correct one, and to their signal urn corresponding to their signal action based on how many others were successful. The end result, at least in the base version of a signal model as described in this paragraph, is that over time the urns take on a makeup that allows for the agents to coordinate better than if they were acting alone on average.

To implement this, the code defines, again, a number of objects and functions, those

being the agents and model on the one hand and the processes they take, as well as processes for collecting data, on the other. Once again I am using NumPy's random number generators with random seeds given uniquely to each agent and each model, in a systematic repeatable way. Therefore, each type of model can be run through the same sets of agents with the same starting urns and random choices throughout. One advantage of this for the network model was that we could look at one network that turned out a particular way in one run and a different way in another between the two runs, and track which differences were making a difference and in what way, since each network had the same structure between each different type of model run on it. This advantage still exists, but to a much lesser extent; the starting conditions are still generated the same way for each model type, but these starting conditions have less of an influence on outcomes than before, or at least a more chaotic influence. The reason for this is that network formation was one random event that would create a lasting structural effect on the rest of the model; with the exception of mingling models, if agent 0 and 1 were linked in network 5 in one model type, they would be linked later in that model, and they would be linked in any other model type. There is no analogous structure in a signal model. Each agent is communicating with each other agent, with some exceptions below that allow agents to change how much weight they give particular agents' signals. Therefore a lot of the power granted by this use of NumPy is lost. However, it is not entirely lost. Recall that it was still possible in the previous model to track individual networks through mingle models. Everything I was able to say about mingle networks was possible despite this loss in power. It follows that the same tracking of individual sets of agents through different kinds of models will still be powerful enough here to give worthwhile conclusions; where previously I was looking at what changed other than the network structure, I now will look specifically at what changed other than particular random decisions, especially near the start.

Each agent object has a number of features that are tracked throughout. Most importantly, each agent has a pair of lists, standing in for the urns. An Urn list contains two lists

of two variables. Each of these two lists represents one urn, and the variables represent how many balls of a particular type are in that urn. When drawing from an urn, an agent generates a random number from zero to one, and if that number is less than the proportion of balls in the zero position, a ball marked “a” is drawn, and if it is greater than that proportion, a ball marked “b” is drawn. Agents also have features that help determine how to behave based on type of model and identity marker, which as before are determined by parameters given in each type of model, and in one case below will also have a fifth urn for tracking their trust in other agents.

The output of the code is once again a massive database formatted by the Pandas library. This time the information being tracked is more extensive, including average number of imperfect signal systems per run at various degrees of imperfection, how likely agents were to communicate successfully, including a separate value just tracking agents with identity markers, variance in that communication success rate, how many successes agents had in signaling and acting, and which agents were listened to the most in model types that allow for difference in listening.

I am measuring one-way communication success according to the following equation,

$$C(a, b) = \frac{1}{n} \sum_{x=1}^n \sum_{y=1}^s U_{a1}(x, y) U_{b2}(y, x)$$

where $C(a, b)$ is the communication success rate of agent a communicating to agent b , n is the number of natures in the model, s is the number of signals in the model, and $U_{ij}(x, y)$ is the percentage of balls of kind y in the urn belonging to agent i of type j , $j = 1$ corresponding to signal urns, the first kind, and $j = 2$ corresponding to action urns, the second kind, corresponding to input x . In words, it is the percent chance that agent b would act correctly in a dyadic signal model with agent a , assuming the kinds of observation are

equally likely. Two-way communication success, then, is

$$C_2(a, b) = \frac{C(a, b) + C(b, a)}{2}$$

$$C_2(a, b) = \frac{1}{2n} \sum_{x=1}^n \sum_{y=1}^s U_{a1}(x, y)U_{b2}(y, x) + U_{b1}(x, y)U_{a2}(y, x)$$

or the average of the two one-way communication success rates between the two agents.

I measured two-way communication success rates only as a statistic in itself. I used one-way communication success rates to determine how close to signal systems the systems were. If two agents have one-way communication success rates neither of which are below p , then I will say that they form an imperfect signal semi-system of degree p . If, for an entire model, there is no one-way communication success rate from one agent in the model to another in the model below p , I will say they form an imperfect signal system of degree p . No imperfect signal system can have degree 1, as then it would be a regular signal system. However, since agents will not forget their last ball of a particular type, perfect signal systems are not possible using this code, and I can safely speak only of imperfect signal systems, and from now on drop the descriptor “imperfect.” To be more plain, a signal system of degree p is a group of agents that always have at least a probability of p of understanding each other. I will specifically track the numbers of signal systems and signal semi-systems of degree 0.95, 0.9, 0.85, 0.8, 0.75, and 0.5.

Finally, the code can run a number of different types of models, using different features as described in chapter 3. I have already described how these features will work in theory, and there is little worthwhile to say about implementing them. It is worth noting the numbers of the different starting urns in the “starting urn” model type. Tagged agents will have their randomly generated starting urns replaced with urns that have one ball of each type, plus a number of balls of a particular type equal to the normal starting urn size for the model. In this case, the observation 0 starting signal urn starts with six “a” balls and one “b” ball, and the observation 1 starting urn starts with one “a” ball and six “b” with the

corresponding a urn having six “0” balls and one “1” ball and b urn having one “0” ball and six “1” balls. Therefore, if there are multiple agents with these differing starting urns, those agents will start with a signal semi-system of degree 0.7551 instead of with random starting urns. At the same time, other agents will have two balls added to their urns corresponding to the opposite configuration, so signal urn 0 will have “b” balls added and so on. This is done to make the other agents less likely to understand the agents that were changed. In all other model types, starting urns are randomly generated by putting one ball of each type in, then a number of additional balls as needed to get them to the same predetermined starting number, with each ball’s type chosen randomly with even chance across ball types, and independently.

4.1.2 Results: Base Model

Because this kind of signal model is novel, more discussion of the baseline is needed than for the previous model.

I first tested parameters on the baseline model for various numbers of agents in order to set those parameters for the other model types moving forward. In these preliminary tests, I only ran 100 runs each, instead of 1000, as there are far more possible parameter combinations, and I did not intend to draw important conclusions from them. Models that involve between two and five agents inclusive all have a subset of parameters for which they almost always converge to a high-degree signal system. From six agents on it is a lot harder for agents to *all* come to the same signaling convention. At seven agents, for example, the best outcome was 27% of runs resulting in a degree .95 signal system, which occurs when starting with 3 balls in each urn, adding 3 balls when learning, and losing 2 balls when forgetting. I chose two sets of parameters that, for the numbers of agents I used, one of which would always result in a rate of convergence not significantly lower than the highest. Those sets were a starting urn size of 3, a learn speed of 3, and a forget speed of 2, which works for every number of agents up to and including five,

and a starting urn size of 5, learn speed of 5, and forget speed of 3, for all numbers of agents higher than five. See tables 4.1 and 4.2 for a partial overview of tested parameters. The reason for the specific parameter choices involved a holistic look at semisystems and average communication percentage as well as the data in these tables. Not reflected in the tables is that I also tested a six-agent model on the 3/3/2 and 5/5/3 parameter sets to determine whether five or six agents is the point at which the better parameter set changes, and found that 5/5/3 is much better for six agents. Rather than discuss the tradeoffs of other possible choices, I will note that the decision is relatively inconsequential as long as the parameters allow some difference in outcome to show up in the statistics, and move on. The only real worry is that strange parameters might cause artifacts in the data, but if something in the data is especially surprising, it is always possible to try again with different parameters.

There is not much of interest that the parameter results tell us, other than that these parameters do matter a lot for whether a model is viable. A signal model where agents never forget will struggle even from three agents, most signal models where learning is only marginally faster than forgetting can do well up to five agents, and starting urn size is less important than these other parameters. None of this is surprising. In order to succeed with more than two agents, agents need to be correcting mistakes that get made, only a little slower than they learn. When there are more agents, there does need to be an increase in the ratio of learning to forgetting, because something needs to get through the noise of all the other agents, but relatively quick forgetting is still necessary, and either way nothing is that successful. This helps demonstrate why it is important to go beyond equilibrium analysis for this kind of model; for some parameters, it is very rare to settle on one equilibrium.

I now move on to the results for 1000 runs at the chosen parameters for 3, 4, 5, 6, and 10 agents. At this higher number of runs none of the numbers of agents actually hit 100% convergence. 99.3% of 3- and 4-agent models had degree .95 signal systems, and 98.7% of 5-agent runs did, with average agent communication percentages of 0.997,

Parameters	3 agents	4 agents	5 agents	7 agents	10 (0.9)	15 (0.5)
2, 3, 0	0.07	0	0	0	0	0
2, 3, 1	0.36	0.01	0.05	0	0	0
2, 3, 2	0.98	1	0.98	0.23	0	0
2, 4, 0	0.08	0	0	0	0	0
2, 4, 1	0.26	0.01	0.02	0	0	0
2, 4, 2	0.78	0.26	0.26	0.01	0	0.02
2, 4, 3	1	0	0.84	0.03	0	0
2, 5, 0	0.08	0	0	0	0	0
2, 5, 1	0.23	0	0.02	0	0	0.01
2, 5, 2	0.42	0.06	0.07	0	0	0.01
2, 5, 3	0.98	0.86	0.69	0.22	0.02	0.16
2, 5, 4	1	0	0.27	0	0	0
3, 3, 0	0.06	0	0	0	0	0
3, 3, 1	0.32	0.01	0.03	0	0	0
3, 3, 2	1	0.99	0.99	0.27	0	0
3, 4, 0	0.05	0	0	0	0	0
3, 4, 1	0.29	0	0.02	0	0	0
3, 4, 2	0.7	0.16	0.24	0.03	0	0.02
3, 4, 3	1	0.01	0.75	0	0	0
3, 5, 0	0.08	0	0	0	0	0
3, 5, 1	0.2	0	0.02	0	0	0
3, 5, 2	0.43	0.04	0.04	0.02	0	0.02
3, 5, 3	0.99	0.87	0.68	0.25	0	0.16
3, 5, 4	1	0	0.4	0	0	0

Table 4.1: The parameters tested for various numbers of agents, and the percentage of models that achieved a signal system of degree .95, or lower degree where specified. The parameters are, in order, starting urn size, learn speed, and forget speed. Continued in table 4.2.

0.994, and 0.997, respectively. None of these values are significantly different from each other, so differences are somewhat likely to be due to statistical variance rather than small differences in actual expected values. In other words, given good parameters, convergence is highly likely for 3, 4, or 5 agents, without much difference between them, so long as there are no complications. On the other hand, there are differences. When there are 3 agents, convergence is a lot faster. After 10 000 steps, agents had on average accumulated 9847.13 successful actions, for a 98.47% cumulative success rate. As well, a full 92.5% had already achieved degree .95 systems by step 1000, long before the model was finished

Parameters	3 agents	4 agents	5 agents	7 agents	10 (0.9)	15 (0.5)
5, 3, 0	0.07	0	0	0	0	0
5, 3, 1	0.36	0.01	0.03	0	0	0
5, 3, 2	1	0.98	1	0.22	0	0
5, 4, 0	0.07	0	0	0	0	0
5, 4, 1	0.29	0	0	0	0	0
5, 4, 2	0.72	0.17	0.24	0.01	0	0.03
5, 4, 3	1	0	0.78	0	0	0
5, 5, 0	0.1	0	0	0	0	0
5, 5, 1	0.18	0	0	0	0	0
5, 5, 2	0.38	0.02	0.11	0	0	0.01
5, 5, 3	0.99	0.9	0.74	0.22	0.01	0.13
5, 5, 4	1	0	0.22	0	0	0
9, 3, 0	0.04	0	0	0	0	0
9, 3, 1	0.36	0.01	0.02	0	0	0.01
9, 3, 2	0.99	0.99	0.99	0.28	0	0
9, 4, 0	0.08	0	0	0	0	0
9, 4, 1	0.27	0	0	0	0	0
9, 4, 2	0.76	0.2	0.27	0.05	0	0.03
9, 4, 3	1	0	0.84	0.02	0	0
9, 5, 0	0.11	0	0	0	0	0
9, 5, 1	0.23	0	0	0	0	0
9, 5, 2	0.38	0.02	0.08	0	0	0
9, 5, 3	1	0.87	0.74	0.25	0.02	0.14
9, 5, 4	1	0	0.33	0	0	0
3, 11, 10	1	0	0.04	0	0	0

Table 4.2: Continuation of table 4.1

running. On the other hand, 4-agent systems averaged 75.56% cumulative success rates, and only 2.5% had converged by step 1000. In other words, while they are both as likely to converge by step 10 000, 3-agent systems converge *far* faster than 4-agent systems.

Also of interest is that 5-agent systems were somewhere in the middle, with 95.76% cumulative success rate and 33.1% converging by step 1000. As later results will also indicate, there is a notable difference in the functioning of even and odd numbers of agents for low populations. Even numbers of agents introduce a chance for ties in number of signals, which can introduce noise to the learning process, and slow things down. This is a notable downside to many-agent signal models for modeling reality, because it introduces very vis-

ible artifacts to the results, which one needs to be aware of when drawing conclusions from them. To avoid such artifacts, and because they support more model types, I will be focused on even population sizes.

The high rate of convergence begins to break down at 6 agents, the same number at which a higher learn and forget speed start performing better. 20.4% of 6-agent systems managed above degree .95, with a significantly lower 92.67% communication percentage and 77.10% cumulative success rate. Once problems started I skipped to 10-agent systems. Only 0.4% of 10-agent systems managed to get to degree 0.95, the average communication percentage was 86.17%, and average cumulative success rate was 73.52%. As I will show below, however, the situation is not entirely dire even at this higher number of agents.

The equilibrium analysis indicated that the mixed-strategy equilibrium was dubiously evolutionarily stable. This equilibrium represents the two-language phenomena sometimes observed with 2-agent systems. A valid proper signal system is for one agent to always signal 0 when getting observation 0 and 1 for observation 1, but for the second agent to do the opposite, with each acting in accord with the other's signaling. This is a proper signal system because both agents are able to coordinate perfectly. However, it is not reminiscent of what one hopes for from a signal system, namely something resembling a language. In such a situation, it cannot be rightly said that signal 0 "means" either state/action 0 or state/action 1, since when one agent uses it they are saying to take action 0 and when the other uses it they are saying to take action 1. Again, it is as if one speaks French but understands Spanish, while the other speaks Spanish but understands French [44]. A third party would not be able to learn to communicate with both agents without the ability to differentiate between the two and switch how they signaled with each.

These were vanishingly rare. The closest 3-agent model was run 530. A typical run will have urns with a single-digit number of balls of one kind, usually only one, and between 4500 and 5000 of the other, with all three agents looking the same. Run 531, for example,

has urns on step 10 000 written $[[1, 4889], [4807, 1]]$ and $[[1, 4961], [4822, 1]]$, for agent 0's signal and action, respectively, and the other agents have almost identical percentages. This corresponds to signal 1 on an observed state of nature 0 and signal 0 on observed state of nature 1 99.98% of the time in either case, with corresponding action urns. Run 530 has agent 0 with urns $[[1223, 3], [4, 1389]]$ and $[[1, 723], [1373, 3]]$, which still give above 99% chance of doing one thing, that being signaling alike to observation but acting opposite to signal, which looks like the two-language behaviour described above. Agent 1's urns mostly correspond, being $[[147, 580], [548, 1726]]$ and $[[75, 1], [1, 1906]]$, which means it usually gives signals corresponding to agent 0's actions, and actions corresponding to agent 0's signals, again as described above. One difference is that the signal urns are clearly not very definite, giving a 20.22% and 24.10% chance of not giving the corresponding signals respectively, although the action urns have at least a 98% chance of taking the corresponding actions. As expected, agent 2's urns do not make sense, being $[[2130, 148], [930, 1]]$ $[[1135, 1], [1, 62]]$, which corresponds to generally signaling 0 no matter what, and acting in accord to only agent 0's signals. Because agent 1's signal urns are so mixed it is not really the situation described.

Keep in mind that this is one run out of one thousand; none of the other four runs with degree below 0.5 can be characterized this way. For example, run 612 had two agents that mostly sent signal 1 no matter what and one that mostly sent signal 0 no matter what, with one of the 1-senders having the opposite-to-signal action urn configuration, and the other two having same-as-signal action urn configurations. I have no good explanation for why this happened, even after investigating individual steps, except that this is also a one-in-a-thousand run, and bizarre unlikely outcomes still come up now and again in a large enough dataset. Run 275, also degree below 0.5, looks different from either, with several very mixed urns. In short, I do not think there is any inclination for runs to end up in two-language semisystems, but something superficially like it, like anything, is possible by pure chance.

The equilibrium analysis noted that while the mixed equilibrium technically met the first-blush definition of evolutionary stability, it did not qualitatively appear stable. The results ended up supporting the informal analysis better than the formal analysis; my definition is unsurprisingly poor¹, and it would take significant work to come up with a better formal definition.

To summarize, while 2-agent systems often end up with agents that signal one way and act the other, the existence of additional agents nullifies this possibility, and when there are high-degree semisystems, they are generally using only a single convention that other agents generally adopt, at least up to five agents. Beyond this number, failure to create overall convention is not due to high-degree semisystems excluding others, but rather due to the difficulty of getting additional agents in line with a convention so many agents already adhere to.

I will say more about this difficulty. When viewing the step 10 000 urns for 10-agent systems, the story presents itself very clearly. In general, seven to nine of the agents are entirely successful, aligning to each other's strategies to extremely high percentages. The small minority of other agents, however, will have action urns perfectly in line with the group's chosen convention, but signal urns that have drifted into incoherence. Once nine agents have a signaling convention, there is very little incentive for the tenth agent to signal in that convention. If an agent is sending only signal 0, or has one urn that is still flipping a coin on what to signal, it is essentially generating noise, but 10% noise to 90% signal is a pretty good ratio. With ten coin flips, there is a 24.61% chance of a 5/5 tie, and a 20.51% chance each of a 4/6 split in either direction. Suppose an agent sends only one kind of signal. In half of 5/5 ties, the agent will get the observation not corresponding to this signal in the larger convention, and signal incorrectly, so that to all other agents it appears to be a 4/6. Because 5/5 ties have their best action determined randomly, half

¹This is unsurprising because I noted in this section that the work of giving a proper definition would be substantial if it is possible, and the I made no pretense that the definition I gave would have any positive qualities except as a convenient starting place.

the time this flip will result in the other agents taking the incorrect action. This scenario only occurs 6.15% of the time. Alternately, only one of the 4/6 splits are splits in which the agent can make it appear to be a 5/5 instead, which happens in only 60% of those 4/6 splits, and results in incorrect action for only 50% of agents, which also rounds to 6.15% of the time. In other words, the unconventional agent is causing disruption for other agents 13.3% of the time, and otherwise there is generally above a 99.8% communication success rate. Agents in perfect signal systems act incorrectly because of nature giving them a 50/50 split 12.31% of the time, and due to the way these possibilities overlap, the total signaling success rate for an agent sending only one signal is 81.54% compared to an ideal 87.69%. This small difference in positive feedback received between giving noise and learning the signal convention, and the fact that agents can learn to act according to the convention without learning to signal according to it, is the reason the last few agents often do not learn to signal conventionally.

Once an agent is sending mostly one kind of signal in a mostly ideal system, it is most likely to continue doing so, because it is seeing its peers acting successfully, and has no reason to change its behaviour. The result is that after enough steps 10-agent signal models have reasonable average communication percentages, reasonable average numbers of high-degree semisystems, and a lot of individually successful agents, but almost never form perfect systems. Any individual agent is likely to be able to communicate with any other, aside from a few exceptions without a strong enough incentive to get with the program. So while 10-agent systems look like they are doing very poorly when looking at number of systems, they actually do fairly well, and have high potential to give good information by running other model types on them. One still must be more careful when drawing conclusions from models with higher agent counts to make sure what is being observed is not an artifact of this strong tendency for otherwise successful models to have individual aimless agents.

4.1.3 Results: Other Model Types

The full data linked to in the appendix includes results for signal models with 3, 4, 5, 6, and 10 agents. The 3- and 5-agent runs introduced artifacts to the results not present with even-numbered agent populations. In particular, the bias model type did nothing and winnings next to nothing, since with an even number of incoming signals, weighting the votes differently made no difference; a tiebreaking single vote matters just as much whether the other two voting agents get a larger number of votes or not. See table 4.3. The neighbour model did not work for 3 agents, and as I will note below, behaves strangely for 5 agents. Therefore, aside from a brief note on 5-agent neighbour models that will require more discussion about how the model type works with even numbers, I will focus below on signal models with even-numbered populations. Some statistics for the most relevant signal model types at these agent populations are given in tables 4.4 to 4.6.

	No Bias: Signal 0	Bias: Signal 0	No Bias: Signal 1	Bias: Signal 1
Both signal 0	3/0 (urn 0)	7/0 (urn 0)	2/1 (urn 0)	6/1 (urn 0)
Opposite signals	2/1 (urn 0)	4/3 (urn 0)	1/2 (urn 1)	3/4 (urn 1)
Both signal 1	1/2 (urn 1)	1/6 (urn 1)	0/3 (urn 1)	0/7 (urn 1)

Table 4.3: Urn choice in 3-agent systems with and without bias. Rows are the decisions of two of the three agents, and the column is the decision of the third agent, plus whether or not that third agent is biased against. The signal count given is the count one of the two agents would have assuming the signal they gave lines up with the urn they would be inclined to choose on their own evidence.

To give a baseline, 4-agent systems are highly successful with 99.3% convergence and 99.4% communication, where 6-agent systems are far less successful, with 20.4% convergence and 92.7% communication. At the individual system level, what this looks like is that the majority of 6-agent systems had a large degree of success but generally had a small number of agents that were unable to fully converge to the convention; effectively, there was one dominant convention, and there is a strong sense in which the signals can be said to have meaning for each system, but that these agents are not entirely with the program. Each lower degree barrier for what is considered “convergence” notably increases percent-

age; 31.9% at degree 0.9, 39.0% at degree 0.85, all the way to 90.5% at 0.5. The remaining 9.5% generally had only one or two agents that by statistical unlikelihood were so far from converging as to go in the opposite direction from the clear main convention, in either signaling or, more often, acting alone. In other words, when one remembers the standards for convergence, it is a good result to have only 9.5% of systems in this category, having one or more agents that communicates clearly against one dominant convention.

Model Type	4 agents			
	% .95 sys	% .95 semi	Com %age	Dif Com %
Base	99.3	99.37	99.38	99.38
Bias	8.9	54.45	87.67	75.62
Winnings	19.2	59.55	90.39	90.34
Neighbours	41.3	71.5	80.81	80.77
Difference	1	3.15	81.78	78.43
Difference 2	0	0	51.84	51.77
Choice	0	0.13	52.4	52.24
Urn	99.2	99.43	99.44	99.44
Bias + Diff	0.1	4.88	81.2	63.55
Bias + Choice	0	0.13	52.4	52.24
Winnings + Diff	0	2.02	80.71	77.41
Winnings + Choice	0	0.15	52.36	52.35
Neighbours + Diff	0	3.84	75.59	69.92
Diff + Choice	0	0.01	52.62	52.47

Table 4.4: Statistics from select 4-agent signal models. “mtype” refers to the model type, where “b” is bias, “w” is winnings, “n” is neighbours, “d” is difference, “c” is choice, and “u” is urns. “% .95 sys” is the percentage of systems that had degree 0.95 by step 10000, “% .95 semi” is the same for semisystems, “Com %age” is the average communication percentage, and “Dif Com %” is the same for only agent 0, or in the case of the “d2” mtype, agents 0 and 1, i.e. agents that may have bias or difference tags or different starting urns.

Bias and winnings model types both had a profound effect. As with the base model, changes in learning dynamics can cause a large percentage of systems to fail to equilibrate, which further demonstrates the importance of going beyond equilibrium analysis. In 4-agent, bias had 8.9% convergence and 87.7% communication, while winnings had 19.2% convergence and 90.4% communication. In 6-agent, bias had 4.7% convergence and 86.6% communication, while winnings had 3.9% convergence and 86.9% communication. For

	6 agents			
Model Type	% .95 sys	% .95 semi	Com %age	Dif Com %
Base	20.4	57.8	92.67	92.52
Bias	4.7	46.32	86.58	72.59
Winnings	3.9	43.84	86.91	86.82
Neighbours	3.6	28.99	71.39	71.2
Difference	31.4	6.53	93.92	93.65
Difference 2	0	0	51.07	51.08
Choice	0	15.27	59.41	59.44
Urn	22.8	58	92.52	92.27
Bias + Diff	6.5	5.21	87.59	72.22
Bias + Choice	0	15.27	59.41	59.44
Winnings + Diff	0	2.92	85.13	80.18
Winnings + Choice	0	15.33	59.47	59.28
Neighbours + Diff	0	2.05	68.75	64.28
Diff + Choice	0.1	1.47	59.01	59.1

Table 4.5: Statistics from select 6-agent signal models. See table 4.4 for definitions.

	10 agents			
Model Type	% .95 sys	% .95 semi	Com %age	Dif Com %
Base	0.4	33.93	86.17	86.23
Bias	0	30.09	82.83	68.47
Winnings	0	29.19	83.35	83.63
Neighbours	0	8.99	59.28	59.48
Difference	0.2	3.4	85.97	85.04
Difference 2	0	0.11	50.71	50.78
Choice	0	0	50.68	50.66
Urn	0	33.17	86.01	85.65
Bias + Diff	0.1	3.1	83.24	68.44
Bias + Choice	0	0	50.68	50.66
Winnings + Diff	0	2.75	84.19	83.76
Winnings + Choice	0	0	50.76	50.73
Neighbours + Diff	0	0.82	58.65	55.74
Diff + Choice	0	0	50.73	50.68

Table 4.6: Statistics from select 10-agent signal models. See table 4.4 for definitions.

both numbers, the difference in communication between the two types was not statistically significant. There is an unsurprising qualitative difference in the resulting urns. There is a strong trend for individual bias model systems that do not have high degree to have the biased-against agent be an outlier to an otherwise high-degree system. Winnings models

are much the same, mostly high-degree with an outlier, but with the difference that this outlier is just as likely to be any of the agents. It would be much more surprising for this not to be true, as bias acts mostly on one agent and winnings does not target one more than any other, but this will be relevant later.

The urn model type, however, which starts one agent with radically skewed signal urns and the others slightly skewed against it, did not significantly change results for any number of agents. To be brief, the evolutionary learning dynamics quickly erase the effects of the starting urns.

Adding differing agents, agents that count their own evidence incorrectly, did have notable effects, almost entirely eliminating high-degree systems, with comparatively small decreases in average communication rate of agents. This is another sharp difference between the statistical and equilibrium analyses, since the equilibria changed very little with the introduction of difference agents, only adding a set of unstable equilibria similar to the mixed equilibrium that was not actually observed in the base model. The end result urns tell one story predominantly; that the much lower chance of success for differing agents makes it impossible for them to keep up with the rate of learning, and their resulting action urns stay nearly empty, but trending in the correct direction. On the other hand, they signal in line with the others, which communicate with each other just fine. In other words, if only one agent makes consistent errors about its own interests, the others will generally still be able to create their own communication system, at the exclusion of the error-making agent. Increasing the number of agents tended to make the difference model type perform closer to baseline, as signals from agents learning adequately started to drown out the divergent agents' mistakes. Adding a second divergent agent mostly meant the model was overall unable to function, with the exception of exactly 5 agents.

The choice model type has a massive impact. It almost entirely eliminates high-degree systems, and tanks communication percentage, for most model types very close to 50% or about as good as chance. The reason for this is that very often each agent chooses only one

other agent to listen to, has some degree of success, and then this creates a feedback loop where the listen urn almost exclusively has balls corresponding to one agent. Since each agent has one agent that it listens to, chosen almost completely arbitrarily, the connections being made are without structure, and often agents get left out with no selection pressure to prefer one strategy or another, and cannot learn. Even the best-off agents have little enough information that they cannot take actions with any real degree of accuracy, especially as number of agents grows, and they fare little better. Combining choice with other model types largely does not matter, because this model type is so destructive to ability to learn.

The neighbour model type is very interesting. On the surface, it appears to just be damaging convergence a fairly small amount and overall communication percentage about as much as difference models—at least for 4-agent models, since 6-agents have difference models do worse, but 10-agents have them do better. In short, while it is generally far easier for a model to be successful with this type of difference, systems that do not converge diverge further than unsuccessful difference model systems do. This may have been expected from the equilibrium analysis, since this model type has a large class of additional equilibria that are not conducive to good signaling. However, when one looks at the character of that failure, the picture is notably different. In non-converging systems, agents are generally alternating language use, as occurs in the mixed equilibrium. For example, agent 0 might signal in line with its observation, and act opposite to the signals, and agent 1 would do the reverse, alternating again for agent 2 and 3. A 4-agent system would at this point have agent 3 communicating successfully with agent 0, whereas 6-agent and 10-agent systems would continue the pattern, finding that loop later on. See figure 4.1 for an illustration of such a system. In such cases, communication is very high-degree when looking only at neighbours.

This result may seem very surprising, as without any structure allowing choice of which other agents to listen to, these agents are evolving strategies in line with the signals of only their neighbours despite receiving other signals. As well, the equilibrium analysis

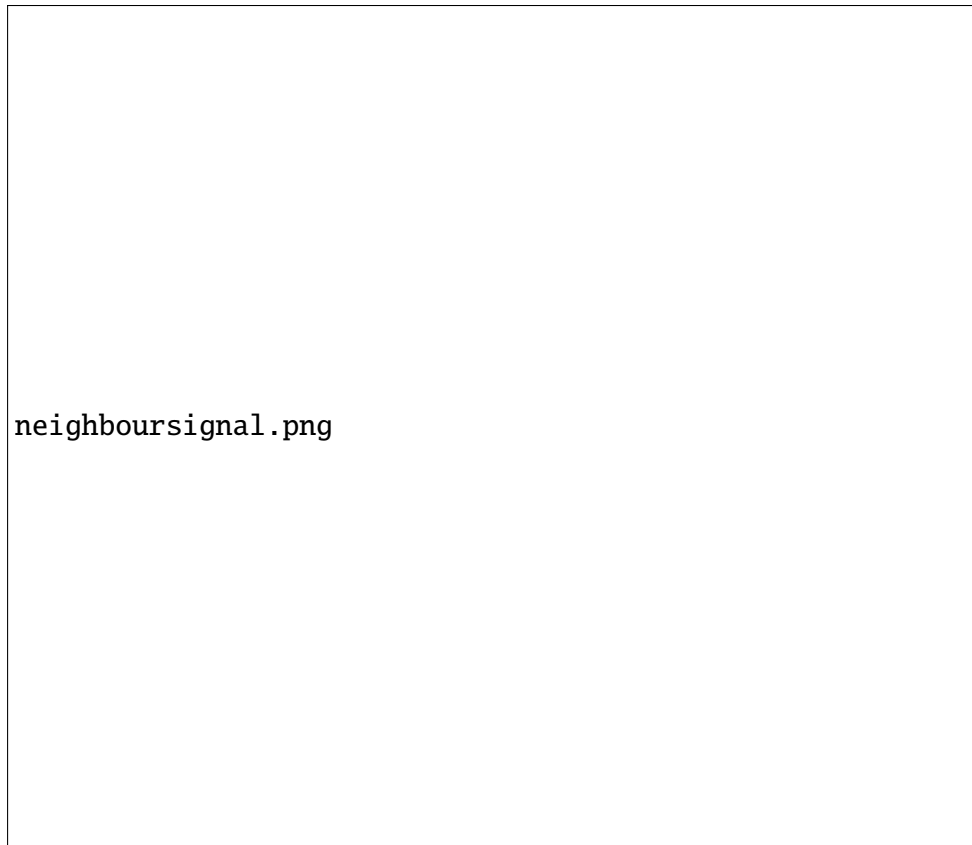


Figure 4.1: An example neighbour model type signal model with 4 agents and alternating convention. Agents 0 and 2 signal in like to their observations, but act opposite, and vice versa for agents 1 and 2. Each agent is most likely to pick its best action urn based on the signals of its neighbours. In this case, orange and purple are used for the different urn setups instead of yellow and cyan, to indicate that they are not the same urn setups that yellow and cyan previously indicated.

did not reveal any reason to expect the model to reach this equilibrium more often than the base model. The reason it happens, though, is that there is no pressure from the non-neighbours to act any particular way, so despite noise being added by non-neighbours, that noise does not drown out the worthwhile signals from an agent's neighbours. Because the two neighbours are not pressuring each other, this functionally creates a string of 2-agent systems. Recall that a common result for 2-agent systems is as described above, as if one agent speaks French but understands Spanish, and the other speaks Spanish but understands French. For 3-agent systems this cannot occur, because there is pressure from a third agent that creates a system of connected pressures forcing convergence to a non-alternating

convention. However, in systems of more than 3 agents pressured only by neighbours, the relationship between agent 0 and agent 1 is not impacted by agent 2 in the same way; agent 2 exerts pressure on agent 1, but not on agent 0, so the pressures do not interact to create a different kind of system. In a normal 4-agent model, agents 0 and 2 cannot use a French/Spanish convention and agent 1 use a Spanish/French convention, because agent 3 will be torn in two directions, and exert its own pressure. It is better for all agents to have only one way to both speak and interpret. As a result, the only stable equilibria are non-alternating conventions. In a 4-agent neighbour model, however, alternating conventions can work, because agent 0 is not pressured by agent 2, so both agents can take the speak French understand Spanish convention, and it will not be a problem that neither understands what the other is saying. This difference in pressure comes from a difference in the game versions that one might expect to come up in an equilibrium analysis, but only becomes sharp when looking at how that pressure affects dynamics.

While this argument runs fine for 6-agent systems as well, there is one step missing for 5-agent. This is not an equilibrium for odd-population games, so it is very surprising that the model settled on something so close to it in the 5-agent model. To explain, things start to settle one way or another once there is an agent that has similar pressures from its two neighbours. Say for example that early on, agent 1 has some pressure from agent 0 and from agent 2 to use an alternating convention, because it has so happened that randomly signaling and acting has both been more successful than not, and has been closer to this convention. This is just as likely to happen as for it to receive pressure from both to use a non-alternating convention, and given enough time one or the other is very likely to happen. Now agent 1 is pressured towards that specific alternating convention, reinforcing the direction agents 0 and 2 are going, and they begin to exert even more pressure on agents 4 and 3, respectively. The problem is that agents 4 and 3 also pressure each other. If both are receiving pressure to speak French and understand Spanish, then they will, by beginning to do these two things, pressure the other to do the opposite. The trick is

that the agents are in randomly determined configurations at this point, and will continue to change with an element of chance. One may be closer to the French/Spanish strategy than the other, and may settle into it. Suppose that is agent 3. Now agent 4 is receiving pressure to play the French/Spanish strategy from one neighbour, and pressure to play the Spanish/French strategy from the other. It is likely to stagnate and not develop into either strategy. The result is an agent 4 that is not exerting meaningful pressure on either of its neighbours, not giving useful information on the efficacy of any given strategy by not having any strategies that work well with it. The system can accommodate this just fine. There is a world of difference between agents 0 and 3 receiving pressure from agents 1 and 2, respectively, and no pressure from agent 4, as opposed to if, counterfactually, agent 4 was exerting a specific kind of pressure. Agents 0 and 3, though receiving weaker pressure than agents 1 and 2, are still receiving steady and specific pressure to prefer a specific strategy. The convention, though leaving one agent out, is stable. On the other hand, while a non-alternating convention would have eventually included agent 4, agent 4 cannot pressure any of the agents that started the chain of events off, so non-alternating conventions are not more likely to be established in the first place. The evolutionary dynamics have given us stable equilibria that are such that one is preferable to the other, but each are stable and about as likely to occur².

Comparing the model types for 4 and 6 agents, neighbour models, difference models, and choice models have larger impacts, each having a significantly lower communication percentage than the last, in that order. The exception is that difference models do not have a significantly larger effect than neighbour models for 4 agents specifically. For 10 agents, difference is not significantly worse than baseline, even for differing agents. After this leap

²I have simplified, of course; how the instigating agents came to their preferences are partially a result of pressure from their own neighbours, so agent 4 does play a role even early on, and it is in theory *slightly* more likely that non-alternating conventions arise as a result. On the other hand, alternating conventions are still plainly present in the data in high proportion, so because this point is getting into minutiae irrelevant to any of my conclusions, and also because it is more difficult to test, I did not spend time testing the extent to which the hypothesis that non-alternating conventions should be more common than alternating ones is borne out in the data.

in agents, the number of signals being received appears to drown out the difference itself. In every other way, however, 10-agent models play out the same way, just at the much lower baseline level of convergence.

There are a small number of interesting combinations. Adding bias to difference, so that the different agent is also biased against, does not significantly impact the percentages of the three typical agents, but drops the communication for the differing agent from 78.4% to 63.5%. In other words, the bias only really affects the biased-against agent. On the other hand, when adding winnings, both values appear to lower, though not a significant amount. The difference, then, is that even though winnings indirectly targets the differing agent because that agent is likely to have lower winnings overall, the direct targeting of bias has a clearly larger impact on those agents. Adding neighbour to difference has a similar effect to bias, hurting differing agents but not the others. At this level it is still possible for neighbour and choice model types to be significantly more destructive than choice alone, though for any higher number choice is destructive enough to render any additions insignificant.

In summary, when controlling for artifacts of number of agents, bias and winnings have almost identical effects on overall communication, usually far less than other model types, although biased-against agents of course do worse than average. Agents are mostly able to create convention around an agent that has difficulty grasping that convention due to internal differences, and internal differences can make a big difference. Adding that agents will select who to listen to and learn based on who has given good advice creates disastrous disconnects that stop conventions from forming, and does not provide a mechanism to overcome problems. When agents have an interest in listening to only their neighbours, they form conventions directly with those neighbours even absent any mechanisms to allow this, due to the pressures inherent to the situation. Changes to starting urn were washed out over time.

4.2 Appraising Signal Models

4.2.1 Discussion of Results

Before once again discussing the lessons to be drawn from the results, it bears repeating that there are serious limitations on how formal epistemic models can be applied to real life. As before, the strongest claims I can really make are that I have established some minimum effect so long as there is no reason to think that effect is an artifact of model design. If one feature has more of an effect than another, it is only weak evidence that the modelled phenomena bear the same relation to each other. Also of key import is that the models are not perfect representations of the phenomena they aim to model; what I call “bias” in the above is not the same thing as real-life bias. Real bias does not manifest in specific percentage decreases in how much attention is paid to biased-against individuals, nor does it stop at attention. As well, real people are not machines, and do not behave so mechanically, so we can only make conclusions about overall trends on average. Finally, as with epistemic dynamics before, meaning conventions are a limited domain and something being worse for meaning conventions does not mean it is worse altogether. However, it is still true that conventions around communication are a site of significant ableism, as explored above, so this is less of a problem here than before.

To start, then, I want to note some things that seem to clearly be artifacts of model design. The results around parameters for base models of different numbers of agents do not tell us anything in particular. As well, the fact that changes to the starting urns did not have further results is due to the nature of formal learning models’ tendency to come to particular conclusions regardless of difference in starting conditions, and does not indicate anything about people’s abilities to adapt to different initial views. Finally, the problems caused by the choice model type were due to a large extent to the way it was designed; people do not decide whom to listen to by keeping track of who has helped the most, and listen to only one other person, but that was the result the model tended to have. On the

other hand, not everything about the choice model was an artifact of design. It is still true that in situations where agents are learning to communicate only paying attention to one other agent, they will fail spectacularly. This has very narrow applicability, unfortunately; when identity groups are ignored it is largely not because of a culture of everyone arbitrarily listening to only one other person, at the very least without reciprocation or popularity playing a role.

There is a point to be made connecting bias and winnings, but it requires qualification. It is mostly true in the results that bias and winnings models had almost the exact same impact, including no significant difference in average communication percentage. As well, the results almost always have one agent singled out as not communicating as well as the baseline. The ready-to-hand takeaway would be that paying attention to how successful someone has been in the past when deciding how much attention to pay to them has the same effect as just being biased against an identity group. Again, this was my expectation due to real-world evidence. For example, neurodivergent people have complained that they cannot advance their careers because of inability to adhere to work cultures and especially job interview norms that have been established by neurotypical people, and that this is exacerbated by the resultant lack of neurodivergent people in roles that might allow them to change this. It is worth considering that giving more power to historically successful people concentrates power further away from people less able under the current social paradigm, which can exacerbate power imbalance, especially related to ableism. However, this requires the qualification that the interaction of winnings and difference models did *not* see a focus of the effect of the winnings model on the differing agents. While bias and winnings function very similarly in the model, winnings does not automatically target the differing agents like I expected. While the model suggests that it is a bad idea to pay attention to prior success on matters where there is not likely a differentiation of inherent skill, it is a bad idea only in the sense of being generically counterproductive, not in the sense of being unjust. All this means is that I have not formally established a minimum

effect size for distribution of power based on past success perpetuating ableism, not that there is good reason to think that it does not do so.

The two most important results come from the same core idea, so before getting into specifics I want to explain that idea. I have previously talked about “pressure,” as distinct from incentive, meaning evolutionary pressure. To draw out the distinction, consider a coordination game. Two players want to go to the same place as one another, and have the same preference about which place. If both go to the movies, they will each get 2 points, and if both go to dinner, they will each get 1 point. On the other hand, if they do not end up in the same place, each gets 0 points. See table 4.7 for the game’s table. The best outcome is obviously for both players to choose “movie,” as it is the highest payout for both players. In a sense, both players are incentivized to pick “movie.” However, in evolutionary game theory, things are more complicated. If my opponent picks “dinner” often enough, they will be changing my expected payout. They are putting pressure on me to pick “dinner” instead. My short-term best option is “dinner” if the 1/1 outcome is at least twice as likely as the 2/2 outcome, so in an evolutionary dynamics, any opponent with more than 66% rate of picking “dinner” is going to make me move towards picking “dinner.” Agents using evolutionary dynamics end up with strategies based on the pressures around them, not based on what would ultimately be the best result for everyone.

	Movie	Dinner
Movie	2/2	0/0
Dinner	0/0	1/1

Table 4.7: A simple coordination game

The first of these pressure-based results comes from difference models. For all numbers of agents, the existence of a differing agent did not significantly impact likelihood of other agents creating a convention that works for themselves, even when the differing agent was unable to conform to that convention. As described in the previous paragraph, it is pressure, not incentive, that matters. While agents were incentivized to come up with a convention

that is inclusive of the differing agents, the existence of an agent that is engaging poorly with a forming convention does not actually place any pressure against that convention, because this would require actively rewarding a move away from the forming convention, not just a lack of rewards for moving toward it. The analogy to the real world would be that people who have atypical communication needs may be disadvantaged in the formation of conventions of communication because the majority is able to create conventions without them, and that this would likely lead to conventions that disadvantage those people in their use.

This posits a clear mechanism by which systemic ableism around communication can be established. If it is true that there are possible conventions of communication that function well for neurotypical people but not neurodivergent people, that neurotypical people are a large majority of people, and that no special effort has been taken to avoid these conventions, then evolutionary pressures will favour those conventions, to the detriment of neurodivergent people. It is worth emphasizing that I am using a fairly strict meaning of “special effort” for the special effort condition; it is not enough for people to be well-meaning, holding no bias against neurodivergent people and possibly even desiring the best for them. The model is structured to reward a typical agent for communicating well with a differing agent, because it receives stronger signal reinforcement if its signals help more agents act well. It is therefore in spite of interest in successful communication with the differing agent that the typical agents exclude it. To overcome this, people need to be paying attention to their communication in a way that the agents do not.

The other conditions hold in reality. Examples of conventions of communication that exclude neurodivergent people are plentiful. One is the prevalence of sarcasm. Sarcasm of course has a place, and many neurodivergent people play with sarcasm in their humour. However, sarcasm is so prevalent in regular speech that it is common for autistic people to misunderstand earnest attempts at communication through sarcasm. For example, I recall one job where I was struggling to close a machine, and I noted this to someone who had

been there longer. They responded, “well keep trying to force it, that’ll work.” This made me think it was expected that I should apply more force, and I broke the machine. I was then reprimanded, and nobody thought that it was reasonable for me to have misunderstood my coworker’s attempt to get me to stop trying the way I was trying. My purpose with this example is not to decry all uses of sarcasm to communicate in particular, but to draw out that conventions of communication exist that seem straightforward to neurotypical people, but can be confounding to neurodivergent people. The other condition, that neurotypical people are the majority, is patent.

The conclusion, then, is that unless special care is taken to stop the formation of conventions of communication that disadvantage neurodivergent people, or effort is spent eliminating ones that exist, there will be cultural evolutionary pressures towards systemic ableism in communication. The virtue of epistemic justice therefore involves active rooting out of systemic ableism, and requires the cooperation of neurodivergent people and majority groups to identify and then cease the usage of these conventions. The only alternative is for neurodivergent people to *create* pressures for neurotypical people to change preferences towards more inclusive communication, through some form of direct action³. Passive measures that play on incentive structures may be less likely to work, because this result was obtained in a model that has positive incentive structures built in.

The other point related to pressure comes from the neighbour model, especially as combined with other features. The goal of the neighbour model was to explore how bias or other factors might interact with agents not having the same information be worthwhile to them, with the expectation that if information differs in value for different agents, already disadvantaged agents will find information most valuable to them further devalued. It instead found that agents with an interest in the information an already disadvantaged agent has will work towards communicating with that agent regardless of those disadvan-

³To be clear, I advocate cooperation, and do not think sarcasm should be our key issue. We *could* do the communicative equivalent to work to rule, and just take everything literally until sarcasm stops, but there are better, if less funny, options.

tages, even absent tools to aid in choosing which other agents to communicate with. This is again because of evolutionary pressure. The disadvantaged agent's neighbours are receiving pressure from only it and one other agent, their own neighbours. As well, the structure of pressures greatly simplifies the relationship between the two agents. Therefore, those disadvantages matter much less. There is some fairly strong minimum amount of coordination that can be forced by pressure, which cannot be undone by the minimum effect of social structure or other disadvantages that have been looked at. The clear exception is the choice model type; if agents are not even paying attention to a specific other agent, that agent is not exerting any pressure. The takeaway for reality should be that it may be more effective to create social change by creating systems that uplift everyone, in order to create self-interested pressure for as many people as possible to join the movement, rather than to try to force people to work together through changes in social structure that lack pressure. Placing disadvantaged people into communication with as many others as possible is always a precursor.

As an example, affirmative action policies to put disabled people into workplaces will not on their own eliminate epistemic injustice around disability in workplaces; those workplaces must also restructure communication in a way that is fully inclusive of disabled workers, but also represents a clear improvement for other workers. This is something that can be carried out at a grassroots level, without policy; people do not need to be mandated to do things that clearly benefit them. Note, however, that the word "clearly" is doing some work here; if people are made to believe that they are not experiencing benefits from a change, there is nothing in the above to indicate whether or not they would thereby resist that change, and it is entirely *prima facie* plausible they would.

It is worth repeating here that in both of these model types, and throughout, the suggestion from chapter 2 that integration into wider society is helpful is borne out again. Difference models do better the larger the number of communicating agents is, and evolutionary pressure in neighbour models is confounded only by complete exclusion. Both

points indicate that creating more genuinely communicative connections for marginalized people is productive for combating epistemic injustice.

Before concluding, there is a weak point to be made that introducing differences in individual agents' interests had a bigger average effect than bias. This point must be made with weaker force than the above because it is looking at differences in the established minimum effect sizes, rather than extrapolating directly from those minimum effect sizes. Establishing a higher minimum for one value than another does not provide very strong evidence that the former value is higher than the other, but it may shift which we think is more likely to be higher. Therefore, the conclusion to draw is that we have weak evidence that the effects of the pressure-based considerations discussed above may be more important than the systemic identity-based biases that individuals hold. This is again in line with Anderson's point that our focus should be on systemic issues, not the virtues of individuals [10].

In conclusion, the data collected by these models points us towards cultural-evolutionary pressure as a partial explanation for systemic ableism present in conventions of communication. Whatever else may have led to our current social context, there is some evolutionary effect detectable in formal models that would push towards systemic ableism. This work, based heavily on methodology advanced by Cailin O'Connor, ends up supporting the suggestions she makes in *Origins of Unfairness* [17]. She points out that to counteract evolutionary pressures toward unjust situations, a good strategy would be to create evolutionary pressures toward just situations. Her description of what this looks like evokes direct action, standing in the way of the machinery of oppression in order to make acting as an oppressor less convenient than uplifting the marginalized. What she adds beyond the observation that direct action is directly effective is that it must be permanent. Evolutionary pressures are a feature of nature rather than a contingent situation, and are not going to go away. Therefore, whatever is done to move the world towards justice must continue even in a fully just world, lest it slip back into the situation it was initially pressured into.

I concur, but have a further addendum. Positive pressure is effective. It is a good idea

to get in the way of oppressive systems and make their continued existence inconvenient to the oppressor. It is also a good idea to have as many people benefit from change as possible. While O'Connor is right to focus on what means are best, it is also important at every stage to know what ends one is striving for, and to choose means that best support a particular end. The advice for a just end goal, then, is that we will be most effective in bringing it about if it is to the relative benefit of as many people as possible, and not just disabled people. Short of that, however, the data supports that negative pressure should also be effective.

4.2.2 Shortcomings

While powerful, this model, like any other, has limits on its applicability. Just as the previous model was limited by looking only at information exchange, this one is limited by looking only at formation of meaning. Fricker gives a large part of the remaining picture by analyzing the results of systemic breakdowns in communication [6]. However, a complete picture of systemic ableism in communication norms will need a lot more than this interaction, taking a more fine-grained look at the specific ways disabilities directly and indirectly impact communication, and forms of oppression not considered here and how they intersect with all of the above and each other. Additionally, just as much as before, any recommendations toward addressing systemic ableism in norms of communication do not imply that a value system that prioritizes this topic is superior to one that does not, and before putting any advice to use, the impacts on other areas should also be considered. Once again, a major mitigating factor of this limitation is that its abstract nature may allow it to map onto other situations; if the conclusion is that a good strategy to resist something that is partially a result of cultural-evolutionary processes is direct manipulation of those processes rather than the results themselves, then this is likely to apply to other partial results of cultural-evolutionary processes. That said, as has already been seen with prior work, more modeling of different phenomena with different formal tools can give deeper

understanding for different domains, and will continue to be worth doing.

This model retains the previous one's limitation that agents are overly mechanical in nature compared to real people. As before, a mechanical explanation for how systemic ableism can arise without specific ill will should not be taken as a naturalization of injustice. It will always be true that people could and should have noticed that the needs of a minority group were not being met far earlier, and those needs could and should have been met before today. It is not a neutral fact that we have only recently come to understand neurodivergence the way we understand it, but a political one; we would have understood much more much earlier had more people in the past with control of the collective hermeneutical resource acted on an interest in the well-being of people who were being marginalized by the norms of communication. The forces discussed above are nothing more than general trends—do not miss that anyone with power has the power to care. While one of the model types above tries to address this by limiting agents' interests to only themselves and their neighbours, the results ended up having implications mostly unrelated to actual free choice. It may not be possible for a formalism that has agents taking only what actions are best for them to fully capture the choices and ethical culpability present in the creation of systemic injustice, but it is at present unclear what utility any other kind of formalism would have. It would be worthwhile in the future to attempt to develop and study such a formalism, if for no reason other than to eliminate them as a possibility and firmly set boundaries on the usefulness of formal epistemology for study of injustice.

The signal model was able to mitigate many of the limitations that the network learning model had, but where it falls shorter is in the fine-grained look it was easy to take with the network model. It was much harder to give graphical or step-by-step depictions of how the models played out mechanically, which was a large portion of the previous model's value. This depiction gave significant detail to the already present understanding of how epistemic networks can create unjust outcomes, which is a big help compared to the fairly limited advice to be gleaned from the statistical results. Conversely, my more involved

statistical analysis of the signal models gave more expansive recommendations, but did not do much to elucidate the mechanics of cultural evolution. Since this difference was mostly the result of the existence of the NetworkX library for Python, the development of a similar tool for visual depiction of signal models would go a long way to dig deeper into what is happening in this chapter's models.

Finally, the model type I termed "choice" was a particular letdown. My design did not do a good job of giving agents a choice about which other agents to listen to, and gave relatively little of any value. Future work should look at more ways to describe social structure in signal models. For example, a future model may try to combine networks and signal models. Another may find a way to give agents a more robust kind of agency in associating with specific other agents. Future investigation into social structures and communication may also be able to touch on rippling consequences of particular kinds of signaling. The "winnings" and "urn" model types did not give much conclusive information on the impacts on future communication of neurodivergent signaling differently early in their development, which seems like a major site of real-world struggle. Another reason to try to combine networks and signal models is that network learning models and signal models cover each other's weaknesses in the realm of epistemic injustice. Just as network learning models leave out most aspects of hermeneutical injustice, signal models leave out most aspects of testimonial injustice. Combining these two structures does not guarantee a remedy, but a fuller picture of how the two interacting forms of epistemic injustice do interact could possibly be achieved by having network learning and signaling both occur.

Overall, where network learning models gave a more detailed mechanical look, signal models gave a larger set of strategic recommendations. Both can be combined for a more robust view of what is happening, and what to do.

Conclusion

I will first address anyone who began reading this without a conviction that ableism is present in our norms of communication. I hope that most such readers would be easily convinced by simple exercises in paying attention to who is listened to in their daily lives—assuming they have neurodivergent people in their daily lives—but this should also not be necessary by this point. Strictly speaking, lack of belief in ableism in norms of communication does not imply positive belief in a lack of such ableism; however, on balance, it should be far more credible *prima facie* that such ableism would be present. Unless the reader has good cause to doubt that the processes described above translate at all to the real world, or good cause to believe that they have been adequately addressed far earlier than they were ever articulated, they should accept that there is a tendency for conventions of communication to be formed in ways that disadvantage groups we would now call “disabled.” That tendency, absent any other data, implies that it is far more likely that our present norms are ableist than not. So if nothing else, at least be convinced that epistemic injustice is real and affects neurodivergent people.

I take it for granted, then, that this is a problem worth solving. From the preceding, at least some causes of the problem are that when cultural objects like conventions are created in a context of cultural evolutionary pressures, individuals with atypical communication needs are unable to place their fair share of pressure on that creation, and individuals that are excluded or biased against are less coherent to others and, whether from mere lack of outward reflection or something harmful replacing it, to themselves. In addressing the

resultant systemic ableism, we have to pay attention to its causes lest they recreate that ableism in the wake of our work. This means that whatever strategy we take, we must make sure we are at least addressing the systems themselves and not just individuals, that we are including disabled people of all kinds in meaningful rather than tokenizing ways, and that we are creating a situation in which movement towards less ableist norms is the path of least resistance for as many people as possible.

It is outside the scope of this project to come down firmly that either grassroots or policy-focused strategies for achieving these aims are better, nor is there a clear reason not to support both, so I will look at advice for either kind of strategy.

Extant efforts to bring disabled people into the mainstream are a good place to start. For example, my mother is a resource teacher in British Columbia. She helps high schoolers with special needs develop the interpersonal skills neurotypical people take for granted, and require for baseline participation, and then facilitates integrating them into workplaces suited to their skillsets. Creating connections across ability lines through which the competency and humanity of disabled people is undeniable creates the baseline of what is necessary for any change of the kind I recommend. The very least we can do is advocate for policy expanding programs like hers, affirmative hiring action, and for a general culture of including disabled people in communities not related to work.

As I have stressed, however, affirmative action is not enough on its own, and especially not if neurodivergent people are placed into environments where they lack the solidarity of other neurodivergent people within that environment. Programs aimed at integrating disabled people into the mainstream should take care that they do so in a way that is actually helpful for those people themselves, at least by ensuring that they will have others like them around. As well, if we can directly break down barriers like bias and communication differences through policy, we should pursue such policy. In corporate environments, for example, making sure some people are trained in various styles of communication and available for facilitating good communication could go some of the way towards making

communication more effective, which would in turn hopefully reduce bias. The risk of excessive othering or counterproductive training would be present, so as with all things, neurodivergent people should ideally be involved in the development of such policies. Outside corporate environments such roles may be hard to include; what further efforts to improve communication and inclusion via policy might look like eludes me.

There is already a movement to include neurodivergent researchers more in research into neurodivergence. I would emphatically support this movement, and support work towards that goal, whether through grassroots advocacy, using our influence to change publishing practices, or through actual policy like grants and affirmative action. One thing that is worth noting for this domain is that there is *not* a need to oust all neurotypical researchers. Talented and well-meaning neurotypical researchers exist, and their mere presence is not actively harmful—it is the *absence* of neurodivergent ones that is harmful. As multiple models show, just adding one connection with a divergent agent is enough to significantly improve communication both ways for a divergent agent. Just a small number of neurodivergent researchers per research project would go a very long way, as long as they have actual agency at each level of the project. If more argumentation is needed, note that part of the problem to be solved is just that neurodivergent people frequently do not discover that they are neurodivergent until late in life; at a large enough scale, exclusion of people who seem not to be neurodivergent will end up excluding some neurodivergent people.

We can also model movements after previous successful movements to combat epistemic injustice. To use one of Fricker's examples, women's groups and speak-outs are credited with creating and popularizing the concept of workplace sexual harassment [6]. Neurodivergent discussion groups and speak-outs could also be helpful. It appears that the effects of digital equivalents are already being seen. Many people on the popular smartphone application Tik Tok describe realizing they were neurodivergent as a result of relating to others on the application who talk about their experiences as neurodivergent people. However, social media algorithms are not designed with the goal of advancing

social activism in mind; these discussions are most likely to reach people who already express interest in neurodivergence or related topics. As my results indicate, direct action is most efficacious, I would look to past examples of direct action like speak-outs. Organizing neurodivergent speak-outs, or similar events for specifically autistic or generally disabled people, could go a long way to reaching an audience that would not receive the same information from social media, disrupting the flow of cultural information as it currently exists.

Finally, when it comes to cultural change, at least a significant amount of the work has to be at the grassroots level. Readers can begin to help just by checking their own biases around disability, and vocally contesting ableism around them. Evolutionary pressure can look like an increase in an attitude that punishes visible ableism, which thereby reduces the viability of visible ableism. In this case “punishment” can just mean being shown disapproval in one instance and not having the ableist party’s ableist comment taken into consideration; if ableism is not efficacious toward any goal at all, it will not be appealing. However, we need to make shifts to systems, not just individuals. A grassroots approach could find ways to communicate *in general* that work better *in general*, but especially for neurodivergent people. Advocacy for such a mode of communication would result in systemic change as opposed to individual by changing the system in use for communication itself. Sketching such a mode of communication is beyond the scope of this project. Visibility is also something that is structural; increasing the actual visibility of neurodivergence increases overall expectation that neurodivergence will be seen, which will hopefully keep visibility at the new level. Neurodivergent people can contribute to their own liberation, then, just by making art or otherwise being present in the public eye, and neurotypical people can contribute by uplifting neurodivergent creators. If an attempt to sketch such a movement is to take anything away, it is that the actions taken by the movement should result in the most convenient action for anyone affected becoming a move away from ableism, whether by making ableism inconvenient, or anti-ableism very convenient.

Bibliography

- [1] Yuanhe Huang, Samuel RC Arnold, and Julian N Trollor. “Diagnosis of Autism in Adulthood: A Scoping Review”. In: *Autism* 24.6 (2020), pp. 1311–1327.
- [2] Dori Zener. “Journey to Diagnosis for Women with Autism”. In: *Advances in Autism* 5.1 (2019).
- [3] Maire Claire Diemer, Emily D Gerstein, and April Regester. “Autism Presentation in Female and Black Populations: Examining the Roles of Identity, Theory, and Systemic Inequalities”. In: *Autism* 26.8 (2022), pp. 1931–1946.
- [4] Jim Sinclair. “Why I Dislike ‘Person First’ Language”. In: *Autonomy* 1.2 (1999).
- [5] G. Raynor and S. Gale. “ADHD and Executive Function Disorders”. In: *Neuropsychiatry and Behavioral Neurology: Principles and Practice*. Ed. by DA Silbersweig, LT Safar, and KR Daffner. McGraw Hill, 2021.
- [6] Miranda Fricker. *Epistemic Injustice*. Oxford University Press, 2007.
- [7] Cailin O’Connor and James Owen Weatherall. *The Misinformation Age*. Yale University Press, 2019.
- [8] Venkatesh Bala and Sanjeev Goyal. “Learning from Neighbours”. In: *The Review of Economic Studies* 65.3 (1998), pp. 595–621.
- [9] Cailin O’Connor and James Owen Weatherall. “Scientific Polarization”. In: *European Journal for Philosophy of Science* 8 (2018), pp. 855–875.

- [10] Elizabeth Anderson. “Epistemic Justice as a Virtue of Social Institutions”. In: *Social Epistemology* 26.2 (2012), pp. 163–173.
- [11] Shelley Tremain. “Knowing Disability, Differently”. In: *The Routledge Handbook of Epistemic Injustice*. Ed. by I. J. Kidd, J. Medina, and G. Pohlhaus Jr. Routledge, 2017.
- [12] José Medina. “Feminism and Epistemic Injustice”. In: *The Oxford Handbook of Feminist Philosophy*. Ed. by K. Q. Hall and Åsta. Oxford University Press, 2021.
- [13] Gail Pohlhaus Jr. “Varieties of Epistemic Injustice”. In: *The Routledge Handbook of Epistemic Injustice*. Ed. by I. J. Kidd, J. Medina, and G. Pohlhaus Jr. Routledge, 2017.
- [14] J. M. Reynolds and A. Silvers. “Feminism and Disability”. In: *Philosophy: Feminism*. Ed. by Carol Hay. Macmillan Reference USA, 2017, pp. 295–316.
- [15] Sally Haslanger. “Objectivity, Epistemic Objectification, and Oppression”. In: *The Routledge Handbook of Epistemic Injustice*. Ed. by I. J. Kidd, J. Medina, and G. Pohlhaus Jr. Routledge, 2017.
- [16] Kristie Dotson. “A Cautionary Tale: On Limiting Epistemic Oppression”. In: *Frontiers* 33.1 (2012).
- [17] Cailin O’Connor. *The Origins of Unfairness*. Oxford University Press, 2019.
- [18] Leslie R. Walker, Anisha A. Abraham, and Kenneth P. Tercyak. “Adolescent Caffeine Use, ADHD, and Cigarette Smoking”. In: *Children’s Health Care* 39.1 (2010), pp. 73–90.
- [19] Csilla Ágoston et al. “Self-Medication of ADHD Symptoms: Does Caffeine Have a Role?” In: *Frontiers in Psychiatry* 13 (2022).

- [20] Søren Dalsgaard et al. “ADHD, Stimulant Treatment in Childhood and Subsequent Substance Abuse in Adulthood — A Naturalistic Long-Term Follow-Up Study”. In: *Addictive Behaviours* 39.1 (2014), pp. 325–328.
- [21] Kevin J. S. Zollman. “The Epistemic Benefit of Transient Diversity”. In: *Erkenntnis* 72 (2010), pp. 17–35.
- [22] C. J. Crompton et al. “Autistic Peer-to-Peer Information Transfer is Highly Effective”. In: *Autism* 24.7 (2020), pp. 1704–1712.
- [23] Kevin J. S. Zollman. “The Communication Structure of Epistemic Communities”. In: *Philosophy of Science* 74.5 (2007), pp. 574–587.
- [24] Bennett Holman and Justin P. Bruner. “The Problem of Intransigently Biased Agents”. In: *Philosophy of Science* 82.5 (2015), pp. 956–968.
- [25] Cailin O’Connor, James Owen Weatherall, and Justin P. Bruner. “How to Beat Science and Influence People: Policymakers and Propaganda in Epistemic Networks”. In: *The British Journal for the Philosophy of Science* 71.4 (2018).
- [26] Jingyi Wu. “Epistemic Advantage on the Margin: A Network Standpoint Epistemology”. In: *Philosophy and Phenomenological Research* (2022).
- [27] Erich Kummerfeld and Kevin J. S. Zollman. “Conservatism and the Scientific State of Nature”. In: *British Journal for the Philosophy of Science* 67.4 (2016), pp. 1057–1076. doi: 10.1093/bjps/axv013.
- [28] Sarita Rosenstock, Cailin O’Connor, and Justin Bruner. “In Epistemic Networks, is Less Really More?” In: *Philosophy of Science* 84.2 (2017), pp. 234–252. doi: 10.1086/690717.
- [29] Bennett Holman and Justin P. Bruner. “Experimentation by Industrial Selection”. In: *Philosophy of Science* 84.5 (2017), pp. 1008–1019. doi: 10.1086/694037.

- [30] Bill Hughes. “Being Disabled: Towards a Critical Social Ontology for Disability Studies”. In: *Disability & Society* 22.7 (2007), pp. 673–684. doi: 10.1080/09687590701659527. eprint: <https://doi.org/10.1080/09687590701659527>. URL: <https://doi.org/10.1080/09687590701659527>.
- [31] Sina Fazelpour and Daniel Steel. “Diversity, Trust, and Conformity: A Simulation Study”. In: *Philosophy of Science* 89.2 (2022), pp. 209–231.
- [32] Erik J. Olsson. “A Bayesian Simulation Model of Group Deliberation and Polarization”. In: *Bayesian Argumentation*. Ed. by Frank Zenker. Springer Netherlands, 2013.
- [33] Cailin O’Connor. “The Cultural Red King Effect”. In: *The Journal of Mathematical Sociology* (2017).
- [34] Connor Mayo-Wilson, Kevin J. S. Zollman, and David Danks. “The Independence Thesis: When Individual and Social Epistemology Diverge”. In: *Philosophy of Science* 78.4 (2011), pp. 653–677.
- [35] Michael Weisberg and Ryan Muldoon. “Epistemic Landscapes and the Division of Cognitive Labor”. In: *Philosophy of Science* 76.2 (2009), pp. 225–252. doi: 10.1086/644786.
- [36] Johanna Thoma. “The Epistemic Division of Labor Revisited”. In: *Philosophy of Science* 82.3 (2015), pp. 454–472. doi: 10.1086/681768.
- [37] Samuli Pöyhönen. “Value of Cognitive Diversity in Science”. In: *Synthese* 194.11 (2017), pp. 4519–4540. doi: 10.1007/s11229-016-1147-4.
- [38] Jason McKenzie Alexander, Johannes Himmelreich, and Christopher Thompson. “Epistemic Landscapes, Optimal Search, and the Division of Cognitive Labor”. In: *Philosophy of Science* 82.3 (2015), pp. 424–453. doi: 10.1086/681766.

- [39] Conor Mayo-Wilson and Kevin J.S. Zollman. *The Computational Philosophy: Simulation as a Core Philosophical Method*. 2020. URL: <http://philsci-archive.pitt.edu/18100/>.
- [40] MW Macy et al. “Polarization in Dynamic Networks: A Hopfield Model of Emergent Structure”. In: *Dynamic Social Network Modeling and Analysis*. Ed. by R Breiger, K Carley, and Pattison P. National Academies Press, 2003, pp. 162–173.
- [41] The NumPy Community. *numpy.random.seed*. 2021. URL: <https://web.archive.org/web/20210502064616/https://numpy.org/doc/stable/reference/random/generated/numpy.random.seed.html> (visited on 05/02/2021).
- [42] Jules Holroyd, Robin Scaife, and Tom Stafford. “What is Implicit Bias?” In: *Philosophy Compass* (2007).
- [43] David Lewis. *Convention*. Harvard University Press, 1969.
- [44] Brian Skyrms. *Evolution of the Social Contract*. Cambridge University Press, 1996.
- [45] Aileen H. Sandoval-Norton and Gary Shkedy. “How Much Compliance is Too Much Compliance: Is Long-Term ABA Therapy Abuse?” In: *Cogent Psychology* 6.1 (2019).
- [46] Gary Shkedy, Dalia Shkedy, and Aileen H. Sandoval-Norton. “Long-Term ABA Therapy Is Abusive: A Response to Gorycki, Ruppel, and Zane”. In: *Advances in Neurodevelopmental Disorders* 5 (2021), pp. 126–134.
- [47] O. McGill and A. Robinson. “Recalling Hidden Harms: Autistic Experiences of Childhood Applied Behavioural Analysis (ABA)”. In: *Advances in Autism* (2020).
- [48] Brittany Garcia Freitas. *Questioning Normativity: Exploring the Experiences of Autistic Adults Who Have Undergone Applied Behavioural Analysis (ABA)*. 2020. URL: https://rshare.library.ryerson.ca/articles/thesis/Questioning_Normativity_Exploring_the_Experiences_of_Autistic_Adults_Who_Have_Undergone_Applied_Behavioural_Analysis_ABA_/14663727/1/files/28149561.pdf.

- [49] Brian Skyrms. “Evolution of Signalling Systems with Multiple Senders and Receivers”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1518 (2009), pp. 771–779.
- [50] John Maynard Smith and George Price. “The Logic of Animal Conflict”. In: *Nature* 246 (1973), pp. 15–18.

Appendix A

Code and Data

The python code and full data tables used for this project are both too large to reproduce here. They can be found in a dropbox by typing the url “bit.ly/Marcotte22” into your web browser of choice.

The code can be run by downloading the python files into the directory in which you would like the output files to appear, opening them in an editor like Notepad++ or PyCharm, and adjusting the parameters at the bottom of the files before running them as normal—a knowledge of how the code works is not necessary. If your machine does not already have an installation of Python 3.9 with the libraries Numpy, Pandas, Matplotlib, Pylab, Multiprocesing, Networkx, and Mesa, you will need to install them first. Other editions of Python 3 are likely to run the code

Curriculum Vitae

Name: Mackenzie Marcotte

Post-Secondary Education and Degrees: Quest University Canada
Squamish, BC
2013-2017 B.A.Sc.

Western University
London, ON
2017 - 2018 M.A.

Western University
London, ON
2018-2023 Ph.D.

Honours and Awards: Keystone Distinction
Quest University Canada
2017

Related Work Experience: Teaching Assistant
Western University
2017-2022

Co-Instructor
Western University
2020-2021