CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Coordinating minds: mindshaping, communication and technology
Authors	Finnegan, Colum
Publication date	2023
Original Citation	Finnegan, C. 2023. Coordinating minds: mindshaping, communication and technology. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2023, Colum Finnegan. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-10-17 08:54:05
Item downloaded from	https://hdl.handle.net/10468/15936

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork

Coordinating Minds: Mindshaping, Communication and Technology

Thesis presented by

Colum Finnegan

0000-0003-3995-5704

for the degree of

Doctor of Philosophy

University College Cork

Department of Philosophy

Head of School: Prof Don Ross

Supervisor: Prof Don Ross

2023

Table of Contents

Declaration **1**

Dedication **2**

Abstract **3**

Chapter 1 – Coordination and Communication **5**

1.1 Communication, Ontology and Minds *8*

1.2 Coordination and Applied Game Theory *15*

1.3 Structural Overview *28*

Chapter 2 – The Mindshaping-Coordination Complex **33**

2.1 Mindreading and Coordination *35*
2.1.1 Coordination Without Mindreading 42
2.1.2 From Reading to Shaping 47

2.2 Making and Maintaining Minds *55*
2.2.1 Whither Mindreading? 61

2.3 The Processes of Mindshaping *64*
2.3.1 Imitation and Pedagogy 67
2.3.2 Conformist Mindshaping 71

2.4 Conclusion *80*

Chapter 3 – Cognitive Evolution and Mindshaping **83**

3.1 Mindreading and Evolutionary Psychology *85*
3.1.1 Paris: Ostensive Communication and Minimal Mindreading 89

3.2 Cultural Evolution and Mindshaping *94*
3.2.1 Cultural Evolution in California 94
3.2.2 Sterelny: Niche Construction and Cognition 100
3.2.3 Heyes: Cognitive Gadgets and Mindreading 107

3.3 Model Choice in Mindshaping *111*
3.3.1 Social Learning in California 113
3.3.2 Learning Socially How to Learn Socially 116
3.3.3 Social Learning in New Domains 122

3.4 Conclusion *126*

Chapter 4 – Coordination Online **131**

4.1 Epistemology for Coordinators *136*

4.2 Communication Technologies and Group Signalling *148*
4.2.1 The First Communication Technologies 149
4.2.2 Modern Communication Technologies 157

4.3 Signalling and Coordination Online *165*

4.4 Re-Engineering Mindshaping Online *182*

4.5 Experimental Resources *189*
4.5.1 Minimal Group Paradigm 190
4.5.2 Confederates 192
4.5.3 Monetary Incentives 192
4.5.4 Online Sample Recruitment 194

4.6 Conclusion *195*

Chapter 5 – Politics, Economy and Spectrums of Possibility 198

5.1 Neoliberal Utopias 201

5.2 Neoliberal Digital Communication 211

5.3 Spectrums of Potentiality 223
5.3.1 Efficient Coordination/Coordination Noise 227
5.3.2 Democratising Discourse/Empowering Gatekeepers 231
5.3.3 Mindshaping Model Diversity/Cultural Homogenisation 234
5.3.4 Liberal Democratic Trade-offs 236

5.4 Conclusion 239
5.4.1 Politicising Philosophy of Mind 239
5.4.2 Mindshaping and Coordinating 246
5.4.3 Evolving to Learn Strategically at Scale 248
5.4.4 Digital Communication Technology: Putting it All Together 250

References 253

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Dedication

I dedicate this work to Alex, without whom this fork would never have been taken.

And to Nora, whose arrival radically reshaped my own mind for the better.

My utmost thanks to Don, whose encouragement and clear guidance made this thesis exponentially better, and whose depth, range and ability in remarkably varied knowledge domains continues to be an inspiration.

Abstract

Traditional explanations of social cognition and coordination are under revision. Accounts that emphasise mindreading, the dominant, internalist and spectatorial framework, are challenged by accounts based on mindshaping, an externalist, interactionist concept. Mindshaping places the interpersonal regulation of minds at the root of coordination, and self-generation. However, if the profiles of selves are socially sourced, as mindshaping accounts claim, then structures that interact with these processes call for close examination in light of the new insights.

One such structure, that now plays a crucial role in facilitating mindshaping, is digital communication technology. Due to the historical dominance of mindreading accounts, online communication has been primarily conceptualised in individualistic terms. As a result, the *content* of information communicated online (e.g. fake news, conspiracy theories) is seen as the primary driver of problematic outcomes. However, the mindshaping-based framework reveals that many pernicious features of online discourse are instead caused by the *form* of online communication and the incentive structures it creates for agents seeking to coordinate action. Specifically, as currently designed, online communication interferes with and modifies the types, range and effectiveness of signals deployed by agents to signal coordination suitability, changing the equilibrium dynamics of quotidian interpersonal coordination. In this new domain, I argue that mindshaping processes like imitation and conformism push agents to strategically adopt increasingly extreme belief sets. This in turn creates significant coordination noise and generates a perception of ecological threat – thereby undermining general social welfare and contributing to political stasis. These outcomes are in part the result of particular interface design choices which are responsive to the economic incentive structure

within which online communication providers emerged and operate. However, this incentive structure could be reformed, and mindshaping-based explanations can play a crucial role in guiding such reforms, delineating specific causal dynamics that underpin dysfunctional online communication. I draw together literature from coordination theory, mindshaping, cultural evolution, political economy and online knowledge dissemination to show how the current digital ecology shapes minds in problematic ways but could be re-engineered by regulators to shape minds less perniciously.

Chapter 1 – Coordination and Communication

Over the previous decade a series of political ruptures and epistemic trends have revealed that the machinery of democratic governance and public knowledge generation are under strain. The rise of extreme demagogic political actors, the emergence of widespread scepticism about professional expertise in the domains of health, economics and the environment, and a general perception of ostensibly zero-sum cultural disputes in the public sphere are all evidence of a deep and growing malaise. Political polarisation is on the increase (Borbáth et al. 2023; Abramowitz & McCoy 2019). Rival political factions have increasingly little common ground, even as their core ideologies are delinked from their traditional values. Geopolitical tension is again increasing after a period of relative stability, and action on climate change commensurate with the scale of the challenge is elusive, reflecting a deep political paralysis born of an inability to effectively bring about large-scale coordination (McCoy et al. 2018).

The causes of this shift to a more unstable, volatile and suspicious political and cultural environment are of course multiple. The Great Financial Crisis of 2008 is at least one proximate cause, unleashing as it did a broad loss of faith in established elite rule and a decade of economic stasis. Another widely cited contributing factor is the recent decisive shift to a new infrastructure of mass communication, enabled by the internet and in particular social media (Bennett & Livingston 2020a). The removal of traditional publication barriers and the spread of so-called “fake news” and misinformation this has made possible is thought to be a decisive factor in generating our current political and epistemic instability. On this view, a primary cause of the various breaks with the political status quo and the increasing public rejection of expertise is the prevalence of falsehoods in the public

domain made possible by social media, and a concurrent decline in the belief that "truth" is an attainable epistemic goal (Muhammed & Mathew 2022; Hochschild & Einstein 2015). Proponents of such a diagnosis thus often propose that the solution is to more strenuously police online discourse to remove falsehoods and to alert people to the epistemic status of various claims (Porter & Wood 2022; Pennycook et al. 2020; Hameleers & Van der Meer 2020; Kim et al. 2019).

In what follows, I join this debate using resources drawn from contemporary philosophy of mind, cognitive science and applied game theory to criticise the commonly encountered fake-news-centred diagnosis of the contemporary epistemic malaise and its proposed solutions. Specifically, I use a theoretically sophisticated and empirically robust explanation of human social cognition based on the concept of "mindshaping" (Zawidzki 2013) to re-examine digital mass communication technology and its influence on individual agents and their ontologies.[1] I show that the theory of mindshaping can help illuminate the novel manner in which these technologies can generate pernicious effects for public discourse. As we will see, the primary issue with online communication lies not with "fake news" or general falsehoods – issues that undermine individual epistemic capacities – but rather relates more directly to the *structures* of communication implemented online. This is because, according to mindshaping theory, interpersonal coordination pressure is the primary driver of ontology formation, not the acquisition of true knowledge of the world, and as such, to understand what determines individual and group ontologies and the relationships between individuals, their cultural groups, and other out-groups requires theorists to pay close attention to coordination pressures and signalling

[1] Throughout the following the term digital mass communication technology is used to refer to internet-based communication platforms, and, in particular, social media.

incentives created by specific environments of interaction. Social media, as currently designed, are particularly invidious on such a view, as they pervasively incentivise the adoption of coordination equilibria that are corrosive for social action and large-scale welfare.

Mindshaping processes play a central role in determining belief sets and the form of coordinating groups that exist in a given society. Thus, the theory has applications beyond philosophy of mind. The role mindshaping plays in structuring understandings of quotidian reality and enabling better or worse outcomes for human groups means that, if correct, it generates a host of novel normative and political implications. Mindshaping theory proposes a radical reconfiguration of how it is that human selves and groups come into being; it illustrates the priority of enabling coordination in ontology formation and it highlights the importance of signalling environments in determining group beliefs. As such it can potentially provide resources that can be used to influence broad societal outcomes and influence policy makers. Yet to date, such implications and uses remain undertheorised, a lacuna this thesis partly aims to remedy.[2] In addition to intervening in the pressing debate about the broader social effects of online mass communication as currently designed, this work, through making concrete, empirically assessable claims, aims to both support mindshaping theory as the most promising explanation of social cognition yet developed, and serve as an example of how it can be used to guide policy.

This introductory chapter carries out three specific tasks: in Section 1.1, I provide a summary of the general topic. In Section 1.2, I situate my work within a

[2] McGeer (2015) and Haslanger (2019) have recognised similar implications in the context of achieving social justice.

key existing literature, namely applied game theory. Finally, in Section 1.3, I sketch a map for the reader of the following four chapters.

1.1 Communication, Ontology and Minds

The central aim of this work is to show how a specific, novel understanding of social cognition can be used to shed light on the design of contemporary communication platforms, showing how they may incentivise a particularly pernicious form of interpersonal communication. This explanation makes a break with the traditional accounts of social cognition and interpersonal coordination which posit a functional faculty known as mindreading (Nichols & Stich 2003). Mindreading is the idea that individual human agents engage in observation of others and use various clues to infer the specific internal brain states driving others' behaviours. Once those states are inferred the agent doing the mindreading can act strategically to try to achieve her specific aims. This species of explanation is still widespread in philosophy of mind and, I argue, partly underpins claims that the unmonitored spread of falsehoods is a primary cause of online communicative dysfunction.

Notwithstanding its popularity, this account has come under challenge. Increasingly philosophers have begun to adopt an alternative paradigm, one that builds on the broader *externalist* reframing of cognitive science and cultural evolution (Hutchins 1995; Clark 1997; Sterelny 2003, 2012; Ross 2005, 2022; Dennett 2017). Rather than the spectatorial modelling of others to generate theory-based predictions, this new approach understands humans as knitted into interacting influence networks that naturally guide and constrain actions (Hutto 2008; Gallagher 2008; Andrews 2015). In general, on these accounts, minds do not exist in isolation

containing discrete propositional attitudes that can be inferred using theory. Instead, they are brought into being and dynamically regulated to remain intelligible by social interaction.

In its most well-developed and empirically rigorous form, the phenomenon described by this approach is known as "mindshaping" (Mameli 2001; McGeer 2007, 2015, 2020; Zawidzki 2008, 2013). On this view, social cognition is enabled by the nature of shared mind *making*: in most contexts people are not opaque agents that require theory to explain, but rather utilise action strategies that exist in public frames of reference, and dynamically negotiate mindstates on the fly via the mindshaping processes of imitation, pedagogy and conformism, examined in Chapter 2. These mechanisms were selected for early in our lineage, in the deep history of *Homo sapiens* (Sterelny 2012; Zawidzki 2013). Human minds do not exist prior to, nor stand separate from, social immersion, instead, they emerged *due* to this immersion and co-evolved with the increasing sophistication of social domains. In sum, minds are culturally evolved entities that allow for dynamic social cognition and interpersonal coordination in complex decision contexts.

One significant, yet largely underexplored, implication of the mindshaping-based account of mindedness is that significant parts of a person's beliefs about the world, particularly its social dimensions, are responsive to signalling pressures emerging from interpersonal action coordination, not to the intrinsic representational utility or practical efficiency of the beliefs in question (McGeer 2015). As a result, much of what we think about the world generally conforms to what other close conspecifics think about the world, otherwise we would not be able to easily coordinate our actions (i.e. adopt aligned coordination equilibria) because the action space would be too large to match strategies.

This is not to suggest that people are unthinking clones of one another. Rather, the range of selves available in any cultural context is responsive to and constrained by broader coordination considerations. Agents act to regulate (shape) one another to adopt and adhere to specific belief sets so that coordination is possible. On the mindshaping account, cognitively mature human agents regulate their beliefs and self-descriptions to fit shared models thereby enabling coordination, and in order to signal coordination status. This signalling and regulation function takes precedence over the content of beliefs, meaning that in many cases it is the function of a belief as an effective signal of coordination suitability that determines its adoption (Funkhauser 2022; Williams 2022a). Thus the mechanisms that enable coordination exert a powerful influence on quotidian ontologies, particularly concerning symbolic, or creedal beliefs (Joshi 2022). Mindshaping *both* restricts behavioural range *and* allows beliefs to function as signals of coordination suitability. Fundamentally, this means that what it is to be a human, and what forms the human self can take, are partly determined by the dynamics of the coordinating groups in which individuals are embedded (Dennett 2017), not by some pre-existing self that enters the social domain fully formed.

The most basic form of mindshaping, which likely enabled the very earliest hominin coordination, is observing and imitating a model (Zawidzki 2013). At its core, imitation is a form of communication. The imitating agent replicates some information about the behaviour of the model. In general, all mindshaping is powered by communication – the model the agent is being shaped to adopt must be transferred in some manner. Contemporary humans make use of a host of rich communicative tools to distribute the shared behavioural models that enable coordination and signal group status. This link, namely between mindshaping and

communication infrastructure, means that the evolution of human society can be partly explained by reference to an upward trajectory of changes in the scale and scope of communication technologies. These technologies enabled the ever-wider dissemination of propositional attitude regulating narratives and symbolic resources used for signalling coordination potential – large-scale mindshaping – thereby enabling coordinating groups to enlarge in size and grow in power.

Thus, the scales of group action possible and the signals groups adopt relate to the particular communication affordances available for use. This is because the structure of communication technologies, their degree of centralisation, scope, efficiency and scale, interact to generate and constrain signalling possibilities and incentives. In other words, the types, or the structures, of the communication technologies utilised by agents seeking to coordinate, play a significant role in determining which coordination equilibria are salient and the types of beliefs that are optimal for signalling group affiliation. For example, in a given communication context, beliefs that undermine large-scale, society-wide coordination can become locally entrenched if their use enhances coordination inside a sub-group (Hardin 1995). In other words, the specific communication structures generated by different technologies of communication interact with and modify interpersonal signalling efforts, in turn enabling, blocking or altering coordination prospects across various levels of social integration.

The advent, rapid spread and adoption of internet-based mass communication technologies is the most recent consequential twist of the ratchet towards larger-scale and more efficient communication technologies. A core claim of this thesis is that mindshaping theory can illuminate how, by interacting with and changing patterns of communication and information dissemination, these technologies fundamentally

modify the processes that generate individual selves and groups. Furthermore, internet-based mass communication possesses several features that make the manner in which it interacts with mindshaping processes, and the types of coordination it facilitates, particularly problematic and normatively significant.

As will be examined in detail in Chapter 4, communication online takes place on a substrate that is opaque, complex and deeply mediated. Social media in their various forms have made interpersonal communication lower cost and faster while significantly increasing the potential audiences for individual speech acts. Simultaneously they have modified how agents are exposed to information about their world and increased the control individuals have over their day-to-day communication partners. It is a communication environment that, due to the volume of information it contains, must be structured by content selection algorithms, which are designed by private entities that control the dominant communication platforms. These algorithms are optimised to further the private interests of these entities, central of which is the capture and maintenance of user attention (Bhargava & Velasquez 2021).

This reshaping of informational dynamics changes the signalling landscape in which people act to coordinate. The upshot is to prioritise and make highly visible signals that capture attention. In effect, the online communication environment acts as a highly efficient selection device for specific types of information, or signals, prioritising information that captures attention regardless of its broader effects on individual and group welfare (Acerbi 2019). Simultaneously, the combination of sheer signal volume online and the manner in which it delinks communication from real-world outcomes means that signals of coordination suitability, in general, are

devalued. In other words, in an environment in which signals are cheap, easily faked, and abundant, it becomes more difficult to honestly advertise group membership.

At the limit, these processes combine to make available and incentivise the use of, more extreme or partisan signals of group membership. These outcomes emerge due to the strategic activity of agents seeking to coordinate action in this specific information domain. However, because the effects of mindshaping mean that the signals adopted by agents to advertise coordination suitability restructure their ontologies when group signalling equilibria exhibit large enough divergences, then the agents in different groups can eventually come to inhabit incommensurable discursive spaces. Beyond increasing the potential for inter-group strife, from a macro perspective, such a set of processes generates significant coordination noise, thereby undermining large-scale social welfare.

In the framework developed in this thesis, outcomes of this sort are linked more robustly to the *structure* of the online communication domain, than to the *content* of information propagated. Thus, the form of communication enabled by online platforms emerges as a more significant driver of pernicious outcomes than the specific information these platforms communicate. Arguably, the shape of public discourse is now mainly determined by the outcomes of processes of this nature.

Across the rest of this thesis, I expand on the claims developed above, arguing that mindshaping theory can be used to shed light on how digital mass communication technologies interact with the formation of coordinating groups of agents. I show that due to specific features of their design, which, as we will see in Chapter 5, emerge from the broader economic incentive structure in which these technologies developed, they now generate significant pernicious outcomes that require investigation and, if possible, regulation and redesign. My explanatory

approach explicitly breaks with the dominant mindreading-based explanation of coordination, which I argue lies at the root of a general misconception of how these new communication technologies generate deleterious effects for societies.

In what follows, I utilise an explicitly pragmatist lens to examine concrete outcomes for human groups generated by their use of specific forms of technology. As such, my work can be situated within an extant sub-discipline known as pragmatist philosophy of technology. Philosophers in this tradition build on John Dewey's work on technology which examines how it extends or interacts with human capacities for inquiry (Keulartz et al. 2004; Pitt 2011; Hickman 1990). For Dewey, advances in technology are the key driver of progress in the modern world, and it is technology, not class struggle, that is the primary means by which humanity could attain a state of functional liberty for all. Following this, philosophers who develop a pragmatist analysis of technology engage in "experimental inquiry into what things do now and how they can be improved so that undesirable or problematic situations can be ameliorated, especially in their social and environmental implications." (Hickman 2017: 500). Fundamentally this approach is "concerned with the creative capacity for the innovation and invention of vocabularies which provide new meanings and open new perspectives" (Keulartz et al. 2004: 18).[3] The normative aspect central to my work echoes this approach, foregrounding the outcomes and effects of technologies on the lives of individual humans. I use mindshaping as a novel and empirically rigorous vocabulary for

[3] Using a pragmatist lens sets my work apart from the two other main approaches in philosophy of technology: analytic philosophy of technology which seeks to understand technology as primarily a practice, the practice of engineering and which often views technology in apolitical terms (e.g. Franssen & Koller 2016); and continental philosophy of technology which generally seeks to understand how technologies alter or interfere with human agency, in particular our moral agency and understands technology in explicitly normative terms (e.g. Verbeek 2011).

describing and attempting to improve or ameliorate pernicious elements that arise as a result of technological change.

Tight relationships between coordinating groups, selves and beliefs allow coordination solutions, once they emerge, to remain powerful and robust. However, this linkage, between ontology and environment, between individual beliefs and collective beliefs, means that we must enlarge the moral framework with which we parse the world. On such an understanding, individual human subjects, though necessary, are not the primary engines of belief formation. In addition, environmental affordances (including other humans and communication structures), which interact with the dissemination of models for actions, beliefs and selves, play a constitutive role in forming human ontologies. As a result, our specific environments of interaction, both designed and natural, largely determine societal structures and the agents which populate them. This is an essential foundational understanding for any effort to normatively reform institutions.

1.2 Coordination and Applied Game Theory

The term coordination, as used in this work, is a technical concept drawn from game theory. Although I will not utilise any formal mathematics, my discussion, especially in Chapter 4 which examines communication dynamics in the online context, draws heavily on the game theoretic understanding of social interaction as strategic and relying on signalling dynamics. In placing coordination at the centre of ontology formation, via mindshaping, my work thus can be understood as applied game theory. To make such an approach intelligible it is necessary to do some groundwork, thus this section outlines the history of coordination as understood by game theorists and the general challenges it raises.

Game theory is the mathematics of strategic interaction. Applied theorists try to use this knowledge to carry out empirical social science. Efforts to pare interaction down to its structural core has revealed that coordination, and the resolution of the dilemmas it throws up, is a foundational feature of social interaction. Coordination is strategic interaction in which agents must choose actions that, at least partially, align to bring about mutually beneficial outcomes. Thomas Schelling, in his early study of coordination (1960), has suggested that pure coordination and pure conflict lie at opposite ends of a spectrum classifying types of strategic interaction. Pure conflict can be understood as "all gained for one" and pure coordination as "all benefitting together" (1960: 84).

Some games can have several different ways for players to coordinate their actions – that is, they can have multiple equilibria – and these solutions may have different payoffs, some of which may be higher Pareto ranked. In addition, agents can benefit from coordination even when the distribution of payoffs from successful coordination is inequitable and more equitable equilibria were potentially achievable. For example, if two people take turns rowing a boat across a lake so large that neither could have managed to make it alone, but one participant does the bulk of the work, both parties still benefit from their coordination, even though the work-shy rower benefits more.

Pure coordination dilemmas, games in which players are indifferent as to which equilibrium is chosen, were brought to widespread philosophical attention by David Lewis (1969). Lewis was particularly interested in the role of coordination around arbitrary conventions in the development of language and scientific reference. Most nouns are phonologically arbitrary, as the sound that refers to the referent could be anything pronounceable just so long as the speakers agree on the

meaning. Thus, speakers must coordinate, but the various solutions (equilibria) are Pareto-indifferent: “cow” could refer to a cat just as well as the reverse so long as we all agree on which means which (and indeed can agree that anything means anything, the unexplained original coordination dilemma that underpins Lewis’s hypotheses).

Lewis recognised that coordination dilemmas are pervasive in human affairs, not just in the case of language, and sought to explain how rational agents, as he understood them, go about solving them. He proposed that such dilemmas are solved by what he called conventions. When humans are faced with a coordination dilemma, they bring a history of interaction to the game – there is almost always some precedent, either with the same player, a similar game, or a shared story – that is perceived as relevant. This precedent operates to make one of the coordination solutions stand out as more salient than the others, it becomes a focal point for the players (Schelling 1960). A salient solution, through repeated successful use, eventually becomes a convention.

The standard example of a convention that resolves a coordination dilemma is the rules of the road. Though now enshrined in law, the convention to drive on the left (or right) is a solution to the coordination problem of how to safely pass in opposite directions. Though it seems trivial – choosing to drive on one side or the other makes sense – the fact that we do coordinate on a single side invests the solution or convention with great power. Because the majority utilise the convention, those who do not, or cannot, face the prospect of harm in the form of a collision. Furthermore, the solution is powerful as the scale of its usage makes the process of switching very difficult (though not impossible – Sweden did so in 1967). This shows how conventions are invested with power by their continued usage. The more

widespread they become, the more power they have. Thus, the alteration of a powerful convention requires large-scale effort to disseminate information in order to make most participants aware of the upcoming switch. Focal points are central to explaining real-world coordination inside cultural groups, and their nature and emergence will be taken up in more detail in Chapter 2.

As we have seen, in pure coordination the specific solution chosen is arbitrary where welfare is concerned. However, interesting coordination dilemmas, of the sort that characterise social and political life, do not usually have this form. In such dilemmas different players prefer different equilibria and potentially all players can get trapped in Pareto-inefficient equilibria. Understanding how coordination is achieved, and how conventions or focal points emerge and are discovered by agents, is, as Schelling and Lewis recognised, a central challenge for game theory. Agents participating in coordination games do not necessarily have access to shared information about states of the world and often they will have imperfect knowledge about other players' utility functions. Thus players must rely on subjective estimations of relative probabilities of the strategy profiles of other players. This makes convergence equilibria theoretically surprising. Yet human agents are immersed in a world of coordination dilemmas – from passing one another on a busy footpath to determining divisions of labour in the home or workplace – which they seem to often resolve fluidly and without apparent cognitive burden. How do people do this? In other words, how can they jointly arrive at specific equilibria in multiple equilibrium coordination games? Adequately explaining this is a requirement for game theory if it is to act as a useful theoretical toolkit for the modelling of small-scale social interactions, i.e. if it is to be of use in experimental social science.

The most widely known and general game solution concept, Nash equilibrium (NE), is too brittle to explain coordination outside of highly rule-governed institutional contexts. As a result, additional theoretical developments were required to account for players' inferences concerning strategy vectors. Leonard Savage (1954) was the first to demonstrate how to incorporate subjective probabilities within the established framework for calculating expected utility. John Harsanyi (1967) then built on Savage's work with a method for solving games made up of utility maximisers of the sort Savage had modelled. In particular, Harsanyi incorporated the fact that players make inferences from observing one another's actions in extensive-form games. In general, it is assumed that players are Bayesian learners, that is, they update their models of the world, or their ideas about the probabilities of various states of the world obtaining, as they observe what other players reveal about their assessments of states obtaining through their moves. Repeated games, of which real-life strategic interaction is often an example, offer players ample opportunities to engage in Bayesian learning about the various states of the world and the games they are playing.

Robert Aumann (1974, 1987), further building on the work of Savage and Harsanyi, proposed a solution concept by which players can solve coordination games by discovering what he calls a "correlated equilibrium". The idea is that if the players know that other players are Bayesian learners with common priors and they receive shared signals then they can identify a coordinated equilibrium prior to play. If established, a coordinated equilibrium limits the possible equilibria agents can choose to a subset of solutions, thereby making coordination tractable. However, Aumann's account brings the propagation of shared signals into focus, for how is it

that actual people can discover shared signals without explicitly communicating their preferences to one another?

The traditional resolution to this problem appeals to the concept of mindreading (Guala 2016), discussed above. This is the idea that in any bout of interaction, agents use their theory of mind skills to determine the intentions, or propositional attitudes, of the other agents, by observing behavioural signals. Essentially, humans observe others and attempt to infer their internal mind states, and their intentions, and once they discover these they then act accordingly, aiming to adopt coordinated strategies to maximise their payoffs. However, such an account faces serious, potentially insurmountable problems, particularly with regard to computational tractability. Specifically, many distinct propositional attitudes can be congruent with some observed behaviour, making it impossible for agents to identify exactly which belief is driving behaviour in any single case, and thus decide which strategy to use. Furthermore, mindreading, if based on Bayesian inference, requires agents to share common priors if Aumann's solution is to apply. However, mindreading itself lacks the explanatory resources with which to provide a way for those agents to establish common priors as the agents doing to reading are isolated in advance of the act. Thus it seems to be undermined by circularity.

The alternative, increasingly influential explanation of coordination resolution relies on minds*haping*. As we saw in the previous section, mindshaping processes allow agents to dynamically form shared beliefs on the fly through dynamic micro-negotiations. This type of shared signalling and preference set alignment neatly slots into Aumann's idea that converging on shared belief sets via signalling lies at the root of the resolution of multiple equilibria coordination dilemmas. His basic idea is that agents use common signals that each has seen in

advance of play to extract probability distributions around strategy vectors. Mindshaping, insofar as it "supplies agents in a social influence network with shared signals that support correlation" (Ross & Stirling 2023: 16)[4], is thus a crucial component in the resolution of coordination dilemmas. If agents enter a game with common prior beliefs, they can solve the equilibrium selection problem as the solution sets are significantly narrowed. In this sense coordination, in particular the coordination of prior beliefs, is central to all strategic games, as it is required to narrow down the range of possible equilibria. We will return to the use of mindshaping as opposed to mindreading for resolving coordination dilemmas in the following chapter.

As we have seen, once a coordination equilibrium is established it benefits all participants to follow it, thus making it difficult to alter, as all must change simultaneously to avoid utility losses. Furthermore, all players are immersed in a vast set of repeated, entangled coordination games, and a move in one affects the others, making established equilibria even more difficult to dislodge. This means that suboptimal solutions can often be preserved if they emerge as coordination-enabling conventions. The possibility of social groups becoming locked into a suboptimal equilibrium can be illustrated using a toy game known as Hi-lo.

[4] Ross and Stirling (2021, 2023) develop a specific, novel, extension of game theory known as Conditional Game Theory (CGT), which they frame as a formal model of mindshaping processes. In addition, they use CGT to resolve two key obstacles faced by Aumann's correlated equilibrium solution concept for application to games among unacquainted people. Correlated equilibrium relies upon the idea that players share common priors (dubbed the Harsanyi Doctrine) and that they adhere to Savage's expected utility theory (EUT). Both of these assumptions have come under challenge. The common priors assumption is an assumption in Aumann's model, and that people generally adhere to EUT-derived reasoning has not been borne out in the laboratory (Harrisson and Ross 2017). Ross and Stirling use CGT to resolve both issues. Specifically, through immersion in pre-play, i.e. mindshaping, agents can create shared signals, thereby providing common priors, and this same pre-play furnishes agents with transmission matrices which factor in potential deviations from EUT, allowing in-game play inference to assume EUT.

In a Hi-lo coordination dilemma there are two coordinated solutions. However, they are not Pareto-indifferent. There is an inferior and a superior solution. See Table 2: with a payout of 10 for both players, (Hi, Hi) is Pareto superior. Coordination dilemmas with Pareto-ranked solutions are more commonly observed than pure coordination games in social and political coordination. However, the logic of coordination sets up the potential for inferior outcomes to become entrenched. The existence of an inferior solution coupled with the stability of solutions to coordination dilemmas means that if an inferior solution becomes established it is difficult to dislodge. The initial selection of a suboptimal solution may occur if a Pareto-inferior solution appears more salient or if external factors alter the payoffs, but not the structure of the game after the solution has been selected. Then the inferior solution can become entrenched, and no individual player can unilaterally flip the game to the other solution.

Table 1

Hi-Lo Game	Hi	Lo
Hi	10, 10	0, 0
Lo	0, 0	1, 1

Making such equilibrium changes especially difficult is the role of norms in maintaining coordination in large-scale society. Social norms can be understood as evolved collective rationalisation of equilibrium solutions to coordination dilemmas commonly faced by cultural groups (Bicchieri 2005). Norms often support high levels of ostensibly altruistic behaviour such as promise keeping or refraining from theft (though as Binmore (2005) argues these equilibria must be compatible with strategic play by self-interested individuals over the long run). It is for this reason

that many norms are often understood as relatively binding rules for behaviour (Guala 2016); otherwise, their violation for personal gain would erode their stability and value as coordination devices. As a result, deviance from many social norms is sanctioned (either by the person themselves, in the form of shame, or by others, in the form of public disapproval – a relatively low cost but highly effective punishment for hyper-social animals like humans) thereby maintaining their utility as robust guides to culturally specific coordination equilibria. In this vein, Brian Skyrms (1996, 2003) presents detailed and influential modelling work that illustrates how, over the long run, game theory dynamics can enable various normative structures relating to commonly observed features of human social life like fairness and property rights to evolve.

From a large-scale perspective, the role of norms as behavioural rules can lock in inequitable, or inefficient social structures. For example, existing power structures or modes of organization may become suboptimal over time, perhaps due to demographic changes or advances in technology. However, due to their status as conventions that often carry normative force, these now inefficient structures are very difficult to dislodge. Moreover, a situation like this can be perpetuated by the phenomenon of preference falsification, whereby people support dominant public opinions or preferences, even though their actual opinions differ privately (Kuran 1997). This can happen if some players initially misrepresent their preferences in order to signal coordination fitness and reap the benefits of group action, or if circumstances change making some preferences less attractive. However, because a majority appear to support the sub-optimal preference, the established norm, individual members of the group are disincentivised from unilaterally dissenting, even if those individuals, unbeknownst to one another, make up a majority.

However, if information dissemination structures can be effectively used to generate sufficient common knowledge of the actual support for the alternative preference, sudden reversals of public opinion can occur (Kuran 1997; Chwe 2001). Mindshaping, as we will see, is crucial to such dynamics.

It is a process of this sort that arguably led to the rapid dissolution of the Eastern Bloc in 1989. In that case, the initially imposed coordination structures were for a time relatively efficient, precipitating rapid industrialization and high levels of economic growth (Ofer 1987). This allowed them to persist and thus become conventions. However, through a combination of mismanagement, changes in external material conditions, and excessive repression, the institutions of authoritarian state socialism became increasingly inefficient. However, the entrenched nature of the coordination conventions, coupled with widespread preference falsification and state control of common knowledge generating media allowed this inefficient state of affairs to persist for several decades. The rise and rapid dissolution of the Eastern Bloc and the USSR is a testament to the power of successful coordination to make and unmake life-worlds. Once the right circumstances came together, such as public communicative displays and widespread knowledge of alternative more efficient coordination solutions, these regimes fell. This example demonstrates the importance of understanding how signalling dynamics, and how signals are channelled or suppressed, may enable or block large-scale coordination.

Fundamentally, in order to coordinate at all, the agents involved must be able to clearly and robustly signal to other agents that they are using similar, or correlated, scripts for action, i.e. scripts that align. The importance of coordination means that signalling pressures have come to determine, in many cases, the structure

of social life, sometimes for the worse. For example, the pressure to efficiently signal coordination status has been suggested to lie at the root of social inequality, particularly inequalities linked to race/ethnicity and gender. This is because dividing a society into easily recognizable types provides low-cost and readily perceived signals which can help interactants break symmetries in cases of complementary coordination dilemmas (O'Connor 2019).

Complementary coordination dilemmas occur when participants need to carry out different but complementary actions to coordinate. For example, to effectively run a typical household a range of different tasks of varying difficulty must be carried out and the agents involved make their lives easier if they do not have to negotiate the division of labour before each task is undertaken – they benefit from coordination around a task division scheme. This type of coordination can be made more fluid or low cost if each participant automatically perceives their role in the interaction, thereby avoiding the use of time consuming and potentially intractable negotiation. Cailin O'Connor (2019) uses a game theoretic analysis to propose that social typing, like gender roles, emerged and now endures primarily to solve coordination dilemmas of this form. The existence of an easily recognizable and difficult-to-fake type such as gender or ethnicity is useful for efficiently breaking issues of symmetry in coordination dilemmas, particularly in repeated coordination dilemmas of the sort that characterise quotidian life. However, due to the power differentials that various tasks can inadvertently generate (for example, hunting with weapons trains an agent to be adept at violence, and childcare significantly limits the caregiver's potential set of social opportunities) the division of labour can generate inequalities. Furthermore, the normative component of

coordination equilibria means that these divisions are seen not as useful tools, but as natural and binding norms, further entrenching any inequality they may generate.

The fact that salient types, and particularly labour division across gender lines, are universal features of human culture suggests that inequity in society is not something that can simply be erased (O'Connor 2019). Perhaps some forms of inequity can be remedied, but other forms will emerge due to the inescapable logic of coordination dilemma resolution. The low cost and efficient nature of type-based resolution of coordination games means that classifications of this sort will tend to emerge over the medium term, even if inequitable categorisation is ameliorated in the short run. In other words, the power that coordination generates may hurt some of those it helps, and avoiding, or minimising, the negative externalities it can create requires continued vigilance and redress coupled with a keen awareness of coordination dynamics and signalling potentials.

The logic of coordination, and in particular the importance of reliable signals of coordination suitability, also plays a significant role in fuelling inter-group conflict. Often group identity is established in opposition to some other group, conceptualised as being inferior or even sub-human. Inside a cultural group, costly other-group harming actions are then rewarded as they demonstrate trustworthiness and thus coordination potential. Effectively violence acts as a very costly, and thus honest, signal of in-group status. The logic of coordination in this case can drive the members of a coordinating group to carry out atrocities to more effectively signal that they are good coordinators. For example, Russell Hardin (1995) suggests that group identification coupled with the gains from coordination played a contributing role in the attempted genocides in Rwanda and Yugoslavia. In both cases, Hardin believes that a general perception of instability related to state failures (and the

cynical urging of self-interested political agents) incentivised the salient subgroups to deploy increasingly robust signals of in-group membership, signals that involved deadly out-group repression. The history of humanity suggests that negative outcomes resulting from successful group coordination leading to intergroup conflict are at least as common as issues arising from failed collective action within groups.

Contributing to such dynamics is the fact that ecological threats tend to tighten group signalling norms, a phenomenon that has been robustly demonstrated to occur cross-culturally (Gelfand et al. 2011). In societies that face high ecological uncertainty, for example, as a result of endemic poverty or environmental precariousness, social norms are more strictly enforced (Gelfand et al. 2017). Conversely, in cultural groups that have not faced significant ecological threats in their recent history, or experienced high levels of material stability, norm enforcement is usually less strict and behavioural diversity is tolerated. We can link this phenomenon directly to coordination: for groups facing peril the value of effective coordination is enhanced; tight norm enforcement enables coordination because it restricts the range of strategies available to interactants – norms serve as effective guides to coordination equilibria. Similarly, once threats meaningfully recede, agents can begin to express a broader range of idiosyncratic behavioural traits as this does not interfere with the prospects of overall group survival. This phenomenon is not necessarily problematic. The ability of groups to make behavioural trade-offs depending on external circumstances is an effective survival strategy. However, the tendency for groups to trend towards looser norms in (materially) good times implies that a somewhat looser norm structure may be preferential when agents can bring such a state about.

This examination of the effects of coordination on the structure of social formations demonstrates that in a surprisingly varied set of cases, negative societal outcomes can often be understood as emerging not as a result of intrinsic malice, but due to the logic of coordination. For this reason, "we should often become suspicious of group success in mobilizing individuals just because individual incentives typically run counter to group action…the incentives for group commitment may be perverse and destructive. Successful collective action can sometimes be a wonderful achievement. But it can also be a dreadful one, the source of great harm, even to those who succeed in the collective action." (Hardin 1995: 214). The potential downsides that can emerge from successful coordination mean that understanding how coordination is achieved and how those processes may be misdirected, stymied, or hijacked is a pressing task. The generation of coordination noise can block the emergence or adoption of optimal society-level coordination solutions, or push agents to adopt more costly and thus honest and easily interpreted signals. Understanding how agents manage to engage in quotidian coordination is therefore an important part of social analysis. Of particular interest, given the topic of this thesis, is how the specific structure of a communication environment may contribute to the emergence and persistence of suboptimal coordination equilibria. Specifically, the manner in which internet communication affects or modifies the incentive structures that determine signal use and coordination equilibria will be the topic of Chapter 4.

1.3 Structural Overview

The final task outstanding in this introductory chapter is to provide an overview of the structure and content of the thesis as it develops from here.

Chapter 2 takes mindshaping as its primary focus. I describe its intellectual history, its role in philosophy of mind, and examine the main processes via which it operates. As we will see, mindshaping can help game theorists resolve the deep dilemma they face in explaining fluid interpersonal coordination, replacing flawed, but still widespread, mindreading-based explanations. Mindshaping can be seen as part of a general trend in philosophy of mind, which moves away from spectatorial explanations of social cognition in favour of more interactionist approaches which understand agents as knitted into and constrained by interpersonal influence networks. Mindshaping, however, goes beyond merely explaining social cognition – it also provides a rich account of the genesis of selfhood showing how social groups constrain and guide its emergence. I describe the main processes of mindshaping, specifically focussing on conformist mindshaping, as this will be especially pertinent in the context of online communication. As we will see, the extant scholarship on the mindshaping-coordination framework has been primarily focused on how the theory can resolve issues that arise for mindreading-based accounts of interpersonal coordination. As a result, surprisingly little use has been made of mindshaping to examine concrete cases beyond philosophy of mind or game theory.

Chapter 3 then places mindshaping in the context of the evolution of complex cognition. As we will see, mindshaping, though it is not explicitly referred to by the leading theories of how complex human cognition evolved, is both consilient with these theories, and in many cases implicitly presupposed by them. I show how mindshaping neatly slots into a range of approaches to explaining the emergence of complex cognition. The two, mutually reinforcing, purposes of such an exercise are: (1) to show how mindshaping aligns with influential and widely cited work in cultural evolution, which serves to strengthen the core claims of mindshaping, and

(2) to use the scholarship in cultural evolution to shed further light on mindshaping itself. This is because theorists working in cultural evolution place significant emphasis on the strategy of model choice in social learning; mindshaping itself is a species of social learning and thus this work is highly relevant, helping explain how agents come to choose what shapes their mind, and the strategic nature of these choices.

Chapter 4 then turns to the more applied aspect of my work. Specifically, I build on the framework developed in the previous chapters to critically examine digital communication technology. First, I examine how beliefs serve a strategic signalling role – the beliefs an agent holds often act as a signal, and, as such, are divorced from representationalist concerns. Thus, largely strategic concerns guide mindshaping and the formation of individual ontologies. However, for most of human history this strategic interaction was restricted to face-to-face peer groups, and the communication technologies required to scale up these processes emerged slowly. I examine how the evolution of increasingly powerful communication technologies has thus closely mirrored the emergence of increasingly complex and politically distinct human social formations – the manner in which signals can be dispersed and centralised constrains the possible scope of coordination and the potential coordination equilibria within reach. This mindshaping informed framework is then used to examine digital communication proper. I show how, given the strategic role of belief formation that mindshaping underwrites, the specific structure of communication affordances interacts with the manner in which agents use beliefs as signals to advertise group status and coordination suitability.

This perspective breaks with the dominant mindreading-driven approach which sees digital technologies as merely tools for the use of agents. The

mindshaping view shifts the primary locus of concern from the specific *content* these networks allow to spread, to the *structure* of these communication affordances and the specific sorts of agents or groups they shape into being. I use this mindshaping informed understanding of communication, signalling incentives and equilibrium dynamics to make concrete proposals for the redesign of these affordances, aimed at minimising inter-group strife, coordination noise, and extreme anti-out-group discourse in online communication. Afterwards, I briefly examine how these claims can be operationalised to help further support the mindshaping framework in general.

Finally, Chapter 5 turns to the possibilities and potentialities of these proposed reforms. Examining the real-world applicability of potential reforms necessarily draws us into the realm of the political[5], a territory not often associated with philosophy of mind. First, I examine the specific economic context in which these technologies emerged – characterised by neoliberalism – describing the incentives this generated which in turn influenced their design. Any effort to reform these affordances will need to understand and alter this incentive structure, thereby pushing providers to implement structural design changes. However, as we will see, the impact of neoliberal governance on the initial design of these technologies may have locked in their problematic form for the foreseeable future.

I then turn to a discussion of possible implications of the proposed reforms. If these technologies can be redesigned to minimise coordination noise and maximise inter-group coordination then they hold significant promise for the prospects of

[5] In what follows, I use the term "political" in a way that is closer to the concept of ideology than the mundane sense of the political. Thus, a claim that is political as I will understand this, is one that either has ideological content or can be used to further a specific ideology. The political aspect of philosophy of mind is its ideological content, or the parts of it that can be used to promote some specific ideology.

realising a more egalitarian social contract. However, this potential promise is counterposed by significant threats. I briefly examine these threats and promises and show how they lie upon spectrums of possibilities. I suggest that liberal democratic polities will need to carefully navigate a path through these various outcomes, and mindshaping-informed design can be of use in guiding such efforts. Finally, I conclude with a summary of the main points argued for and the core themes that run through the work.

Chapter 2 – The Mindshaping-Coordination Complex

This chapter outlines, defends and examines the implications of the claim that the human ability to engage in fluid interpersonal coordination is due not to mindreading – a spectatorial form of interpersonal interpretation that utilises a rich theory of mind – but rather relies on the processes of mindshaping. Mindshaping proposes that human minds are culturally evolved, virtual entities that are partly constituted and dynamically shaped by scaffolds beyond the brain, including other humans. The role of these external, but shared, entities in the constitution of minds ensures that humans remain sufficiently alike to fluidly interact without utilising an unwieldy and underdetermined theory of mind. Mindshaping thus explicitly seeks to bring about an externalist reconfiguration of our understanding of human social cognition, setting itself in opposition to internalist and still widespread mindreading-based accounts of such competencies.

In what follows, in Section 2.1, I examine in more detail the role of mindreading in explaining interpersonal coordination. As we have briefly seen, such an account is faced with intractable obstacles when posited as a means of resolving coordination dilemmas. Game theory thus requires an alternative explanation of coordination if it is to be of use in the social sciences. An account that understands humans as knitted into influence networks which constrain their action sets, thus determining their coordination options, is a more viable explanation, one which avoids the problems faced by a mindreading-centric account.

In Section 2.2, I examine the broad historical and intellectual trajectory that led to the emergence and influence of mindreading theory, tracing its rise to prominence and current proposed (partial) eclipse. Several recent attacks on mindreading-centric accounts of human social competence emphasise the key

importance of interaction in human life and cognitive processes (Leudar, Costall, & Francis 2004; Gallagher 2004; Hutto 2008; Fenici 2017; Fenici & Zawidzki 2021). We will see that these interactionist accounts see social cognition as a form of, and as reliant upon, shared mind making, which in turn, suggests a reconceptualization of the role of propositional attitude (PA) attribution – a shift from epistemic appraisal to interpersonal regulation. As we will see, the views clustered around the concept of mindshaping offer the best yet formulated conception of the regulative role of PA attribution (Mameli 2001; McGeer 2001, 2007, 2020; Zawidzki 2008, 2013, 2020).

Following this, in Section 2.3, I examine the theory of mindshaping in greater detail, setting out its core claims, specifically concerning the "mind". This account vitiates any appeal to a standalone, pre-existing mind possessing PAs to be "read" in advance of mindshaping processes. Instead, social coordination relies on the pluralist craft of folk psychology (Andrews 2015), culminating in the intentional stance (Dennett 1987, 2017; Zawidzki 2018), a skillset that sidesteps appeals to discrete mindstates as causally responsible for individual behaviour.

Having reconceptualised PAs, in Section 2.4 I examine mindshaping beyond the regulative attribution of beliefs and desires to other agents. Broadly speaking, mindshaping can be broken down into the processes of imitation, pedagogy, and conformism (Zawidzki 2013). As we will see, conformism will be especially relevant for the coming investigation of mass communication technology. This is for two reasons: first, its mechanisms operate pervasively, automatically and unconsciously, making them opaque and thus difficult to counter. Second, they operate to disseminate and regulate the use of beliefs as signals, a feature that plays a crucial role in generating pernicious outcomes in the context of digital information communication technology.

2.1 Mindreading and Coordination

As we have seen a central concept used to explain real-world coordination in the game theory literature is that of a focal point. First proposed by Thomas Schelling, in his ground-breaking *Strategy of Conflict* (1960), a focal point is a feature of some strategy that an agent finds salient, and thinks will also be salient to the other player. For example, in Schelling's original experiments participants were asked to coordinate a meeting the following day in New York with another person, without having agreed on a time or location in advance. Most participants suggested Grand Central Station at noon, a time and location which functioned as a salient focal point for interactants in the 1960s. However, despite the subsequent widespread adoption of focal points by theorists, the account Schelling gives of them has resisted successful formalization. In other words, how it is that focal points function in interpersonal coordination is not clearly understood. Thus, the concept remains informal, offering at best an ad hoc explanation (Larrouy & Lecouteux 2018; Guala 2018). Moreover, Schelling's influential description of focal points has been characterized as metaphorical and somewhat mysterious (Sugden & Zamarrón 2006). How is it that agents can identify and choose a focal point in a game? If focal points are going to be central in explaining coordination then we need some answer to this question.

At various points in his book Schelling proposes that rather than trying to *anticipate* each other's choices, as in a standard understanding of strategic interaction, during coordination dilemmas the players instead *reason together* in an effort to find a focal point. This explanation has been vindicated in the lab by Mehta et al. (1994), where investigators set out to try and discover how focal points emerge.

First one group of participants was asked a series of questions, for example, to name a number or a year that stands out to them. The aim was to identify primary focal points for various categories, things that naturally appear salient to people (or at least to the subjects of the experiment). The participants tended to respond with examples that were personally relevant to them in some way, for instance, their year of birth or a number they had chosen before in a random number generation task. Then a separate group of people were asked the same questions, but they were instructed to try and coordinate their answers with a partner with whom they could not communicate. In this case, the players gave answers that were much less idiosyncratic, for instance, the year the experiment took place, or the number 1. In this treatment, the two participants often succeeded in coordinating, whereas in the first treatment, the answers rarely matched up.

The results of these experiments suggest that when people explicitly try to coordinate, they do not rely on some predetermined focal point known to all prior to the interaction. Instead, they use a type of socially engaged reasoning applicable to the context – in Schelling's terms, they really do seem to *reason together*. Focal points thus appear to be partly the product of the dynamic interaction of agents in the moment. Participants do not just draw on what appears privately salient to them. Instead, they choose strategies that they think others would also choose. Yet explaining how people could successfully do this seems to require some account of how people could access or predict the contents of other minds.

In general, when describing focal points Schelling makes use of phrases and examples that attempt to *show* us, rather than *tell* us, how they emerge. For example, he describes them as the result of a "meeting of minds" (1960: 83) or a "spark of recognition" between two players (163). It is as if "Schelling is struggling to

articulate a vividly perceived idea…which resists literal formulation" (Sugden & Zamarrón 2006: 612). Following this line of thought Sugden and Zamarrón propose that Schelling's account is driven by pragmatic concerns rather than deductive ones. Schelling pragmatically draws our attention to a feature of our social practice, focal points, that, once identified, has the potential to be useful in explaining coordination dilemma resolution.

However, Schelling did not just choose to present focal points metaphorically because this was the most effective way to characterize them. There is a more specific reason for this non-standard approach: Schelling is forced to *show* us focal points in action rather than analyse them in formal terms due to the inadequate conceptual toolkit available in the wider literature of the era. The primary framework for understanding human social cognition at the time was as cognitively solipsistic agents who are the possessors of private internal states, labelled as propositional attitudes. However, such an approach is insufficient to explain how an interactive and necessarily shared entity like a focal point could emerge.

This individualistic approach to game theory creates an explanatory problem in the case of coordination as the players' strategies are not determinable from a solipsistic position: one player's best response depends on what the other chooses and vice versa. The internalist conception of human cognition is thus faced with the problem of how agents can access the contents of other minds, how they can tell in advance what the other player might choose. It is this conception of cognition that forces Schelling to talk about a mysterious "meeting of minds". Schelling's work amounts to an exhortation to pay attention to focal points but provides no explanation as to where they come from. To see how this issue is at the core of the

efforts to explain focal points, we can turn to how David Lewis expands and attempts to formalise Schelling's work.

In his (1969) Lewis sets out some guidance as to how players come to form the required common beliefs about the relevant focal point:

> We may achieve coordination by acting on our concordant expectations about each other's actions. And we may acquire those expectations, or correct or corroborate whatever expectations we already have, by putting ourselves in the other fellow's shoes, to the best of our ability. If I know what you believe about the matters of fact that determine the likely effects of your alternative actions, and if I know your preferences among possible outcomes and I know that you possess a modicum of practical rationality, then I can replicate your practical reasoning to figure out what you will probably do, so that I can act appropriately. (27)

On this account, to coordinate we need to know what another "practically rational" person believes and their preferences about possible outcomes. Then we simply run those data through our internal simulator and output likely actions and salient focal points. Here we have a much more explicit statement of how coordination is supposed to occur. And, unlike Schelling, Lewis is explicit about his internalist presuppositions. A philosopher by training, Lewis broadly endorsed an internalist conception of cognition. On his view, mental content is primary and the contents of our linguistic utterances are determined by mental states, not vice versa (Weatherson 2021). What this means is that, according to Lewis, there are private internal states of mind that others, in attempting to coordinate their actions with ours, have to discover. Schelling's "meeting of minds" becomes for Lewis the "knowing of

minds". The attribution of mental states to other players is core to Lewis's account, but how this occurs is only vaguely explained in his text. This, in turn, has been dubbed the "problem of mindreading" for Lewisian-style focal points (Guala 2016, 2018).

Gaula goes on to suggest that the tensions and elisions in Lewis's account can be resolved by recourse to mindreading. As we have seen, the central claim of mindreading is that humans attribute mental states, or PAs, to one another to predict their behaviour. Thus mindreading is an act of prediction ventured through observing and engaging in cognition about the beliefs and desires of another agent (Graham 1993; Baron-Cohen 1995; Goldman 2006; Nichols & Stich 2003). Mindreading is inherently "spectatorial" as it aims at an objective description of the mental states of others and the self (Zawidzki 2019; Fenici 2017). Thus, individual agents use mindreading to engage with an external and separate world. This spectatorial stance implies a reductionist conceptualization of human social interaction wherein individual brains try to determine the contents of other brains. It is reductionist insofar as it proposes that the drivers of action and thoughts can be decomposed into discrete PAs. Thus, for mindreading to truly be at the root of human social competence, PAs must be fixed entities that drive behaviour and can be discovered by other agents.

The main two explanations of how mindreading occurs propose that humans either simulate the thinking of others in their minds or use an explicit theory of how minds work in general to predict the contents of other minds. These are known as simulation theory and theory-theory, respectively. Guala argues that Lewis's talk of "putting ourselves in the other fellow's shoes" can be interpreted as a being a type of simulation theory (2016, 2018, 2020). Specifically, he draws an affinity with the

work of simulation theorist Adam Morton. Morton argues that when human agents try to coordinate, they first think of the most obvious solution, and if there is one, then they attribute identical reasoning to the other player (2003). We can see how this is a simulation: the player assumes that their thinking is a mirror of the other's and uses an answer they would themselves have given as their answer.

The problem with such an explanation, however, is that if it is to be useful in human interaction mindreading must be both fast and accurate, otherwise, it wouldn't be able to quickly resolve the myriad coordination dilemmas faced by human agents. More importantly, it also wouldn't match up with the lived experience of fluid human social interaction. The complexity and range of human behaviours, dispositions and social situations strongly suggests that a general theory capable of linking observed behaviour to beliefs and desires in all cases is likely to be extremely unwieldy (Morton 1996). This undermines the idea that mindreading could be fast enough, particularly in its theory-theory incarnation, to explain human social interaction. Indeed the task of linking up a behaviour with a discrete PA quickly becomes computationally intractable once we realize that any number of PAs are typically compatible with a single observed behaviour (Zawidzki 2013).

Making matters worse, in the context of coordination the intractability issue is more pressing. This is because, unlike the restricted settings of static game theoretic models, human coordination takes place in dynamic contexts. Coordination games are themselves embedded in broader coordination games, and so a move in one game is typically also a move in countless other games (Ross 2005). This adds further fuel to the intractability argument against mindreading: it appears unlikely that humans, using either theory or simulation, could interact as efficiently as they evidently do if they were required to predict the behaviour of their conspecifics in a

world of recursively embedded non-parametric coordination dynamics. Theory-theory is a non-starter in such a context, as any theoretical apparatus would be either too inflexible or too unwieldy to deal with the range of coordination scenarios humans face in social interaction. Morton's idea that to predict your behaviour I merely imagine what I would do, and then assume it will be more or less what you *will* do, is undermined by the evident variability in human behaviour and motivation (Zawidzki 2013: 81). Simulation theory, then, seems to depend on some prior unexplained processes that render agents cognitively alike enough to be able to take their internal reasoning as a good model of their conspecific's reasoning or, to put it differently, depend on the generation of types of humans that could be accurately simulated using one's own conception of the world.

Appearing to implicitly recognize the weaknesses of strong mindreading, Guala's work displays an interesting and suggestive internal trajectory in the way it applies Morton's arguments to explain focal points. Across a sequence of papers Guala gradually gravitates towards a rejection of full-blown mindreading as an adequate explanation of the emergence of focal points. In his (2016) Guala explicitly refers to full-blown mindreading (89, 96) as key to resolving coordination dilemmas but by his (2020) this has been whittled down to "minimal mindreading". The idea here is that "belief-desire attributions are harmful" as they inhibit successful behavioural prediction and instead their main function is not to explain or predict but rather to "justify behaviour, by attributing to agents a set of reasons for doing what they did" (2020: 163). This later approach interprets Morton's simulation theory[6] as

[6] Of the main theorists who argue for simulation theory (see: Gordon (1986) and Goldman (2006)) Morton (2003) is the friendliest to the idea that humans shape one another's minds to conform to a model to enable coordination, namely the mindshaping perspective that informs this work. His (1996) is explicit about the role coordination plays in determining the content of an agent's beliefs and desires. This in part explains the path down which his work has taken Guala, minimising the idea that the reading of pre-existing minds occurs at all.

claiming that agents familiar to one another engage in "belief-less coordination" (Guala 2020: 160). Thus, people coordinate efficiently because they do not think about beliefs *at all*. There is no theory, and indeed no conscious simulation going on, in such an account.

This approach requires an *extremely* minimal concept of mindreading, to the point that characterizing it as such is perhaps misleading. If one agent merely does what seems sensible, for example choosing the higher payoff in a Hi-lo game, and this works because the other agent does likewise, with no access to or representation of the other agent's beliefs, it is unclear what mindreading is being done at all. Indeed, we appear to be back at something like Schelling's insistence that we heed focal points, and don't think too much about formalising how they emerge.

This seems to miss the explanatory mark. How could it be the case that two agents attempting to coordinate could just do so, in a fast and frugal manner, without modelling the mind of the other agent? How can creatures endowed with complex cognitive faculties come to be able to coordinate around one of a huge array of possible actions just by simulating the thoughts of the other agent? We need some explanation of this that does not appeal to an intractable and inflexible mindreading faculty.

2.1.1 Coordination Without Mindreading

Fundamentally, the complex social worlds that humans occupy appear to preclude the viability of fast and accurate mindreading – it seems unlikely that we discover internal states that drive quotidian behaviour in others by using theory application or simulation. Yet, despite this, humans do rely on PAs to explain behaviour, both their own and that of others. However, discarding mindreading does

not mean discarding PAs, or the language of internal states as drivers of behaviour. Instead, we can reformulate the role they play in human cognition. Rather than positing entities that exist "inside" others to be discovered to predict their behaviour, we can understand PAs as playing a primarily *regulatory* function in human affairs (Zawidzki 2012; McGeer 2015). As Tadeusz Zawidzki puts it: "propositional attitude ascription enables human beings to set up regulative ideals that function to mould behaviour so as to make it easier to coordinate with." (2008: 193). So, rather than try to infer from evidence which exact PA is in which individual mind, we instead use PAs as models which we shape our messy and non-transparent inner cognitive and emotional states to fit. PAs work to regulate our behaviours, making them more easily computable. They can do this because in a given group or society PAs are in the public domain. We learn to call a certain set of hypothesised inner states or dispositions "sadness" because we learn that that is what our group calls something approximating these states, and then we go on to fit a range of discrete mental states into such a category.

Core to the feasibility of this regulatory perspective is the phenotypically plastic nature of human cognition. This allows for the utilization of external environmentally-based scaffolds (Hutchins 1995; Clark 2003; Sterelny 2012), one type of which are propositional attitudes (Zawidzki 2013; Dennett 2017). Scaffolds are entities that enable human brains and bodies to engage in actions unavailable to an unscaffolded individual. For example, we scaffold cognitive tasks with objects like a pen and paper or a calculator, and physical tasks using tools like hammers or saws. In the case of social cognition, the task of predicting the behaviour of potentially intractably complex conspecifics is scaffolded through sharing and enforcing models of behaviour. Human phenotypic plasticity means that scaffolds

can reach "in" to shape human minds, thereby bootstrapping the emergence of complex cognition and a host of coordination enabling tools (Sterelny 2012), whilst also ensuring that the agents interacting with the scaffolds are alike in important, coordination enabling, ways. This is needed because cognitive plasticity is a double-edged sword: it makes a species flexible, but too much variation raises the risk that joint projects can never get off the ground. Insects such as termites have very little cognitive plasticity but this makes them expert coordinators (within restricted domains). For humans, potentially coordination-disrupting plasticity is reined in by our use of *shared* scaffolds. We are flexible and our contemporary PAs are ultimately contingent, but in a given time and place their shared nature and mutually applied pressure makes them powerful determinants of behaviour.

Scaffolds of this sort are not taken from the environment and replicated within an individual brain. They exist in the social environment, in shared practices, and in culturally evolved institutions (Sterelny 2012). Without immersion in deeply structured environments replete with culturally evolved, fine-tuned scaffolds of this nature human agents as we know them would not exist. If you remove the scaffolds, the forms of mindedness and complex cognition they support collapse. By being both outside individual brains, and yet *constitutive* components of the cognitive routines and models brains shape themselves to fit, the scaffolding model of cognition makes the contents of human minds fundamentally external. Taking this idea seriously is a core and crucial difference between the mindreading and the mindshaping perspectives.

Understood this way, we can see that the fundamentally internalist mindreading conception of coordination overly complicates the explanation of social cognition. Evolution, cultural or biological, has not installed mindreading modules in

our brains, though it often may feel like this to individual coordinators (or game theorists). Instead, our ability to coordinate has the form of a magic trick: we regulate one another to conform to established patterns, thus getting coordination without complex and costly cognition. We do not interpret agents, instead, we jointly create the agents who participate in coordination. Just as explanations of consciousness run into "the hard problem" if we use our phenomenological experience as a starting point (Dennett 1991a), so too does coordination throw up problems if we take the phenomenology of its execution as our starting point for explanation. Like consciousness, to understand coordination we must deflate the intuition that how it appears is the way it is. Though it may feel like it, social interaction does not rely on solipsistic agents applying mindreading skills to determine the beliefs and preferences of co-interactants.

Viewed from a different perspective, an externalist one, where minds are meshed in a web of interaction at once regulating themselves and others to conform to models that exist in the common domain, the problems faced by Schelling, Lewis and Guala's accounts of coordination are dissolved – we do not need to discover internal strategy sets in order to coordinate, instead the nature of our cognitive practices means that we come pre-wired for coordination thanks to interpersonal self-regulation. In other words, we do not *solve* coordination dilemmas, but rather coordination solutions *emerge* as a by-product of how our selves are constituted. As social creatures enmeshed in a web of mutually regulating rules for behaviour via which we interpret the actions of ourselves and others using models which we sanction deviance from and fear being sanctioned over ourselves, coordination comes cheap. In its most fully developed form this explanation of human

sociocognitive competence has come to be known as the mindshaping model (Zawidzki 2013; McGeer 2015).

Mindshaping is an example of what Clark (1997) calls "epistemic action", a process whereby agents act on the world to make complex tasks easier. By regulating conspecifics to be alike, mindshaping transforms the informational environment in which coordination takes place. Other agents become more amenable to efficient understanding. Action coordination is "made more tractable not by providing interpreters with a more powerful theory of mind but by making targets of interpretation easier to interpret using low-cost computations capable of tracking observable behavioral dispositions" (Zawidzki 2013: 69). Indeed, regulation is what allows for accurate behavioural prediction because agents whose minds are the product of pervasive shared mindshaping processes can be appraised as agents acting rationally within a given, constrained domain. It is this narrowing down of possibilities through cognitive homogenisation that allows for widespread, low-cognitive-cost, coordination.

As we saw above, the purported use of mindreading to deduce a course of action in a coordination dilemma is faced by serious obstacles. Strategy choice is significantly underdetermined by the evidence, as such choice requires the agent to know what the other person will choose in advance, but the other player also cannot choose until they know what the first player will choose. With mindshaping such a dilemma is avoided: instead of trying to predict the strategy choice of an interactant, in most quotidian settings the dilemma comes presolved, or at least the range of strategy choices is severely delimited, allowing for low-cost techniques like type-based differentiation of agents, or the tracking of observable behavioural patterns, to inform optimal strategy choice.

Mindshaping bears some relationship to Clark and Chalmers' extended mind hypothesis (1998). This is because, according to mindshaping, individual minds are extended into the world in the sense that their content and states emerge from the skilful use of entities beyond individual brains, in particular socially sourced scaffolds that constrain and guide behaviour. However, mindshaping goes beyond the extended mind hypothesis, which preserves the mind as a discrete self-standing entity that can sometimes be extended.[7] In mindshaping the mind itself emerges from and is continually reshaped by cultural processes, and this interaction is bidirectional. Epistemic action of this sort distributes the cognitive burden of achieving complex tasks onto the world and this division of epistemic burden requires division of epistemic credit. Humans are not master social cognizers as a result of in-built cognitive tools, instead they manage to effectively coordinate due to their immersion in richly scaffolded environments. Fundamentally, unlike the mindreading-based paradigm, mindshaping recognises that human social cognition does not result from our onboard theory application skills, but rather through the effective epistemic engineering of the social environment to transform the task, coupled with the reciprocal engineering of our own minds by the environment.

2.1.2 From Reading to Shaping

That a concept like mindshaping has taken so long to emerge as a plausible explanation of social cognition can be explained by the tight links between

[7] Indeed Chalmers, co-originator of the extended mind hypothesis, also coined the now infamous "hard problem of consciousness", thus reanimating an old Cartesian spectre. This fact perhaps explains the curious reification of the "mind" that the extended mind hypothesis itself smuggles into its otherwise radical proposals. Additionally, Chalmers's most recent work *Reality+* (2022) further develops his general solipsistic thrust, proposing that minds can, and indeed soon will be, uploaded to artificial substrates to participate in virtual worlds. On the mindshaping view such proposals radically misconstrue the nature of the "mind", and the manner in which it emerges from the social. There is no self-standing mind to upload without interpersonal co-creation and reinforcement.

mindreading and the dominant approach in philosophy of mind during the latter part of the 20th century, namely computational functionalism. Mindreading owes its origins as a concept to an intersection between work in psychology and philosophy in the late 70s and early 80s. Comparative psychologists Premack and Woodruff (1978) discovered that chimpanzees seemed to be able to infer the intentions underlying the actions of another agent, which they took as evidence for the existence of a theory of mind in these animals. On the prompting of philosophers (Bennett 1978; Dennett 1978), this experimental protocol was then further developed to become the "false belief test", presumed to demonstrate that agents, in particular non-linguistic agents, could have beliefs, and thus possess PAs. These results were then widely taken to show that some social agents possess a rich theory of mind (TOM), attributing beliefs, desires and intentions to other agents.

This interpretation of the empirical work, in turn, seemed to support the reigning paradigm in philosophy of mind at the time, namely variations of computational functionalism (Fodor 1975, 1980; Pylyshyn 1984; Chomsky 1959). This is the idea that the mind is a type of encapsulated computer that utilises a language of thought in some form with which it represents states in the world. Effectively, the computational functionalist paradigm understands social interaction as akin to computers engaged in the spectatorial probabilistic assessment of the possible states existing in the minds of other computers. If PAs are cashed out using a representationalist language of thought, then mindreading and the discovery of such discrete language-like entities in other minds becomes highly intuitive. Thus, the mindreading paradigm built on the applications of the false belief protocol and a functionalist conception of cognition to propose that folk psychology works through the spectatorial application of a TOM, allowing humans to infer the concrete internal

causal states that drive the behaviour of others. Because they denoted more or less the same thing, the terms folk psychology, theory of mind, mindreading and mentalising were often used interchangeably in this literature (Hutto 2009).

This early majority opinion set the tone for the ensuing debate over the coming years which orbited around the idea that folk psychology, equated with TOM, aims at the prediction and explanation of other agents. These debates took place primarily between the aforementioned theory-theory (TT) and simulation theory (ST) variants of mindreading and sought to demonstrate, using a combination of a priori argumentation and empirical work in psychology, how some variety of TOM enables humans to engage in fluid and largely accurate social cognition. As we saw above, the substantive differences between TT and ST relate to how exactly people attribute PAs to other minds. TT proposes that people literally use a theory of mind, one that may be innate, learned or subpersonal, to make predictions about other minds. Simulation theorists agree that we make such predictions, but argue that they emerge as a result of simulating in one's mind how others would behave given the behavioural and circumstantial data available.[8] Over time the two camps have begun to merge, with hybrid accounts that have a place for both simulation and theory in mindreading becoming influential (Nichols & Stich 2003). However, beyond the disagreement about method, both approaches accept that the key role of folk psychology or theory of mind is the *epistemic appraisal* of other agents' minds. The apparent dominance of this presupposition led José Luis Bermúdez to claim about mindreading, as late as 2003, that "No philosopher has, as far as I know, denied that this set of attributive, explanatory and predictive practices exists, nor that these practices implicate a network of psychological concepts." (25).

[8] Davies and Stone (1995) provide an extensive overview of this debate.

As accurate as Bermúdez's description of the field may have been at the time (though see Mameli (2001) and McGeer (2001)), this TOM-based prediction-explanation idyll wasn't to stand for much longer. Beginning in the early 2000s (indeed Bermúdez (2003) is itself an attack on TOM) the consensus around mindreading began to break down. Various theorists began to propose new, orthogonal, descriptions of the role and function of folk psychology (Leudar, Costall, & Francis 2004; Gallagher 2004; Hutto 2008). These approaches began from the claim that social interaction plays a crucial role in developing folk psychological expertise, and questioned the idea that anything theory-like is part of human social cognitive competence. Furthermore, the TT/ST TOM paradigm was characterised as an artefact of bad philosophy, one sustained by self-serving interpretations of the results of false belief experiments (Leudar & Costall 2009).

In place of the classic TOM-based interpretation of folk psychological competence, these theorists advance their own positive proposals. Broadly speaking these accounts place interaction at the centre of human cognitive competence, and thus have been grouped as *interactionist* theories (Fenici 2017). For instance, Gallagher (2008), building on the phenomenological tradition, proposes that PAs are the subject of direct perception as opposed to third-person theoretical inference. Hutto (2008) suggests that appreciation of the social context, or the narrative structure in which agent explanations are situated or embedded, allows for social understanding without mindreading. Yet another approach (De Jaegher 2009), inspired by enactivism, argues that social cognition emerges from dynamic processes of interaction between the body and the environment. A more recent proposal (Andrews 2015) draws together a range of approaches to outline a pluralist account of folk psychology according to which agents use various heuristics, including the

use of minimal mindreading (Butterfill & Apperly 2013), appreciation of normative contexts and salient stereotypes, and teleological reasoning, to predict and explain action.

Despite the (relative) heterogeneity of positive accounts on offer, those collected in the interactionist camp are united in a negative proposal, namely a rejection of the epistemic claims of mindreading, according to which "successfully navigating the ins and outs of our social world depends critically and primarily on the on-going exercise of a *purely epistemic capacity* for reading other minds." (McGeer 2020: 4, emphasis in original). The rejection of this presupposition is motivated by the fact that the use of folk psychology is faced by intractable obstacles if its function is an epistemic one, i.e. the prediction of behaviour. In general, the interactionist theorists aim to explain how *existing* psychologically rich agents carry out fluid, if not representationally heavy, social interpretation. Yet, crucially, these theories do not provide a full account of how biological individuals (our pre-hominin ancestors) *became* these agents with complex cognitive profiles[9]. In other words, on either an enactivist or phenomenological account, how did the agents capable of being appraised using the craft or skill of folk psychology come into being? As *interactionist* theories, they generally presuppose a degree of social constructivism in the generation of the subjects of folk psychology. This opens room for folk psychology to be understood as itself emerging from dynamic interpersonal processes of interaction.

Mindshaping theory can provide a plausible account of how such subjects came into being, or rather, evolved. It places at its core the idea that the *practice* of folk psychology itself partly *constitutes* the domain of folk psychology. The broad

[9] Andrews (2015) is an exception to this general criticism.

idea is that our folk psychological expertise, however it is cashed out, depends on interpersonal regulation of mind states, itself often carried out by the use of culturally specific folk psychology (Zawidzki 2013; McGeer 2020). By regulating agents to fit into culturally evolved scripts mindshaping processes allow agents to emerge with complex, socially regulated and thus shared and delimited cognitive profiles. In other words, mindshaping theory can explain the initial emergence of agents with rich psychological profiles through a process of gradual bootstrapping which ensures the ongoing effectiveness of folk psychological expertise.

The human developmental environment is saturated with social scaffolds that reliably bootstrap into being the agents to be assessed using the craft of folk psychology (McGeer 2020). Pervasive immersion in interactions that regulate cognitive profiles acts as the scaffolding basis that keeps the skill of folk psychological expertise effective. On this account, folk psychology functions *primarily* to regulate others; its use shapes interactants to be interpretable to one another. However, mindshaping *comes first*, allowing for various folk psychological heuristics and techniques to be effectively used in social interaction. This is because mindshaping processes must ensure that the domain folk psychology tracks is not independent of the culturally evolved capacity to track those states. "The continual feedback we thus provide to one another constitutes a kind of external scaffolding for maintaining recognizable ways of being minded that cannot be kicked away" (McGeer 2020: 16). In this sense mindshaping is the lynchpin of human socio-cognitive capacities (Zawidzki 2013).

Mindshaping allows that folk psychology sometimes enables prediction, but it does much more than this: it partly *creates* the subjects that it helps predict. Folk psychology, through mindshaping mechanisms, brings into being the agents that are

its subject matter. The reason that it evolved to do this is because without being shaped to fit shared, constrained cognitive profiles, humans wouldn't be able to engage in quotidian coordinated interaction at all because the range of strategy choices available to unconstrained, hypersocial language users is too large. Without some way to determine possible strategy choices in advance of, or during, interaction, the effective, coordinated, social life human groups rely on would be impossible. Thus, in line with interactionist theories more broadly, the mindshaping approach shifts our understanding of social cognition away from being a primarily epistemic faculty, i.e. a means to know the minds of others, to an understanding of it as serving a regulatory role.

Of course we often, especially in retrospect, explain behaviour using folk psychology cashed out in the language of beliefs and desires, but this is not the passive application of a theory of mind that correctly divines the internal drivers of behaviour (McGeer 2020). Instead, believing that someone did something because they desired a specific outcome and believed that the chosen action would bring it about, serves to both rationalise their actions, and act as a model for one's own behaviour in future. Acting *as if* individuals hold discrete PAs that cause their actions serves to make folk psychology effective in general, both because it determines how you treat the person, thereby reinforcing specific PA interpretations by the individual to themselves, but also because it acts as a guide for oneself about ways of behaving, a guide of accepted action in one's cultural milieu. Thus PAs are logically posterior, not prior, to effective mindshaping. In other words, we are shaped to behave in a manner that approximates that of someone driven by our culturally specific PAs and this makes their use effective, as opposed to having PAs

already inside our heads driving our behaviour that our theory of mind comes to discover.

Evidence that this is the case comes from cross-cultural comparisons showing significant differences in how individuals use and understand psychological states (Lavelle 2021). Cultural variance in folk psychology shows us that through time cultural groups develop bespoke social scripts for interpreting behaviour. A person's culture teaches them a specific folk psychology that they use to rationalise behaviour, both that of others and their own. Folk psychology shapes them to act in accordance with that culture – its use makes it true, not its fidelity to actual internal brain states. Thus, folk psychology plays an important role in mindshaping; however, it functions to enable interpersonal mind-making and to help explain away inconsistencies in behaviour, not for the individualistic discovery of pre-existing mental states.

As we have seen, the dominant TOM-driven paradigm has tended to run together the concepts of mindreading, TOM and folk psychology. Indeed, on this paradigm these three terms are the same thing – mindreading is the use of folk psychology which is the application of a theory of mind. However, because mindshaping, and interactionist approaches more generally, reject the traditional paradigm's TOM component, this equivalency dissolves. On the new schema, mindshaping comprises a set of processes or expertise that aim to make minds fit models and folk psychology is just one component of mindshaping. Furthermore, the content of human folk psychological expertise varies depending on what specific flavour of interactionist explanation one prefers[10]. In what follows I will focus on the

[10] Andrews (2015) provides a pluralist account that incorporates various cognitive mechanisms in the use of folk psychology. Her account is compatible with the broad contours of Zawidzki's IS conception of folk psychology and extends it in an equally evolutionarily well-informed manner.

work of Zawidzki (2008, 2013, 2019; Fenici & Zawidzki 2021), and specifically his Dennettian explanation, appealing to "the intentional stance" (IS), as the subpersonal basis of folk psychology. This is because, when correctly conceptualised, an IS-based explanation of the processes underlying social interaction is ultimately compatible with a wide range of enactivist-inspired interactionist approaches to explaining the content of our folk psychological expertise (Zawidzki 2012). As we will see, attributions of goals and contexts from the IS can perhaps be referred to as a kind of mindreading, but the "mind" being read bears little relationship to the mind of classic mindreading.

2.2 Making and Maintaining Minds

Zawidzki (2008, 2013; Fenici & Zawidzki 2020) proposes that the interpretation of agents from the IS can replace TOM-based mindreading and that this, coupled with pervasive mindshaping to limit the threat presented by the holism problem, ultimately explains human social cognition. To adopt the IS in behavioural prediction is to treat a system as practically rational and to assume that it will adopt the most efficient means to achieve its goal. We can attribute PAs from the IS to explain our prediction but "to attribute propositional attitudes is *not* to speculate about the concrete causes responsible for behaviour. Rather, it is to situate behaviour in a *rational*, *normative* framework, to see it as a reasonable response relative to goals and available information" (Zawidzki 2013: 14, emphasis in original).

Thus, the IS allows behaviour to be interpreted as goal-directed without speculating about the internal states of the agents or systems doing the acting. When we engage in behavioural prediction from the IS, we don't literally talk about what is inside the heads of our co-interactants. Instead, we predict their behaviours with

reference to our social milieu and their observable actions. Given mindshaping, if the agents come from sufficiently similar social worlds, these predictions will be (roughly) accurate. Individual agents thus emerge in the context of social structures within which their actions are interpreted by others as instrumentally rational. One upshot of this schema is that people do *not* have "intentional states, specifically preferences and beliefs, that remain assignable properties of them in isolation from the normative expectations of others with whom they interact. Individuals…are products of social structure, not components into which social structure can be analyzed" (Ross 2014: 268). The causal components driving behaviour do not reside solely within the brain-boxes of individual humans – the social structure within which they are embedded, including other agents, is a *necessary* element required to assign intentional states to individual organisms. That this stopped being obvious to philosophers was the result of the cognitivist and computationalist repudiation of Skinnerian behaviourism during the 60s and 70s: a misguided rejection of the necessary role played by the external environment in determining an agent's patterns of behaviour (Ross 2015).

Behavioural interpretations from the intentional stance can stand in for (at least quotidian) mindreading. On this account, we need not correctly attribute propositional attitudes that exist in other minds for prediction at all, and indeed it remains up for debate if such propositional attitudes exist in any form as causal drivers of behaviour[11]. Zawidzki's position is that it is an empirical question as to whether mindshaping coupled with IS interpretation obviates the need for mindreading entirely or merely that PA ascriptions are made functional by prior

[11] Chater (2018) proposes that all human interaction is "flat" in this sense: it does not rely on pre-existing PAs in the mind that drive behaviour instead we retrospectively invent suitable PAs on the fly to maintain coherence. This is eliminativism about the mind which remains a legitimate explanatory possibility in behavioural science (Ross 2022).

mindshaping (2008: 205). Either way the intentional stance should not be read as giving "an adequate analysis of our mature concepts of the mind and mental states" (Zawidzki 2018: 49). Instead, it provides us with a framework with which to understand behavioural prediction *without* invoking the concept of determinate PAs linked up to discrete brain patterns.

The existence of this type of low-level, quotidian interpretation and prediction of behaviour, without the attribution of full-blown PAs, is supported by empirical work in developmental psychology. Gergely and Csibra (2003) show that infants interpret behaviour as being goal-directed even when they are too young to be able to use the linguistic concepts that underpin full-blown PAs or mature concepts of mind. As Gergely points out:

> Young infants are ready to interpret unfamiliar entities such as inanimate objects, abstract 2D figures, humanoid robots, unfamiliar human actions, and even biomechanically impossible hand actions...as goal-directed as long as they show evidence of context-sensitive justifiable variation of action obeying the principle of efficiency of goal approach (2011: 87).

This is taken as evidence that human infants, and perhaps some other social animals, are born with an ability to attribute instrumental rationality, the type of interpretation of which the intentional stance is a variety.

Instrumental rationality of this sort would have been selected for because it allows for the efficient pursuit of goals in social settings and because the world is populated with intentional agents that it pays to track. That nature is populated with agents acting rationally is a given, in the sense of acting in ways to achieve goals relatively efficiently, because natural selection favours efficient replication. As our success in the world attests, people do aim at ends, and their brains help them do

this. The intentional stance allows us to understand humans as intentional systems without delving into the specific neuronal basis of how exactly we are so. This is because interpreting complex physical systems from the intentional stance allows us to see the patterns in their behaviour. It allows for explanatory purchase, but it does not reveal anything about the causal states within nervous systems.

As agents capable of engaging in complex language use, humans can be seen as second-order intentional systems (Dennett 2017). This is because they track intentional systems, including their own selves, and use language to describe and annotate behavioural patterns using the intentional idiom. By attributing intentional states to themselves as self-rationalisations for behaviours, they regulate themselves to more closely approximate intentional systems. This type of second-order intentional stance taking is crucial for maintaining intertemporal consistency in complex social domains, which is necessary for coordination.

The picture being developed here suggests that the majority of quotidian human coordination is achieved with the use of simple heuristics that construe systems as having goals, information and being instrumentally rationality, not on accurate ascriptions of discrete mental states driving behaviours. There is no inner computational effort aimed at modelling the minds of others to better predict their behaviour. Instead, we largely act in the social world as we have been conditioned to expect it to function, and, thanks to the processes of mindshaping, this conditioning makes it generally function in this way. The assemblage of IS-derived descriptions that regulate an individual's behaviour and inform their self-conception can be understood as forming a (relatively) consistent narrative that a self tells itself about itself. This publicly affirmed narrative then allows others to predict behaviours while being relatively assured about future behavioural consistency and thus potential

continued success in coordination and also allows an individual to bargain around coordination conventions with future versions of their own self (Ross 2008).

The use of narrative, a story of how one's self *should be*, allows humans to interpret their messy, complex and stochastic brain states as approximating various specific, socially derived patterns. The self, the protagonist of one's story, acts to shape one's mind to fit a model. This idea, that humans rely on self-regulating narratives to structure and interpret their own cognitive and affective states, has become increasingly influential in recent years (Dennett 1991a; Ross 2007, 2014). The linear, digital structure of self-directed speech provides a template through which humans can interpret and regulate their inner states to form a coherent story (Zawidzki 2013). Indeed, Dennett famously compares the construction of a human self to that of an author constructing a fictional character, like Conan Doyle's Sherlock Holmes (1991a). The use of language to tell a story about who you are and what you want allows the brain to structure perceptual input, prioritise sensory data and aim at consistent goals through time. On Dennett's account, a mind is nothing more than "a narrated compression of patterns in the dispositions, embodied in differential neural network weights, conditioned by an individual's history of encounters with the environment while she tries to keep herself relatively secure, calm and occupied" (Ross 2015: 31).

Thus the mind of contemporary humans is a virtual object, one that is fundamentally socially constructed and maintained. But like money, an archetypical virtual object, this does not mean that the mind is not real. The reality of the dollar is not in the bills, it is in the structure of use in which money is embedded. Using dollars as a store of value makes them real. Likewise, the reality of the mind is not in the brain, but rather in the role of minds and their selves as abstract centres of gravity

that regulate behaviour and allow for prediction. Like the dollar, treating one another and ourselves as if minds are real makes them real. Thus, IS interpretations of behaviours and dispositions are not instrumentalist and do not eliminate the mind. Instead, they track real patterns, ones that would be unavailable were one not to use the IS as a lens of analysis (Dennett 1991b).

The upshot of this characterisation of the mind is that its content is only determinate after probing by oneself or another with a request for an explicit account using language. In advance of action in the world it is not determined. It is for this reason that there is no reading of minds in advance, only interpersonal and self-mindshaping in the moment. This insight is central to the enactivist tradition that is broadly allied with the interactionist conception of our folk psychological expertise surveyed above. Taking such enactivist impulses to their logical conclusion[12], this understanding of the mind also informs Dennett's radical naturalisation of phenomenal consciousness (1991a) – the proposal that self-narration is all that underpins the experience of consciousness. The fictional characters that people co-create in social contexts are what they experience as inner selves. Intriguing as this approach is, accurately explaining the phenomenology of consciousness is not central to my project here. What is important is the idea that self-regulating narratives are *socially sourced*, and thus significant components of what we understand to be our essential nature are in fact socio-historically constructed scripts developed by culture and installed onto our neural processing machinery.

Once established, these culturally sourced self-regulating narratives are intrinsically motivated: for a human self-construction does not further a specific end.

[12] Though in so doing generating a proposal that wholesale rejects the value of phenomenology as an empirically valid source of data, a body of work that is central to the work of many enactivists (i.e. Gallagher 2008).

It is a good in itself and a prerequisite for participation in human society. But, when viewed from the macro perspective of cultural evolution we can see that this developmental process was selected for because it enables large-scale complex coordination. By defining one's self using resources developed socially, agents remain alike enough, in significant enough respects, to be able to pervasively and easily engage in coordinated activities. If this regulative function was not a necessary component of self-creation the design space opened up by complex language powered self-formation would make quotidian coordination extremely difficult. Coordination then is not just a happy by-product or epiphenomena of mindshaping via cultural self-formation, it is its primary function. As a result, the selves that humans make for themselves, inside cultural blocs at least, resemble each other in key ways. Coupled with social norms and micro-signalling mechanisms of behavioural conformity, language-based self- and group-constituting regulative frameworks form the most advanced supporting structures of human mindshaping.

2.2.1 Whither Mindreading?

Zawidzki minimises the role of mindreading in social cognition. According to him, mindshaping is the primary mechanism for enabling social interaction. However, this is not the only way to cash out mindshaping. Victoria McGeer (2001, 2007, 2015, 2020), another leading proponent of the approach, which she also calls the regulative role of folk psychology, explicitly rehabilitates mindreading as a functional component of our suite of cognitive tools. McGeer reminds us that when minds are effectively shaped to local circumstances this allows for relatively accurate, context-specific prediction. For example, rugby players understand the actions and goals of other rugby players in a unique manner; they have been shaped

to be able to "read" the minds of other players. Folk psychology, or as McGeer calls it, "mentalising", on this schema serves three interrelated ends: pedagogical, regulative and descriptive. The pedagogical and regulative aspects are captured by Zawidzki's conception of mindshaping – folk psychology both shows agents how they are to behave and regulates them to conform to this model. However, McGeer proposes that the descriptive function of folk psychology *is* a form of mindreading, by which she means that humans socialised into specific cultural milieus do successfully predict the behaviours of their interactants. Their culturally installed folk psychology allows them to make descriptive claims about the behaviours and motivations of their co-interactants.

As she puts it, immersion in the space of reason giving and asking trains humans to become "inveterate and prolific mentalisers" (McGeer 2020: 19). Rugby players and fans become skilled users of the intentional schema defined by the rules of rugby. Normal human agents likewise become skilled users of the schemas installed by their cultural groups. Rugby is intelligible to the initiated just because its set of rules are constrained. Likewise, human social life is intelligible to the initiated because its own set of rules are constrained. Without constraints there would be no way to coordinate agents, no (complex) human social life. Thus, via this account of prediction in constrained spaces McGeer's account explicitly pushes back on the idea that nothing like mindreading ever occurs, a stance she misattributes to Zawidzki, among others (2020: 18-19) and which has been the driver of sustained criticism of the mindshaping framework (Peters 2019). McGeer aims to accommodate mindshaping, a conception of folk psychology as primarily regulative, to the lived reality of human affairs, whereby humans do often attribute PAs to one another, and even to inanimate objects.

However, Zawidzki does not necessarily disagree with the descriptive use of folk psychology, but rather carefully avoids calling this skill mindreading. Instead, he describes it as the successful use of the intentional stance in a specific context. For Zawidzki, following Dennett, our folk psychological craft does allow for behavioural prediction, for descriptive use, but it does not track discrete causal states that exist in nervous systems. Rather, it tracks "abstract posits, akin to centres of gravity in physics, which help track robust patterns of observable behavior." (Zawidzki 2013: 14). Through self-directed application of the intentional stance human selves attempt to better match their behaviour to these abstract posits, thereby making themselves more predictable. However, calling the prediction of behaviour based on the use of such interpretation *mind*reading is potentially misleading. The "mind" being "read" is not an entity that exists separately from the web of action and prediction in which it is embedded. The descriptive mindreading McGeer proposes is more akin to a type of skilled situation reading, a situation populated by agents that can be successfully interpreted using context appropriate application of the intentional stance. Thus the apparent disagreement between McGeer and Zawidzki turns on a broader semantic point about the use of the term "mindreading".

Mindshaping was formulated to respond to a specific problematic conceptualisation of human social interaction in philosophy of mind. However, in the act of dethroning mindreading, and with it the concept of pre-formed minds to be read, mindshaping also licences a wholesale reconfiguration of our understanding of how human selves are created and maintained. Thus, the claim that minds are the product of mindshaping is more than a redescription of social cognition, how it emerged and is possible. It is also a radical theory of self-formation and the processes underlying the generation of groups, coalitions and human social

ontologies. To get a better grasp of how mindshaping plays this role the following section examines mindshaping beyond the use of folk psychology. We will see that it encompasses a range of processes that allow for the emergence of rich human selves that can engage in fluid, effortless coordination, despite inhabiting complex social worlds.

2.3 The Processes of Mindshaping

The regulative role of folk psychology is just one component in a suite of processes and behaviours that human agents engage in, learned or innate, that aim to ensure groups reliably conform their minds to approximate shared cognitive profiles and signal in-group coordination suitability. Indeed, the regulative attribution of PAs and the use of narratives of self-construction are processes that require a broader set of pre-existing mindshaping processes to themselves emerge. These processes may be innate or hard-wired, or the products of immersion in the human socio-cultural niche. However, regardless of their genetic or cultural origins they now reliably and cross-culturally emerge in human populations. Broadly speaking, mindshaping processes can be broken down into three distinct categories: *imitation*, *pedagogy* and *conformism*. These categories are unified by an overarching concept – that of making a target match their behaviour to a model. This goal, a mind matching a model, is central to Zawidzki's formal definition of mindshaping: "a relation among four relata: a model, a target, a mechanism, and a set of respects in which the target can match the model." (2013: 31). Mindshaping takes place when

> a mechanism aims to make a target match, in relevant respects, a model. The target is always a mind…The mechanism can be some pattern of activity in an individual brain…The model can be an individual agent, but it can also be

> something more abstract, like a possible pattern…[or a] purely fictional agent. The respects in which the mechanism aims for the target to match the model are properties of the model that the mechanism can represent or track. (Zawidzki 2013: 31-32).

For example, in imitation, mindshaping occurs when the target – the person doing the imitating – matches the model, *what* they are imitating, in the various salient respects tracked by the brain processes guiding the imitation. The model contains the unique information that shapes the target, the form that the target mind is encouraged to match. Mindshaping models can persist through time, and be codified in cultural objects, thereby allowing for displaced mindshaping, not merely in-person one-to-one, or one-to-many examples. Models can range from simple, like an imitated action, to complex, as in the case of aiming to approximate the behaviours of fictional agents, for example an idealised hero, like Queequeg or Cú Chulainn. The ease with which mindshaping models are transmitted and replicated between human agents, coupled with their scalable complexity is what enables them to pervasively and cross-temporally regulate human life: once an effective mindshaping model is established it can structure and regulate the behaviours and thoughts of potential targets indefinitely and be widely propagated to influence other targets.

The transfer and alignment of behavioural and dispositional models is the primary outcome of mindshaping. By pervasively and reliably causing humans to match their minds to specific, shared, models it underwrites the emergence of many other distinctive human behaviours and cultural achievements. By keeping behaviour relatively homogenous within cultural groups it facilitates prediction and coordination thereby providing a stable environment for advanced cognitive competencies to emerge. By highlighting the functional role of the model in

mindshaping we can see how information transfer, broadly construed, is central to its three main processes – imitation, pedagogy and conformism. In many cases these three categories are interwoven. For example, an apprentice will consciously imitate, receive pedagogical instruction and unconsciously conform their behaviours to the model provided by the master. Nevertheless, it is possible to roughly demarcate the separate processes operating in a bout of mindshaping. This is because we can isolate the different types of information transferred[13] and thus (at least partially) examine the effects generated by the specific models being propagated and see why specific types of information are adopted, and what incentive structures govern such adoption.

For example, imitation tends to transfer information about physical behaviours or appearance, and pedagogy usually aims to transmit explicitly conceptualised practices and skills. The information transferred by conformism is more complex to demarcate as the models being transferred by such processes are often implicit and various conformism behaviours operate subconsciously. Importantly, even in the case of fine-grained, skills-based imitation or pedagogy, the specific information contained in mindshaping models isn't only of functional use. Adopting the behaviours, skills and beliefs of conspecifics, or aiming to approximate a shared idealised model also acts as a *signal*. Thus, the specific content of mindshaping models is often secondary to the manner in which adopting shared models, whatever they may be, enables group coordination. Sharing mind models

[13] The dynamic and immersive nature of mindshaping-based negotiation of basic ontology means that the isolation of discrete acts of mindshaping is necessarily a theoretical enterprise. The point is to use the idea of discrete incidents of mindshaping as a heuristic device to investigate how mindshaping interacts with information propagation in general. Indeed, in the sense that a mindshaping model is a culturally-based way of doing something in the world it bears close affinities with Dennett's concept of a meme (2017), and like meme theory the specific demarcation of mindshaping models is a fraught process.

with others both makes coordination easier, and also acts as a hard-to-fake signal of in-group membership, and thus coordination suitability. The use of models as honest signals on the mindshaping-based conception of coordination directs attention to how information dissemination structures may interact with such signalling, and thereby coordination, a point I return to in Chapter 4.

2.3.1 Imitation and Pedagogy

Imitation and pedagogy are tightly bound processes of mindshaping and often occur in tandem. In many cases the aim of a pedagogical interaction is to teach another agent how to carry out fine-grained imitation, especially in the primarily skills-based lifeways of our early ancestors. Likewise, the imitation of a bout of behaviour can be pedagogical for the mindshaping target if it demonstrates new skills or culturally specific ways of acting, even if the aim of the imitation was not explicitly pedagogical or even consciously apprehended. The models and the respects in which the target is supposed to match the models in the cases of imitation and pedagogy are relatively easy to infer. This is because for purposeful imitation to occur the agent must consciously perceive the model and aim to match their behaviour to it. Likewise, a teacher must set out to explicitly instruct, and thus have some model in mind making the mindshaping at least partially intentional.

An example of explicit imitation is that of a child imitating a parent in the execution of a household chore, for example vacuum cleaning. The child explicitly tries to match their behaviours to the model, namely the parent's movements, in the salient respects. The mind of the child is being shaped in a novel way, in the sense that they are learning to engage in a complex set of behaviours, ones that would be incomprehensible to someone not familiar with the technology, or contemporary

human lifeways. Human imitation differs in important respects from the types of fine-grained imitation observed in other animals, for example, great apes (Whiten 2017), dolphins (Zamorano-Abramson et al. 2023) and some birds (Zentall 2004), as humans also engage in "overimitation" (Nielsen & Tomaselli 2010), the imitation of non-functional behaviours. For example, children often copy arbitrary and inefficient actions in a bout of behaviour, focusing on what the person being imitated is doing, rather than the specific outcomes they are aiming at.

A well-known set of experiments (Meltzoff 1995; Carpenter et al. 1998) demonstrate this effect. Specifically, these experiments show that infants imitate an adult switching on a light using their head, even though the task can be completed more efficiently using one's hands. Thus, the children faithfully copy apparently non-functional and inefficient behaviours. However, in a further set of experiments, (Gergely et al. 2002) it was shown that the infant's imitation is not mindless and that the non-functional behaviour was only imitated if it appeared intentional. For example, if the adult's hands are occupied, or if the adult implies that the action was not intentional by saying "Whoops!" the children turn on the light using their hands.

Thus, in imitation children discard inefficient, non-intentional behaviour as not relevant, but do copy inefficient, *intentional* behaviour. This suggests that, from an early age, humans are predisposed to selectively copy behavioural models in high fidelity even if the function of those behaviours is opaque, so long as the agent appears to be acting intentionally. Thus, for humans imitation serves not just to propagate functional behaviours but to also ensure that individual agents assimilate culturally specific ways of doing things, even when those behaviours are merely symbolic. Overimitation ensures that the information transferred remains robust, even if it appears non-functional. Thus agents remain alike and refrain from

individualist efficiency-directed modifications of behaviours that would dilute the signalling efficiency of culturally learned behaviour.

Effectively, intentionality acts as a cue to agents that specific components of the behaviour have value, even if they do not have obvious efficiency benefits. The value being indicated relates to signalling in-group status. The reason that specific, culturally evolved ways of doing things are useful is not only because they help achieve efficiency goals, and preserve techniques across generations, but also because fine-grained group consilient behaviour signals in-group status to interactants. To learn signals of this sort it is crucial to be able to tell the difference between arbitrary idiosyncrasies in action and important but causally obscure steps that are to be copied. The fine-grained non-functional imitation that emerges from close attention to the intentionality of behaviour makes the faking of in-group knowledge extremely difficult. Even if an agent seeking to deceive knows about the functional aims of an action that signals in-group status, unless they have learned it through closely observing other in-group agents, their behaviour will likely lack crucial inefficient, yet honest signals.

As with imitation, in the case of pedagogy, the models and thus behaviours it aims to transfer are usually explicit as it involves the act of instruction and indeed skills-based pedagogy can be seen as a type of guided imitation. In the contemporary world, the majority of explicit pedagogy takes place in school settings where the models are consciously designed and often state mandated. But pedagogical interaction has long been central to hominin life and humans have evolved to intuitively pay close attention to potential sources of information. For example, Csibra and Gergely (2006, 2009) demonstrate that infants are primed to pay attention

to caregivers and interpret specific actions, particularly if preceded by eye contact or an audible cue, as pedagogical.

This receptiveness to pedagogical interaction has likely been a feature of the hominin lineage at least since the transition from Oldowan to Acheulean stone tool manufacture about 1.7 million years ago, the artefacts of which are complex enough to require instruction to be robustly replicated (Morgan et al. 2015). Sterelny (2012) suggests that pedagogical acts of skill transfer, in the form of master-apprentice relationships, have been a central feature of human cultural evolution and the driving force behind our ecological dominance across diverse settings. Apprenticeship learning involves pedagogical instruction, explicit imitation, and also the implicit assimilation of a set of ontological claims about the structure of the environment, i.e. materials, animals and geography. The apprentice learns not just a set of skills but also implicitly an ontology in which the skills are situated. The mindshaping processes of imitation and pedagogy likely predate the emergence of language. Indeed, the pervasive coordination and informational transfer that such behavioural complexes enable are preconditions for language itself to emerge (Planer & Sterelny 2021).

The combination of infant attentional priming for pedagogical information and the propensity to engage in overimitation generates a powerful force for mindshaping in early development. Children assimilate the knowledge of their group through interaction with caregivers, and later through explicit apprenticeship-style pedagogy. These forms of mindshaping are aimed primarily at physical and explicit skills, though the fine-grained imitation they enable also acts as a signal of group membership and coordination suitability. In addition to nominally efficiency-directed information shared by mindshaping processes, there also exists a realm of

information that does not have transparent instrumental value, yet pervasively populates individual ontologies and is structured by cultural groups, namely symbolic beliefs. Information of this sort is often transferred by automatic conformist mechanisms. Thus, it is not intentionally transferred, but adopted as a side-effect of immersion in a specific cultural milieu. These processes play a key role in populating the manifest image of humans, automatically and pervasively transferring mindshaping models aimed at behavioural conformity, the adoption of which is not usually explicitly directed by either the mindshaper or the mindshapee.

2.3.2 Conformist Mindshaping

Human social life is primarily populated with explicit content: gestures, utterances, intentional pedagogical actions, narratives, etc. Together these form the wide range of ostensive communicative practices that serve to enable much human coordination. However there is also an implicit, yet no less important, form of information transfer that occurs in human social interaction. The value of this channel of information transfer is to bring about a basic level of automatic conformity of behaviour within human groups, conformity which acts as a powerful signal of in-group membership, further enabling fluid, low-cost coordination. We can split the processes that bring about conformity into three broad categories: low-level automatic matching behaviours, norm following and enforcement, and the regulative role of language (Zawidzki 2013: 50). In what follows we will look at each of these categories in turn, examining both how they operate to bring about intragroup conformity, and how they allow for unfamiliar humans to negotiate coordination strategies on the fly.

Low-level, unconscious matching behaviour is a well-documented phenomenon that emerges in the course of quotidian social interaction (Bargh et al. 1996; Chartrand & Bargh 1999; Heyes 2011). The basic idea is that humans frequently mimic one another when interacting in subtle, non-functional ways. For example, in a non-hostile interaction, individuals may unconsciously reflect characteristics back at one another like facial expressions, body language, or accents. Additionally, beyond the mimicry of perceptible cues like facial expressions or movements, which can sometimes be the result of conscious effort, evidence has begun to emerge that humans also synchronise their bodies on an autonomic level, a phenomenon known as physiological synchrony. For example, it has been shown that interactants sometimes match pupil diameter and heart rhythms (for a review see Palumbo et al. 2017). This type of autonomic nervous system synchronicity is especially prevalent at sporting events, where viewers conform their emotional reactions with the crowd thereby enhancing feelings of interconnectedness (Baranowski-Pinto et al. 2022).

Interestingly, given the context of the present project, it has been shown that this type of autonomic synchronisation can also occur without physical co-presence, and has been demonstrated to occur during competitive online gaming (Wikström et al. 2022). These types of unconscious matching behaviours are related to imitation, but unlike imitation they emerge spontaneously without effort on the part of the interactants. Their existence provides evidence for evolved, hardwired cognitive faculties that cause humans to unconsciously conform to specific subtle models of behaviour and even cognitive activation profiles of conspecifics. Such mechanisms also demonstrate how fine-grained mindshaping processes are, automatically modifying highly granular, non-functional behaviours to work as a type of signal.

Furthermore, it has been shown that this type of low-level conformity seems to enhance cooperative and coordinative dispositions in groups. For example, experimental work has demonstrated that when groups of interactants are instructed to march together or sing in unison they cooperate more on follow up tasks than control groups who though interacting did not partake in shared activities that involved matching behaviour (Kirschner & Tomasello 2010; Wiltermuth & Heath 2009). This suggests that imitation not only explicitly aligns people's actions but also implicitly operates as a signal that demonstrates openness to coordination.

Norm following and enforcement is the second category of conformist behaviours that acts to shape individual cognitive profiles. The pervasiveness and power of norms to regulate behaviour, present in all recorded human groups, provides clear evidence of how social structures determine individual behaviour. Norms are behavioural models that are primarily disseminated without explicit teaching[14] and undergird human life, providing scripts for action, and in the case of moral commitments, rules of behaviour that are binding. The existence of these normative structures greatly facilitates coordination and cooperation in groups (Boyd & Richerson 1985). If the majority of interactants can draw on preestablished conventions for behaviour they are not required to engage in complex cognition of strategy choices during interaction. Adhering to the norms that govern a cultural group, which can often be relatively costly, such as food-based prohibitions, or restrictive dress and gender codes also acts as a powerful signal to other group members that one is a bone fide member of the group, further aiding coordination. To this end group norm packages often also track and enforce visible markers, such

[14] Of course many normative systems are codified and can be taught, but this is a recent phenomenon. For the majority of the existence of the human species normative codes were implicit.

as hairstyles or styles of dress. Thus, norms enable fluid social interaction by proscribing and prescribing specific behaviours and thereby simplifying strategy choice in interaction, both by acting as signals of group membership and by ensuring broad behavioural conformity within groups.

The crucial importance of normative cognition for human social life is demonstrated by its early emergence in ontogeny. Cross cultural experimental work suggests that the faculties for making simple normative inferences related to fairness and hierarchy are innate. For example, young children have been shown to punish counter-normative behaviour after being taught a simple game with rules (Rakoczy et al. 2008). Indeed, children have been described as practising "promiscuous normativity", whereby they tend to construe behaviour they become familiar with, such as rules during play, in normative terms (Schmidt et al. 2016). This suggests that humans automatically and pervasively imbue the natural environment with normative content, applying rules for behaviour inferred from what they perceive as common. Thus, children do not just overimitate behaviour, as we saw above, they also imbue the behaviours they copy, functional or not, with normative force, further reinforcing widespread replication and their mindshaping and signalling power.

Further evidence that norms play a key role in shaping the minds of humans enmeshed in group settings comes from experiments demonstrating that exposure to specific, though wholly impromptu, normative social pressure can even alter perception. For example, in a well-known experiment (Asch 1955) a group of subjects and confederates were asked to choose correct answers as a group, for example to choose the longest line out of several, sometimes choosing with confederates instructed to make an incorrect choice. In the confederate treatment, the subjects often followed the confederate's choices, apparently misjudging the lengths

of the lines. In a follow-up experiment, it was demonstrated that this effect was not just the result of a judgement bias. Some participants misperceived the length of the lines under social pressure (Germar & Mojzisch 2019). The internalisation of socially produced norms is thus thought to lead to "a lasting perceptual bias towards norm-congruent sensory information" (Germar & Mojzisch 2019: 12). What these experiments show is that norms are not just rigid, preestablished rules for behaviour, but can also emerge dynamically to reinforce group judgements in the moment in order to homogenise interactants. In other words, when norms of behaviour are established, even on the fly, they can make their content "true" for those governed by the norm.

Lab experiments investigating equilibrium selection in resource distribution coordination dilemmas have demonstrated similar outcomes. In these games, the participants judged the specific resource distribution solutions that emerged towards the end of the games (each of which was an NE) as being fair, despite there being variation in solutions chosen across different groups (Binmore 2005). This shows us that fairness norms, though reflecting NE solutions, are socio-historically contingent, the specific distributions emerging as the product of accident. However once conventions are established, even artificially in the lab, they become imbued with normative force. Supporting this conclusion, Guala and Mittone (2010) show how after only nine rounds, players in groups who had been incentivised to coordinate continued to cooperate in a final round where they were incentivised to defect. This demonstrates how a shared coordinated experience can lead to the generation of norms, even in short term interactions. An influential series of experiments in the field extends these conclusions (Henrich et al. 2005, 2006). Experimenters conducted resource distribution games across a range of different cultural groups

showing that though individuals from varied cultural backgrounds are willing to pay costs to punish norm violators, the specific punishment thresholds differ widely across cultures. In other words, norms about resource distribution are present cross culturally, but the exact content of those norms diverges in a manner that is not determined by material outcomes alone. The divergence in fairness norms across cultures, coupled with the ability to manipulate their emergence in the lab shows us that *enabling coordination* takes priority, *not* the specific *content* of individual norms.

Thus rather than being the result of pre-existing preferences, norms often emerge from chance discoveries of coordination-resolving behavioural patterns. Once established such solutions are enforced and this makes them preferred and shapes the members of the group to see them as natural in origin. This conclusion is troubling because suboptimal equilibrium solutions that emerge may become enforced norms and be perpetuated by agents who see them as inescapable. This is the same phenomenon that, as we saw above, appears to underpin the emergence and endurance of gender and racial hierarchies (O'Connor 2019). This suggests that close attention should be paid to the incentives that govern the development of norms, for example signalling group membership or status, and to how those incentives are related to the structures of communication, lest problematic norms emerge and become entrenched, or naturalised within groups. Furthermore, we can view norms on a continuum with respect to degrees of moralisation. Due to the contingent emergence of norms more generally the content of morality is also (largely) socio-historically contingent. Indeed, morals are just binding or non-negotiable norms, and emerge via the same means, i.e. through socialisation into a specific group that has alighted upon some coordination enabling strategy.

Supporting this fluid conception of morality is the research referenced above, which shows that a greater proportion of norms becomes moralised when groups perceive themselves as facing ecological or existential threats (Gelfand & Lun 2013). Gelfand and Lun suggest that this tightening serves to strengthen in-group coordination. By punishing deviance from cultural scripts more harshly the group enforces a stricter set of coordination solutions. Thus, in times of crisis many norms are hardened to become moral injunctions and this serves coordination. Problematically, the more a group moralises idiosyncratic norms in the face of a perceived threat, the more often out-group members will appear to act "immorally", thereby further reinforcing a group's perception of being under threat. Ratcheting, or self-fulfilling, processes in norm tightening like these will be examined in greater detail in Chapter 4, as the generation of perceived threats to social groups has become a serious problem for contemporary society and one that has been harnessed by media content producers and platforms to generate revenue.

The regulatory role that the use of language plays in human mindshaping is the final and most powerful of the conformist processes that unconsciously operate to shape human minds to adopt shared models. This includes the mindshaping effects of PA ascription discussed above. PA ascriptions are naturally tied up with language. The belief/desire psychology that drives mindreading is cashed out in linguistic terms. Likewise, according to the mindshaping hypothesis, humans regulate themselves and others to fit PA ascriptions made about them. Naturally, without language, there would be no shared explanatorily rich PA ascriptions to conform to. Thus language is what allows for the transmission and internalisation of rich mindshaping models, ones that may also be displaced or fictional. This then allows

for complex self-mindshaping that reinforces the regulative models and keeps behaviour consistent through time.

The mindshaping processes described here – imitation, pedagogy and conformism – enable human group activity by keeping human game spaces constrained enough to enable fluid coordination. These processes are intrinsically motivated, not merely a way to propagate directly fitness enhancing behaviours. Instead of enhancing direct fitness, mindshaping processes, like imitation, aim at the fine-grained replication of constrained behavioural patterns at various levels of cognitive sophistication across cultural groups, the adoption of which both enables coordination and acts to signal in-group status. By socialising infants to become competent agents through the regulative ascription of culturally sourced narratives, human agents are made alike in important, coordinated ways. The primary outcome of these practices is the construction of a social niche in which complex group cooperative and coordinated projects can be carried out with minimal onboard processing.

However, though essential for generating mass coordination power, and thus instrumental in ensuring the material stability of quantitively more people than at any other time in the history of our species, mindshaping processes can also contribute to negative outcomes. This is because coordination is the primary goal of mindshaping, above epistemic validity, efficient action, or broader social welfare. Repressive or discriminatory ideologies can become entrenched or naturalised within coordinated groups. In-group and out-group discrimination often relies on what are understood by the perpetrators and victims as natural facts. “Facts” of this type are propagated by the conformist processes discussed above.

Ultimately, people face powerful incentives to adopt ontologies that maximise their personal coordination payoffs. In specific types of communication environments perverse incentives may emerge that push agents to strategically adopt ontologies that advocate repression or out-group animosity to act as a signal of coordination suitability with what are perceived as close conspecifics. Processes like these can lead to inter-group strife, as occurred in the case of the Yugoslavian and Rwandan atrocities, where a key common knowledge generating structure in the form of the state collapsed, pushing agents to adopt and reinforce pre-existing ethnic markers as a means of enhancing coordination, at the expense of out-groups. Likewise, if the communication environment is noisy, or incentivises the use of certain types of extreme or partisan signals, this can interact with mindshaping processes and potentially lead to the emergence of groups that, though coordinating well internally, adopt beliefs that serve to undermine large-scale social welfare.

Fundamentally, the generation of mass-scale coordination power will be required to confront the many challenges facing humanity, not least the climate crisis. The specific form a society that could generate this power will take, or whether it can emerge at all, relates to the individual-level signalling environment agents are embedded within. Whether this coordination efficiency is generated in a relatively peaceful and equitable social world or one that emphasises and rewards the adoption of a repressive ontology, or fails to emerge at all, with society splintering into cultural silos locked in deadly competition, will be determined in part by the structure of the communication technologies agents utilise and the types of mindshaping they allow for, and incentivise.

Despite these potential implications, little effort has been made by mindshaping theorists to address the normative and political concerns its

reformulation of social cognition generates. In an effort to begin to address this theoretical lacuna, in Chapter 4 I advance the claim that mindshaping theory can be used to reveal how the specific design of contemporary digital mass communication technologies is interacting with coordination to undermine the generation of coordination power and misdirect it towards pernicious and destructive ends.

2.4 Conclusion

This chapter has aimed to show how the theory of mindshaping sidesteps fatal issues faced by the mindreading-based explanation of social cognition, in particular the issues of holism and computational intractability while lending support to recent pluralist accounts of folk psychological expertise. But mindshaping is not merely a replacement concept that neatly slots into the old picture. Rather, it functions as a new schema, one that requires sweeping changes in our explanation of not just social cognition, but the constitution of selves and groups more broadly. The viability of a mindshaping-based explanation of social cognition shows us that the dominant internalist approach implied by the mind*reading* hypothesis can be discarded, and with it the Cartesian picture of discrete encapsulated minds as the sole causal drivers of behaviour.

However, despite the viability of the mindshaping-based explanation, full-blown mindreading, be it in a TT, ST or hybrid form, is still a respectable theory in philosophy of mind. Such a state of affairs would merely be unfortunate if the use of mindreading as a concept was confined to debates within philosophy. However, the continued acceptance and use of the concept of mindreading within philosophy provides it with an aura of validity, and as a result, it continues to be used to facilitate "intellectual trade across a range of disciplines when discussing social

cognition" (Hutto 2009: 224). As a result, the mindreading metaphor serves to smuggle bad philosophy into other fields, making a theory that faces intractable obstacles appear respectable, and arguably supporting an understanding of humans that underwrites a broader atomistic political ideology common in public discourse, a point we will return to later.

The mindshaping metaphor is an important corrective to our conception of the mind: it explicitly presents the mind as something that is *made*, dynamically shaped into being, not something that exists independently to be *read*. To coordinate with one another we must predict behaviour, but such prediction does not rely on the cognition-heavy *discovery* of discrete PAs existing in other minds. Instead, we get by through assuming that bouts of behaviour are goal-directed and constrained by the available information in a context where agents are shaped by social structures and self-attributions made using the IS, which operates to make such interpretation efficacious. Minds are not pre-social entities that contain isolatable causal drivers of behaviour, instead, they are co-created by social interaction when agents ask for and give reasons for behaviours.

From the perspective of philosophy of mind, we have seen how mindshaping is a plausible explanatory paradigm for explaining social cognition using a more interactionist and less spectatorial perspective. In the next chapter, I aim to further strengthen and expand this account using resources from a different discipline: the science of cultural evolution. We will see that, though not conceptualised as such, the processes and mechanisms that drive mindshaping are already accepted as core features of our best accounts of the evolution of complex cognition. As a result, mindshaping naturally fits into this promising and empirically supported work. And furthermore, cultural evolution can provide us with valuable resources to further

explain mindshaping. In particular cultural evolution takes seriously the issue of model choice in the spread of information, a core part of mindshaping underexplored in the existing literature.

Chapter 3 – Cognitive Evolution and Mindshaping

The following chapter sets out to examine how and where mindshaping fits into the evolution of human cognitive competence with the aim of strengthening its core claims and shedding light on some underappreciated features. Like all biological systems, human brains are the product of a long and complex evolutionary history, one influenced by environmental selection pressures. In theory, it should be possible to link that selection environment with features of contemporary human cognition, to reverse engineer its processes (Dennett 2017), and in so doing support specific hypotheses about how cognition operates. An evolutionary story for mindshaping is required because a key claim of the account is that it comes *first*, that is, *before* the emergence of anything like mindreading. If mindshaping comes first, when and how did it emerge? And is it plausible that it emerged first, before mindreading? In what follows I will show how mindshaping makes a better fit with the leading theories of the evolution of cognition than mindreading. And though these theories frequently refer to mindreading, on closer inspection they do not rely on a notion of this concept that clearly separates it from mindshaping.

The most widely accepted framework explaining the evolution of complex cognition is known as cultural evolution. This explanatory framework strongly emphasises the role that culture plays in enabling human ecological success and in the generation of modern human cognitive competency. As we will see in this chapter, despite the many references to mindreading this literature makes, mindshaping can be neatly integrated into its broad hypotheses. Furthermore, I aim to demonstrate that theories of cultural evolution, which emphasise the processes of strategic model choice in social learning, can feedback into mindshaping, helping to clarify the strategic nature of mindshaping's core processes. Thus, the integration of

mindshaping into the cultural evolutionary paradigm is a two-way street: it allows theorists of cultural evolution to provide a more plausible explanation of core processes that underpin their theories, while shedding light on the strategic nature of mindshaping. This expansion of mindshaping to incorporate strategic concerns will in turn allow us to better grasp how a mindshaping-based explanation of social cognition can be used to illuminate contemporary online communication.

This chapter will proceed as follows: Section 3.1 examines the most explicitly mindreading-friendly account of cognitive evolution, that of evolutionary psychology. If accurate this account seems to provide crucial theoretical support for the mindreading explanation of social cognition. However, this approach faces serious criticisms, which underminine its claims. A more recent and increasingly influential approach descended from evolutionary psychology avoids these criticisms, and implicitly relaxes reliance on strong mindreading. This account, due to the "Paris School", relies on a theory of ostensive communication that can be reinterpreted as a form of minimal mindreading, in the process making room for mindshaping. Incorporating mindshaping into this framework for explaining cognitive evolution is required because, as we will see, it is likely that its account contains valuable insights. Section 3.2 then examines cultural evolution, the other main approach to the study of cognitive evolution. Here I will propose that, though these accounts often refer to mindreading, the cognitive processes they rely upon are better understood as a very minimal form of mindreading which does not conflict with theories that emphasise mindshaping. As we will see mindshaping is sufficient to carry the load these theorists place on mindreading and the central processes of mindshaping – imitation, pedagogy and conformism – are already components of cultural evolutionary theory. Then, in Section 3.3 I then draw attention to resources

developed in both theories of cultural evolution that will be useful for my later examination of the interactions between digital mass communication and mindshaping processes. In particular, both schools of cultural evolution emphasise model choice in social learning, examining how the domain of interaction may modify the strategies deployed. Focussing on the structure of interaction domains can enrich our understanding of how mindshaping acts to determine our manifest image. Section 3.4 then concludes this chapter drawing out the epistemic implications of taking cultural evolution seriously.

3.1 Mindreading and Evolutionary Psychology

Social cognition lies at the core of all contemporary accounts of human evolution with the substantive disagreements between the various theories primarily relating to how much the demands of social learning have altered the genetic substrate.[15] One key empirical result that underpins the claim that humans are uniquely adept at social learning comes from cross-species intelligence tests with human infants, chimps and orangutans. Researchers gave the three species a battery of 38 cognitive tests designed to assess abilities related to space, quantities, causality and social learning (Herrmann et al. 2007, 2010). Children only clearly outperformed the other species in tests of social learning. This provides strong evidence that humans possess either hardwired or reliably culturally installed social intelligence skills.[16]

[15] Unsurprisingly, Darwin was the first to propose an evolutionary explanation of specific features of human cognition in *The Descent of Man*, yet, despite this long pedigree it remains a controversial area of study with some disagreement as to whether mammals share higher cognitive capacities as a result of sharing a common ancestor or convergent evolution (Bolhuis & Wynne 2009).

[16] However recent research comparing intelligence across a wider range of species including Corvids may undermine even this supposed unique human ability (Pika et al. 2020).

In general, the importance of mindreading to these theories varies depending on the degree of regulative feedback the cultural environment is thought to have on individual cognitive faculties. The greater the role that innate faculties which generate a universal "human nature" play in these explanations, i.e. the less of a role given to the environment in altering or enhancing cognitive capacities, the greater emphasis they place on strong mindreading. Viewing minds as encapsulated computational devices generates the problem of knowing other minds, and requires something like innate mindreading faculties or modules to overcome the Cartesian impasse it generates.

One early approach to explaining the evolution of cognition places this problem of knowing other minds at its core (Humphrey 1976). Since dubbed the Machiavellian intelligence hypothesis (MIH), this theory proposes that human evolution has been especially responsive to selection pressures arising from social deception. Deception of this sort was supposedly widespread due to our ancestors' immersion in worlds in which each individual aimed to gain maximum benefit for minimum energy expenditure. In this schema free riders are a constant threat – those who prospered did so through their ability to accurately read social situations and determine the motives of their interactants. Those who were better at this task were, thanks to their avoidance of free riders, fitter. This set up a feedback loop which pushed hominins up a ladder of increasing socio-cognitive competence for detecting and policing free riders, with increasing *general* cognitive complexity emerging as a by-product of this process.

The valuable outcome of these Machiavellian machinations is the enabling of successful group cooperation among self-interested agents disposed to free-riding. However, the MIH relies on fundamentally internalist presuppositions: the faculties

with which humans achieve success in the world reside within their heads and each individual is the bearer of discrete mind states to be known. As a result, the Machiavellian account has obvious affinities with the mindreading-first conception of human social cognition. After all, mindreading is an excellent skill to have in a game of motive detection. Thus, we can see how, from the early days of empirically informed evolutionary explanations of human cognitive success, mindreading was implicated, indeed in the case of the MIH mindreading forms a central plank. However, despite their early prominence, and relative accord with folk understandings of interpersonal interaction, attempts like this, which seek to link human sociocognitive competence to selection for mind-state detection, have fallen out of favour. Not least, as we will see below, due to an implicit commitment to an implausible cognitive architectural model, based on the hypothesis of massive modularity.

A related, highly influential approach that also takes seriously selection pressures resulting from the social nature of human life is evolutionary psychology, also known as the Santa Barbara-School (Lewens 2015: 147), initiated by Barkow, Cosmides and Tooby (1992). Like the MIH this approach claims that the selection pressures generated in our environment of evolutionary adaptiveness (EEA) led to the pervasive genetic adaptation of human cognition to aid survival and reproduction. The outcome of these pressures is a mind populated with an array of special purpose modules, custom designed to solve problems faced by our ancestors.

A key claim of evolutionary psychology is that, though these modules were adaptive in the EEA, subsequent alteration of our ecological circumstances has rendered some of them maladaptive. For example, the contemporary obesity epidemic is explained as the result of a malfunctioning cognitive module that drives

humans to overconsume high-calorie foods where available as they were both scarce and highly valuable in the EEA. Maladaptation happens because the set of cognitive modules is fixed, and changes in the environment have been too fast to generate new, better-adapted cognitive modules.

Evolutionary psychology thus understands human nature as the product of innate hardwired modules. This approach thus gives succour to proponents of a universal human nature (Pinker 2002), and gives relatively less weight to the idea of cultural processes contributing to, or modifying core cognitive faculties. Additionally, many proponents of evolutionary psychology postulate a dedicated theory of mind module (e.g. Tooby & Cosmides, 1995) that allows for fluid mindreading, though this commitment is not entailed by the theory per se.

Overall, then, evolutionary psychology, at least in its classic formulation, supports a mindreading-first conception of human sociocognitive competency, and in de-emphasising the influence of culture on cognition, rejects the idea that minds are shaped at all. Thus, if borne out, evolutionary psychology undermines the mindshaping hypothesis underpinning this thesis. However, this threat is less pressing than it appears, as the evolutionary psychological research program has been the subject of repeated, trenchant critiques (Buller 2005; Dupre 2012; Sterelny 2003, 2012).

One core issue raised relates to evolutionary psychologies' commitment to a modular conception of cognition (Fodor 1983). There are serious issues with this idea, known as the massive modularity hypothesis, which suggest it is a bad empirical bet about the structure of the mind. For example, it has been empirically demonstrated that high-level cognition displays individual-level differences and that these differences are strongly correlated across domains within individual minds,

suggesting that domain-general lower-level processes may underpin their emergence, not individually specialised and genetically endowed modules (Rabaglia et al. 2011). Additionally, the pace of genetic evolutionary change in humans is much faster than previously thought, pushing against the idea that the mind has innate unchanging modules that underpin its processes (Byars et al. 2010). Furthermore, it is increasingly accepted that the skills that underpin human ecological success are likely to be domain-general and continue to flexibly adapt to new circumstances, as opposed to being the product of specialised and now maladapted modules (Buller 2005; Heyes 2012).

These various critiques serve to demonstrate that evolutionary psychology, and particularly the commitment to a domain specific theory of mind module, is at the very least problematic. These same critiques also apply to the MIH, as it is also committed to a dedicated, domain-specific social detection module in the mind. In sum, though two of the historically most prominent accounts of cognitive evolution explicitly support a mindreading-centred account of human sociocognitive competence, while implicitly rejecting mindshaping, both face serious challenges.

3.1.1 Paris: Ostensive Communication and Minimal Mindreading

Recently, an account has emerged that broadly endorses the strong emphasis on genetic adaptions that characterises evolutionary psychology but is significantly more amenable to the mindshaping hypothesis. This approach is known as cultural attraction theory, or the "Paris School", due to its institutional basis in France (Sterelny 2017b; Heintz 2018). Broadly, theorists in the Paris School accept that "mental skills, even the most automated and domain-specific, may spring from domain-general learning processes. The fact that specialised cognitive processes

with dedicated neural resources could arise through learning…is now taken quite seriously." (Morin 2019: 531-532). In endorsing domain-general learning processes, the Paris School can avoid the critiques facing massive modularity, and allow for theory of mind to emerge not due to an innate module, but rather as part of a domain-general learning process.

Following the example of Sterelny (2017b), I will use Olivier Morin's *How Traditions Live and Die* (2016a) to represent the views of the Paris School (see Heintz (2018) for a detailed overview of the various views that comprise the Paris School). Expanding on Sperber's cultural attractors hypothesis (1996), Morin formulates a unique approach to the emergence of human culture and contemporary human lifeways. He proposes a "strange vision: there could have been human populations, societies just like we know them, with humans communicating and cooperating like we do, but whose cultural repertoire would resemble those of modern chimpanzees." (2016a: 11). Thus, according to the Paris school, everything cognitive that humans require for rich contemporary culture was present in our distant ancestors. This explicitly rejects the idea that culture may feed back and modify human genes (at least in the case of cognition). How then have humans travelled so far (at least culturally) from our supposedly chimp-like forefathers? Morin pushes his chips on the bet that complex cultures (or "traditions" as he refers to persistent cultural patterns) owe their origins to a genetically evolved ability to engage in ostensive communication.

Following a broadly Gricean account of language, the Parisian understanding of ostensive communication is communication that is both voluntary and overt. The idea is that this special type of communication allowed our ancestors to communicate specific information to one another about features of the environment,

allowing for attention to be directed economically. As Morin puts it, “To engage in voluntary transmission is to act deliberately to permit, improve, or canalise someone else’s learning of a piece of information, a knowhow, or a motivation.” (2016a: 65). This allowed for the selective reconstruction of attractive practices or know-how which over time ossify to become traditions. This approach allows the Paris School to reject high-fidelity imitation as central to cultural replication, and propose instead that the use of ostensive communication allows human agents to selectively communicate information to one another, information that receivers then reconstruct. However, ostensive communication seems to rely on advanced theory of mind skills, as Morin explains: “Communication in our species does depend on a set of very specific adaptations, [including] our capacity for mind-reading” (2016a: 241). Thus, it seems that a dedicated theory of mind module was present in our pre-cultural ancestors, and as a result the Paris School seems to be offering a theory of cognitive evolution that places a capacity for mindreading at the root of human cognition.

Though coherent and developed in great detail by Morin (2016a) this view has been criticised on a couple of fronts. Sterelny (2017b) suggests that it significantly over-intellectualises early human communication patterns and Richerson (2017) accuses Morin of deploying straw person arguments against gene-culture coevolution theories in cultural evolution. Certainly, the idea that before culture emerged humans were capable of modern communicative practices seems far-fetched. Chimp-like lifeways don’t seem to require complex language, thereby making the selection pressures required for its emergence somewhat mysterious.

However, it may be possible to strengthen the Parisian account using resources more amenable to mindshaping. There are grounds to suggest that the ostensive communication at the core of their account does not necessarily require

rich linguistic capacities. Instead, we can recast ostensive communication as a variety of the intentional stance, described above as a core component of mindshaping. Moore (2016) provides an account of Gricean ostensive communication that only requires *minimal* mindreading to get off the ground, *not* the attribution of rich PAs. On that interpretation all one needs is "a basic understanding of others' purposive activities and desires, operating in conjunction with some tracking what others had or had not seen" (Moore 2016: 19). Thus, perhaps Gricean-type communication can emerge if some capacity for inferring intentionality is in place.

As a result, the kind of ostensive communication at the center of the Parisian account does not necessarily need to rely on the attribution of internally represented PAs and the genetic development of a mindreading capacity that this would require. Furthermore, perhaps dropping this requirement could help its proponents face down the charge of over-intellectualising the communication our ancestors engaged in. In fact, Morin's own conception of mindreading is already quite thin and he accepts the socially structured nature of the capacity. For example, he endorses Franks' (2011) externalist friendly corrections to the received view, agreeing that mindreading is aided by social regularities and scripts, that information about the mental lives of others is apprehended with little effort and that mindreading in general is more perception than reasoning. Morin also grants that our ability to engage in mindreading at all is scaffolded by our collaborative propensities (2016a: 67).[17]

Thus, the type of ostensive communication necessary for the Parisian account need not rely on an internalist capacity for mindreading in order to operate and can

[17] Further evidence for a minimalist interpretation comes from a comment made by Morin, wherein he claims that Zawidzki's monograph (2013) has convinced him "that the scope and power of "weak" mindreading, without recursion or the attribution of fully fleshed beliefs and desires, is underestimated, and of great promise" (see comment on Zawidzki (2015)).

instead be adequately explained using the resources of minimal, and PA representationally-light, mindreading. Furthermore, this can help such accounts avoid the charge of over-intellectualising interpersonal cognitive practices. That Morin and the other proponents of the Parisian, cultural attractors framework, continue to use the term "mindreading" to refer to what turns out to be closer to minimal mindreading, is, I believe, due to terminological inertia, likely originating in the early and more explicitly internalist work of Sperber (1996). For the reasons cited at the close of the previous chapter, relating to the metaphorical role played by terminology, the use of the mind*reading* concept is problematic. In the case of Morin's theory it risks implying that our early ancestors may have had a rich theory of mind before they had even language. Prima facie intentional stance attributions of the sort core to mindshaping make a more defensible fit with his approach, and at the very least a strong conception of mindreading is not a feature of Morin's work.

The early proponents of evolutionary psychology were strong nativists about human cognitive faculties. This required them to posit a dedicated theory of mind module that enables rich mindreading to reliably develop in humans. However, as we have seen, later work in the tradition of evolutionary psychology has dropped this nativist requirement, accepting that a cognitive faculty like mindreading may potentially emerge as a result of domain-general learning. This revision means that approaches of this sort are less antithetical to the mindshaping hypothesis.

It is important to show that mindshaping is at least minimally compatible with approaches in the Paris School for two reasons. First, its approach and hypotheses are not demonstrably false and likely supply some correct and important insights (Sterelny 2017b). Showing the compatibility of mindshaping with the main schools of evolutionary theory is a way to hedge our bets concerning mindshaping's

ultimate validity. Secondly, though the accounts presented in the two main schools of cultural evolution are often presented as mutually exclusive or opposing, they can instead be viewed as complementary, though perhaps using different grains of analysis (Acerbi & Mesoudi 2015) or explaining different periods of evolutionary history (Sterelny 2018) or domains of culture (Mesoudi 2021). The next section turns to the second of these schools, the California School of cultural evolution, again showing how mindshaping can be promisingly integrated with this body of work.

3.2 Cultural Evolution and Mindshaping

In what follows I examine a set of interlinked accounts of the explanation of human culture and cognition that can be broadly grouped under the heading of cultural evolution, showing how they are significantly more amenable to incorporating the mindshaping account than their terminological commitments to mindreading suggest. These accounts are united in understanding elements that exist externally to *Homo sapiens* as core to enabling complex human cognition and our contemporary ecological dominance. In general, these accounts take their lead from a specific research program developed in California, beginning with Cavalli-Sforza and Feldman (1981) and expanded by Boyd and Richerson (1985), which draws on the tools of population genetics to model cultural change as an evolutionary process, though not identical to genetic evolution.

3.2.1 Cultural Evolution in California

Variously referred to as gene-culture coevolution (GCC), dual-inheritance theory or the California School, this strand of research into cultural evolution is the most formally rigorous. The central idea is that culture alters the selection pressures

faced by groups of humans, allowing culture to drive genetic change. The changes in lifeways caused by culture interact with the process of genetic selection, tweaking genetic inheritance to better fit the demands of culture. Furthermore, GCC theorists think that "Gene-culture coevolution enhanced first and second order accumulation: enhanced social learning, and learning how to learn socially." (Sterelny 2017b: 49). So, the selection pressures generated by culture lead to selection for skills for social learning.

In what follows I will use Joseph Henrich's comprehensive treatment of GCC (2016), to represent the California School. This approach presents an account of humans as a species that uses its social learning abilities to generate a rich cumulative culture. Humans are separated from other species by a store of "good tricks" that social learning allows us to preserve and transmit across generations. Humans guide their learning using a set of rich social learning heuristics, for example, discovering and copying what the majority of one's group is doing, or imitating group members perceived as successful or powerful. Through this selective social learning increasingly useful techniques accumulate through time, eventually culminating in rich, ecologically fine-tuned cultures. The secret of our success then lies in our capacity to engage in selective social learning, coupled with feedback into genetic processes that amplify this social learning faculty.

Henrich proposes that one (relatively minor) part of this facility for social learning is the skill of mentalising, the "ability to make inferences about the goals, preferences, motivations, intentions, beliefs, and strategies in the minds of others." (2016: 70). This is mindreading, and later Henrich proposes that this faculty or skill emerges from the "pressure for high-fidelity cultural learning", leading to "sophisticated abilities to infer others' mental states – theory of mind, or mentalising

and overimitation" (2016: 81). So, on this view, mindreading is primarily useful for enhancing the fidelity of social learning, allowing learners to acquire the norms, skills and know-how required to be part of a rich cumulative cultural milieu.

Thus, the California account appears to explicitly rely on mindreading, thereby causing problems for the incorporation of mindshaping into GCC. However, a reliance on a variety of rich mindreading generates some problems for the GCC account, which call this specific commitment into question. Firstly, the power of mindreading to ensure high-fidelity transmission is questionable and even if it was useful for such a task, the general importance of high-fidelity, one-to-one transmission for GCC is also disputed. Second, the types of empirical support Henrich calls upon to demonstrate the early development of mindreading can also be explained in terms of minimal mindreading. Thirdly, the idea of an innate theory of mind that is primed to discover mental states in others pushes the GCC account problematically close to the nativist evolutionary psychology it sets itself up to oppose.

The GCC account calls on mindreading to help enable high-fidelity social learning. However, there are reasons to be sceptical of the power of mindreading to underpin the exact replication of information gained through social learning. As we have seen, one of the main critiques of mindreading relates to the issue of holism. Any bout of behaviour is compatible with a potentially infinite set of PA ascriptions. This calls into question the idea that by employing the faculty of mindreading a social learner can get a reliable conception of their interlocuter's state of mind. Thus, it is hard to see how mindreading could facilitate the reliable, efficient and high-fidelity transmission of information, or act, as Henrich puts it, as a means "to make inferences about the goals, preferences, motivations, intentions, beliefs, and

strategies in the minds of others." (2016: 70). Reading the mind of a master knapper to determine if a movement was intentional or not is not a straightforward task. The intractability of mindreading *novel* content brings an essential element of unknowability to the operation. Thus, the introduction of full-blown mindreading into an instance of attempted social learning should be predicted to *hinder* the efficient transmission of high-fidelity information rather than aid it.

However, even if we grant the claim that mindreading helps enable high-fidelity information transfer (as proponents of mindreading indeed would claim), there are strong grounds to reject high-fidelity transfer of this sort as a crucial component of GCC in general. Indeed, this is one of the features of GCC theory that Morin (2016a) takes aim at, suggesting that in general the idea that cultural information is replicated with high fidelity is false, a critique that, given how cultural items evidently mutate, has some bite. Furthermore, the claim that GCC is always and only committed to high-fidelity information transfer has been disputed by Richerson, a key figure in the research program. In his review of Morin's book, Richerson claims: "Morin does not think that cultures are…imitated faithfully…he is correct on these issues" (2017: 206). Instead, rather than high-fidelity transmission GCC theorists tend to rely on a (relatively) large population allowing for repetitive exposure to the same cultural variant to create redundancy and overcome issues with high-fidelity transmission (Sterelny 2017b: 46). Thus, it seems then that insisting on the importance of high-fidelity imitation is a weak link in the GCC account, thereby undermining the importance of a strong conception of mindreading it supposedly requires (even though as we saw rich mindreading is unlikely to be of much use in social learning anyway: it isn't effective, nor is it needed).

To support his claim that mentalising plays a role in ensuring high-fidelity transmission, Henrich draws on empirical work that investigates theory of mind skills (2016: 71). This work documents children participating in false-belief tasks, which as we saw in the previous chapter is an approach often used to underpin the mindreading conception of social cognition in general. Henrich highlights experiments that show that infants can infer intentions and determine who has knowledge in a learning task. For example, children copy a model's intention in a task, such as grasping a toy, even when the model fails to achieve the goal and they appear to devalue subsequent reports by models who clearly mislabel objects (Henrich 2016: 71). However, the general conclusion that Henrich relies upon, that success in false belief tasks of this sort is evidence of the operation of a rich theory of mind module is disputed (Fenici 2015; Hutto et al. 2011). Indeed, the same results can instead be explained with a minimal mindreading or functional inference-based account.

In his analysis of the false belief protocol in general Marco Fenici proposes that "the data are compatible with the view that these tasks simply tap action prediction abilities, that is, infants' capacities to form and update their expectations about other agents' behaviour." (2015: 498). This suggests that what the infants are demonstrating is perhaps "not a capacity to attribute false beliefs but a more basic ability to predict goal-directed behaviour, which is refined in the time to consider what other agents are looking at." (499). This alternative interpretation of the same data shows that Henrich will need to provide more evidence if he wants to maintain the claim that mindreading is a hardwired cognitive faculty.

There is also a more general criticism of the incorporation of mindreading into the GCC account: the idea of an innate mindreading capacity pushes such an

account dangerously close to the evolutionary psychology position it explicitly opposes. The evolution of an innate capacity for mindreading presupposes the existence of a given mind that can be discovered. For this to be the case, i.e. for a theory of mind faculty to be adaptive, the cognitive architectures it discovers must be uniform and slowly changing. But this is exactly what GCC, broadly speaking, argues against. As Lewens (2015: Chapters 4, 5) highlights, the idea of an innate human nature that is static across groups and cultures is what separates evolutionary psychology and GCC. GCC is committed to the idea that feedback from the cultural environment develops the minds of a species like humans in situ, and thus the idea of drawing a sharp line between culture and human nature is incoherent. The idea of an innate, rich capacity for TOM developing in infants which allows them to discover the inner states of their interlocutors seems to imply that there is a fixed human nature that this mechanism can be trained upon to read. But according to GCC, the sort of intertemporal stability of mind that would have been required for selection to produce an effective hardwired TOM capacity is absent.

Ultimately, mentalising or TOM constitute only a minor element of the overall theory of GCC and the critiques outlined here suggest that it should not and need not be committed to the existence and use of such a capacity. Given that it carries little structural weight for the theory, supplying only the means for high-fidelity imitation, a skill that is in any case not central to GCC, it seems that the California school should drop this claim.

Yet human adults do engage in something like mindreading. If it is not the result of a hardwired module where does this practice come from? The Californian account doesn't seem to provide us with an easy answer to this question. The story it tells is interactionist insofar as it recognises the social aspect of human knowledge

generation and cultural accumulation. However, its claims place emphasis on the spread from person to person of knowledge and know-how, practical skills like complex food processing techniques or canoe building that can often be assessed for value or efficiency on the fly. There is little reference to feedback into individual cognitive development by cultural processes, or of environmental factors canalising learning and underwriting the emergence of advanced cognitive features. This is partly due to the modelling commitments of the California School; it is easier to develop models of agents influenced by conformist learning or prestige transmission biases than of individual cognitive development due to environmental factors. Yet their account does not rule out this type of interaction. To see how human cognitive development may be canalised by environmental resources we can turn to the work of Kim Sterelny. Sterelny takes seriously the role of the built environment in altering the cognitive capacity of human groups, and in so doing provides us with the tools to understand how initial pervasive mindshaping by a specific environmental niche could itself enable later "mindreading".

3.2.2 Sterelny: Niche Construction and Cognition

Of the various accounts of cognitive evolution considered here Sterelny's (2003, 2012, 2021) is the most explicitly amenable to the concept of mindshaping (Fenici & Zawidzki 2021). Known as the cooperative forager hypothesis his account presents early human in-group relations as characterised by cooperative information sharing as opposed to competitive Machiavellian interaction. He provides a detailed reading of the archaeological record in support of his proposals and his account has been well received in the literature (Farina 2013; Lewens 2014).

The key elements in Sterelny's story are a combination of niche construction and phenotypic plasticity (2012, 2021). Cumulative, pervasive niche construction shapes the environment of humans, making certain skills more likely to be acquired or specific inferences made. The phenotypic plasticity of human ontogeny means that a specifically structured environment can reliably canalise cognitive development. Since the genesis of their lineage hominins have modified their environments, a form of human intergenerational information transfer Sterelny places on par with genetic and cultural transmission. Immersion in deliberately modified physical and cognitive life worlds is a central part of what turns unadorned biological humans into selves. Acting in a specific niche means that even ostensibly self-directed learning is constrained and guided by prior actions of the cultural group. Humans are not just physical engineers, they are also often inadvertent epistemic engineers, modifying the epistemic environment to make specific knowledge more or less salient. Tim Lewens provides a nice example of such a case:

> Suppose adults have cultivated a fairly large area close to a village, and cleared it of poisonous vegetation. As in the case of extreme individual learning, juveniles are…left to find foods for themselves, with no instruction. Because their local environment is fairly safe, this 'fending for oneself' may also be fairly safe. Perhaps they develop tastes for the plants they find close by, and when adult they cultivate the same plants again, and remove others. (2015: 98)

This example of niche construction shows how the apparently natural environment is itself embedded with implicit models for learning. Thus, the domain in which learning takes place is a crucial factor in determining what is learned. And as a result, built or modified features of the lived environment robustly canalise the

development of certain forms of knowledge in groups across generations. Downstream niche construction is possible because human brains are highly phenotypically plastic and thus the environment they are immersed in plays a role in determining how they come to perceive the world.

Of the various theorists of cultural evolution canvassed here, Sterelny is the only one who places excessive emphasis on the role of the environment and its specific affordances in the evolution of human uniqueness. This emphasis reflects his assimilation of the main tenets of the externalist approach to conceptualising human cognition (Sterelny 2010). The other theories of cultural evolution I've reviewed tend to place excessive emphasise on resources inside the heads of individual human agents, forgetting or rejecting the fact that many of our capacities are partly constituted by elements in the environment, be they other humans or the social environment.

According to Sterelny, cognitively prehistoric humans become behaviourally modern through the skilled use of external elements in the social and informational environment. These environmental scaffolds guide knowledge accumulation, helping agents and groups prosper in informationally translucent worlds. In turn, this enables them to scale the ladder of cultural complexity without significant genetic or morphological changes in their cognitive machinery. Central to his story is a shift away from understanding complex human cognition as driven by selection pressures faced by *individual* human agents locked into a competitive intelligence arms race, to seeing it instead as a *group* achievement. On such an account, humans did not face selection pressure (at least initially) to become strong mindreaders aiming to outwit, or effectively police one another. Instead, they faced pressures to generate stores of knowledge that, in being shared, facilitate the complex coordination that profitable

cooperation requires. Additionally, by existing exogenously to any individual mind, this hard-won environmental knowledge benefits from a high degree of redundancy: individual agents could be (and inevitably would be) lost but, if the removal rate was slow enough then the knowledge would be preserved. The types of cognitive capacities that would be selected for in such an environment would likely have been ones which enhance social learning, like imitation and pedagogy, not ones geared towards discovering errant free-riders.

Sterelny's (2003: Chapter 11) account of the folk psychology that underpins this type of group formation suggests that our environmental scaffolds in part determine the contents of our theory of mind. On his account, folk psychology is not hardwired, rather it is built upon a set of existing perceptual processing mechanisms. These mechanisms, which allow for the determination of a conspecific's aims based on perceptual cues, are supplemented with scaffolded learning that trains individuals to be able to interpret the minds of others, minds that are themselves the products of prior scaffolding. Sterelny thus rejects the idea that humans are born with an innate theory of mind module, instead claiming that

> A good deal of our predictive efficiency may rest on other cognitive adaptations for interpreting others. These include our capacity to understand social environments and the ways these constrain behavioural options. And they include perceptual mechanisms. The way someone stands and moves can signal perceptually that they are drunk and belligerent. We do not engage in belief-goal modelling to keep out of their way. (2003: 269)

On this account, much quotidian human interaction makes use of skills that can be characterised as, at most, minimal mindreading without belief-goal modelling, i.e. without PA attributions. Indeed his account is very close to the type of IS

functionalism outlined in the previous chapter, and which lies at the heart of mindshaping. Beyond interpretations that are primarily perceptual, the folk psychology required for more complex interpretation is scaffolded into being by immersion in shared environments: mindshaping is prior to mindreading.

Sterelny's externalist focus on the importance of the developmental environment, coupled with phenotypic plasticity, sheds light on how cognitive skills like mindedness, and later, selfhood, can emerge. Humans who already are minded fill their world with mind talk. Such talk thus populates the lived world of children and serves as a model with which they can reliably construct the virtual entity that is the conscious mind. Thus, humans are apprenticed (Sterelny 2012: 35) into the world of mindedness by an environment that is seeded with examples. For example, children in forager societies are often provided with toy models of adult tools and encouraged and advised on their use (Hewlett et al. 2011: 1174-1175). These toy models allow them to learn how to use important artefacts in a low stakes setting. The rich traditions of lore, myth and folk tales present across all recorded cultures that detail the exploits of fictional characters, play this same role in constructing minded agents with selves: fictional accounts of an agent's exploits are toy models of a self. Additionally, children are encouraged to interpret inner states in socially approved terms and given the opportunity to explain away aberrant behaviour in terms of propositional attitudes. Thus, the tools of self-creation are socially learned but not through explicitly designed pedagogy, but rather due to immersion in a specific niche replete with a range of fine-tuned, culturally evolved scaffolds. This niche socialises individuals to become what their specific folk psychology regards as competent human agents.

If this gradualist account of the emergence of human uniqueness is correct, it suggests that the processes of mindshaping predate the modern mind that is taken to be the target of mindreading. The potential early existence of mindshaping, probably before "minds" as we know them existed suggests that many thousands of generations of mindshaping gradually brought the virtual entity that folk psychology calls the "mind" into being. Gradualism rules out the idea of a dedicated theory of mind module. If there is no bright line in early evolution between minded and un-minded, then there is no target to train a genetic module on.

Indeed, Sterelny's account is in direct opposition to the proposal in evolutionary psychology that contemporary humans operate in the present with unadorned prehistoric minds. Instead, humans, leveraging their phenotypic plasticity, deploy environmental resources in order to reshape the cognitive profiles of infants. Thus "the same initial set of developmental resources can differentiate into quite different final cognitive products" (Sterelny 2003: 166). Our minds are products of our times, not only of the savannah. This point can be extended to understand (some) human brains as formatted for the kind of rational thought that existence in contemporary lifeworlds requires (Wolfendale 2019). Phenotypic plasticity enables culture, or socially accumulated intergenerational knowledge, to install thinking tools into the brain. Cognition is not a *deus ex machina*. With careful ethnographic and archaeological work, we can trace the steps that allowed for complex cognition to emerge.

Sterelny grants that the Parisian, cultural attractors approach, is instructive with regards to the cultural items they are most concerned with, things like "stories, jokes, recipes, discursive verbalised items in the public domain." (2017b: 47). However, he observes, correctly, that their intellectualised conception of social

learning makes little room for the vast domain of cultural knowledge acquired passively through membership in a specific cultural group. Human life is, and has been for many thousands of years, structured by norms, norms that we acquire not through acts of ostensive communication but rather through the automatic unconscious social learning processes that interest the California School and are of key importance for mindshaping. Sterelny's account, by explaining cognition as scaffolded by the environment and relying only on thin, intentional-stance-based interpretation, is thus particularly amenable to the integration of the mindshaping hypothesis (Zawidzki 2013).

In the next section, we will examine one further, cognate, account provided by Cecilia Heyes (2016, 2017, 2018) of how advanced mindreading skills may have developed. Though Heyes rejects various elements of the GCC account, in particular the idea that culture has hardwired domain-specific abilities in humans for social learning, her account isn't necessarily in conflict with the modelling claims of the California School and extends Sterelny's proposals about the role of environmental affordances in an empirically rigorous manner. In addition, her account has the added value of laying out exactly how a capacity like quotidian mindreading may have emerged, providing an intriguing description of its origins that lines up remarkably closely with the account of mindshaping, an account that is missing in Henrich and Sterelny's work. Furthermore, it has been suggested that Henrich is also beginning to move in Heyes's general direction and emphasise the role of domain-general processes, as opposed to specifically genetic modifications in driving cultural evolution (Sterelny 2020: 770).

3.2.3 Heyes: Cognitive Gadgets and Mindreading

Coming from a background in cognitive science, Heyes presents a unique account of the evolution of human social cognition. Though contentious (Morin 2018; Sperber 2017) it is taken seriously in the literature (Sterelny 2017b, 2020) and thus warrants exploration here. Heyes proposes that humans are endowed with what she calls "a genetic starter kit consisting of enhanced social motivation, attentional biases (e.g. to faces and voices) and souped-up domain-general mechanisms of learning and memory" (2018: 7). This universal domain-general toolkit then allows humans to acquire a set of what she calls cognitive gadgets. These gadgets are domain specific skills, things like language, imitation and what she calls mindreading. Thus, these capacities are not the product of cognitive changes rooted in genetic evolution. Instead, they evolve culturally and are reliably installed in human minds immersed in modern lifeways in the course of development.

Though Heyes explicitly refers to mindreading throughout her work, her understanding of the capacity differs from mainstream TT/ST interpretations. For one thing, her claim that human competency with mindreading is not innate has obvious affinities with the mindshaping view. Indeed, Heyes explicitly rejects the idea that mindreading is a cognitive instinct or that it is the result of trial and error theory building by children as proposed by theory-theory. Instead, like Sterelny, she thinks that "Children are taught about the mind by members of their social group...[forming] a conceptual structure enabling the ascription of mental states to the self and others" (2018: 147). She proposes an analogy between mindreading and print reading: just as learning to read structures writing, so too learning to "read" other minds structures mind-making. Drawing on the work of McGeer we encountered in Chapter 2, Heyes suggests that being taught to mindread serves a

regulative function: children learn “not only that behaviour *can* be, but that it *should* be, produced by rational interaction between beliefs and desires” (2018: 148). The inclusion of a regulative dimension in her theory shows that Heyes implicitly recognises that mindreading is in part also mind*shaping*.

The hypothesis, for which Heyes provides strong empirical data (2018), that the skill of "mindreading" is learned in a manner similar to reading undermines the idea that there are rich "minds" that pre-figure the emergence of this teaching. If there were already “minds” before teaching, mindreading would come for free, with no instruction necessary, as my mind, to which I have privileged access, would serve as a model for reading your mind or at least understanding that you have a mind to read.

Children are taught to be able to read what words represent, but words do not represent anything outside of this learned structure, they are merely marks. Likewise, children are taught to "read" what other minds "represent", but minds do not represent anything without this representative structure. Thus, if we take Heyes’s analogy seriously we see that the domain mindreading aims to read is constituted by this learned practice of reading; the PAs that mindreading aims to track do not exist independently of the culturally evolved capacity to track them (Fenici & Zawidzki 2020; Clark 1994). This is an artificial structure laid down over the biological substrate that is the human brain. The free-floating rationale (Dennett 2018) for installing such a regulative script is the efficient achievement of joint action, i.e. coordination. However, the phenomenology of mindreading gets in the way of this minimalist interpretation of what is happening.

Heyes goes on to reject “individualist conceptions of mindreading” describing it instead as a “collective achievement” (2018: 154-155). On her view,

this collective achievement is not underpinned by the representation of mind states. What she calls "implicit" or basic mindreading can be explained by the use of domain-general mechanisms, what she calls system 1 processes, none of which represent mental states (2018: 163). This account of implicit mindreading is very similar to the functional/instrumental rationality Zawidzki relies on for mindshaping (2013). She goes on to say that explicit mindreading, the rich PA attribution that characterises human life, is built on these non-representational system 1 processes (Heyes 2014). In other words, humans come to build a rich vocabulary of mind states, that determines what is considered "rational interaction between beliefs and desires" (2018: 148) upon a thin substrate of instrumental interpretation. Overall, Heyes's cognitive gadget conception of human cognition bears strong affinities with the mindshaping account. Heyes, however, does not, so far, take her proposals regarding the similarities between mindreading and print reading to their logical conclusion. If she does, then she will find herself in the realm of mindshaping: the domain that mindreading discovers is itself constituted by the practice (Fenici & Zawidzki 2020).

So far we have seen how, despite appearances to the contrary, the current leading theories in cultural evolution need not rely on a strong conception of mindreading, or posit an innate theory of mind. Though the terminology remains intact, the concept of mindreading these theorists rely upon can in each case be plausibly redescribed as something more minimal, relying on perceptual processing mechanisms or instrumental rationality. As a result, it seems plausible that mindshaping can potentially be incorporated into the foundations of human cognitive evolution. This is an upshot of the negative component of the mindshaping theory,

i.e. the deflation of the mindreading concept in a manner that detaches it from a reliance on PAs that reside in the minds of others to be discovered.

Beyond this negative project, the positive content of mindshaping also makes a good fit with the accounts of cultural evolution just reviewed. Though mindshaping is not referred to directly by Henrich, Sterelny or Heyes, their theories presuppose and utilise the same fundamental mechanisms that drive mindshaping. Namely, the core processes of cultural evolution are types of information transfer between agents guided by attention to perceptual cues and social learning strategies like prestige bias or conformism. Thus, there is a sense that mindshaping merely *names* an implicit feature of the cultural evolutionary story: once we take the plasticity of mind seriously, then the information accumulated through social learning influences or shapes the minds of the agents involved.

That humans possess a high degree of phenotypic cognitive plasticity is empirically well supported. For example, work comparing chimpanzee and human levels of myelination, a lipid that locks in axonal connections between neurons, demonstrates that human brains are born with lower levels, and are thus more plastic, and remain so throughout their lives, relative to chimps (Miller et al. 2012). Once we grant that humans are not born with an innate theory of mind module or rich mindreading skills, and that functionalist or perceptual interpretation suffices to interact with conspecifics, while at the same admitting the power of cultural information to influence cognition via phenotypic plasticity, then we are led to an explanation of mindreading that looks very similar to that proposed by mindshaping theory. This is that our facility with mindreading is learned and that the theory of mind it uses is a cultural construction, one that emerges through interactive processes and which through interpersonal regulation makes itself "true".

The degree of consilience that mindshaping displays with leading theories of cultural evolution lends support to the mindshaping project in general. However, the work just surveyed in cultural evolution can be of use to the mindshaping account beyond providing it with increased plausibility. In particular, contemporary theories of cultural evolution generally draw attention to the mechanisms of model choice during social learning. These processes are of interest because, fundamentally, mindshaping is a form of social learning. Understanding how social learning is guided will thus show us how mindshaping processes constrain individual beliefs, and how beliefs are strategically linked to and determined by cultural groups.

3.3 Model Choice in Mindshaping

Both Zawidzki and McGeer discuss the processes by which various regulative mindshaping models are propagated in quotidian interaction – conformism, imitation, immersion in regulative folk psychologies – but they do not pay particularly close attention to how individual agents sort and strategically adopt information, or to how different contexts may interact with the processes of mindshaping. Not everything in an individual agent's social environment shapes their mind; some filtering process must guide model choice. This filtering process is of great interest to cultural evolutionists as it is a central driver of cumulative cultural evolution. In what follows I want to examine work in cultural evolution that looks at model choice and the role it plays in social learning to show how it can be used to better understand mindshaping.

At first glance, the heterogeneity of approaches to strategic social learning in cultural evolution appears to preclude a unified story. These range from thin conceptions, like Henrich's (2016), who models social learning as the result of

unconscious, automatic processes that emerge early in human development, to thicker conceptions like that of Morin (2016) who sees social learning as strategic and driven by individual choice coupled with the intellectualised reconstruction of information in social interaction. This fault line between the Californian and Parisian interpretations is often seen by the theorists themselves as crucial (Morin 2016c) and the apparent difference between these views can again be partly explained by reference to their differing empirical bets. Henrich and the California School idealise the processes of social learning in order to be able to use formal models to gain explanatory traction. However, this idealisation necessarily involves a degree of simplification. This is opposed to the Paris School which seeks to understand contemporary traditions in their full richness and thus must reach for models that attempt to incorporate more complex capacities, ones that are arguably more cognitively realistic (Andre & Morin 2011). The dominance and success of the California approach suggests that idealisation is not misguided, however, as they acknowledge, this does not necessarily preclude the existence of more complex and strategic social learning (Richerson 2017; Sterelny 2017b).

This apparent discontinuity can perhaps be explained away if we take on board a suggestion of Dennett (2016) that proposes a steady "de-Darwinization" of cultural evolutionary processes through time. In other words, through the evolution of more powerful thinking tools, or "cranes" in Dennett's terms (2016), humans have come to supplement bottom-up automatic social learning strategies with a complementary suite of more intellectualised, or top-down strategies akin to those proposed by Morin. These capacities emerged in succession, thanks to the cultural evolution of more complex cognitive capabilities. This Dennettian synthesis can perhaps shed light on how social learning, and as a result mindshaping, has changed

through time and how it may operate in our contemporary social world, mediated by succeeding communications technologies from reading to social media. By adopting a gradualist view of this nature we can see that both of the main schools of thought seeking to explain cognitive evolution – Paris and California – can be of use in explaining how social learning is strategic. In what follows we will examine these accounts in turn, showing where and how they interact and how they can shed new light on mindshaping processes.

3.3.1 Social Learning in California

Henrich argues that humans are primarily set apart from other species by their ability to engage in strategic social learning that allows for the accumulation of useful knowledge across generations. This accumulation, however, is potentially undermined by individual innovation as this can inadvertently lead to the loss of valuable information. Thus he emphasises social learning that is automatic, early emerging and high fidelity. His core hypothesis about learning can be summed up as "various human groups do things a certain way for ecologically tuned reasons, and, more or less, the best bet for survival is to copy those around you and don't think too much about it".

Henrich's flagship example of this sort of transmission is the processing of Manioc for safe ingestion. Manioc was (and is) an important staple crop in many marginal environments. However, due to residual cyanide content, it requires a complicated series of steps to be rendered edible. As Henrich reports it, these steps are both laborious and their benefits causally opaque, because the poisoning effect of small amounts of cyanide is cumulative and takes many years to become perceptible. This combination of causal opacity and labour cost could easily lead an innovative

agent to discard one or more of the necessary steps. Such an innovation would appear sensible in the short term, but its long-term effects would be harmful due to the accumulation of cyanide, which can result in chronic illness. For Henrich, this is a paradigmatic case where "follow-the-crowd" learning is highly adaptive. Individual innovation would decrease fitness, and destroy the rich cumulative information contained in the established manioc processing routine.

He extrapolates from this example to show how a series of innate social learning strategies act to fine-tune from who, what and when to learn. For example, these heuristics direct agents to imitate successful, prestigious or older members of their cultural group. These cues of skill and success drive (relatively) selective imitation and allow for the accumulation of cultural know-how that is often causally opaque but tuned to the specific ecological circumstances faced by a cultural group. Henrich's discussion of social learning dovetails with the processes of mindshaping examined in the previous chapter. Indeed, imitation and conformism alone suffice to drive this Henrichian story of cultural evolution forward. Likewise, as we have seen, the bulk of Zawidzki's account relies on these same automatic, unconscious processes.

However, this focus on automatic social learning strategies that unconsciously and pervasively track and copy salient group members has been criticised on two fronts: Morin (2016) suggests that the idea that humans are pervasive automatic imitators is implausible as the sheer amount of information contained in the cultural environment (even if the imitation is focused only on the successful) means that the majority of it is ephemeral and discarded; and Heyes (2017: 91-92) points to research showing that the various strategies implicated in Henrich's account are observed in other species. Both commentators then propose

that human social learning is more complex than Henrich seems to assume. While Heyes agrees that humans make extensive use of automatic social learning strategies, she supplements this account with what she calls metacognitive social learning strategies (MSLS). Morin, however, largely dismisses Henrich's story, instead placing the burden of our unique cultural capacities upon ostensive communication, as discussed above.

The empirical success of California's modelling strategy suggests that Heyes's bet is more reasonable, yet Morin's criticisms can't be entirely dismissed – the cultural environment is awash with information, the vast majority of which is never copied. In his recent overview of the contemporary lay of the land in cultural evolution, Sterelny (2017b) crystallises the dispute: does the difference between automatic learning heuristics and the more cognitively rich/discerning approach of the Paris School actually make a difference at the level of population analysis?

To this end, Sterelny proposes four alternative empirical conditions: (1) perhaps the existence of populations of either trusting social learners or discerning social learners doesn't make a difference to the fidelity of cultural transmission. Or, it does make a difference and (2) either discerning learning reduces the fidelity of cultural transmission, (3) or alternatively, it increases its fidelity. Or, finally, (4) perhaps both approaches are partially correct and initially simple learning heuristics have become more nuanced over time (Sterelny 2017a: 146). It is this fourth possibility I want to explore as, building on Dennett's de-Darwinization strategy, it can capture the strengths of, and continuities between, the two dominant models of cultural evolution. Additionally, Sterelny's own account is beginning to lean towards

a two-phase explanation (2021)[18], and, as we will see, Heyes (2016, 2017) reaches a similar conclusion.

3.3.2 Learning Socially How to Learn Socially

The criticisms of the Californian approach to social learning can be partially defused by recognising that they purposefully idealise the processes of social learning in order to render them tractable from their modelling perspective. Taking their thin conception of the processes of human social learning strategies as a full descriptive account is misguided, and to do so is to construct a straw person argument (Richerson 2017). Morin in his (2016a), for example, seems to do exactly this – a strategy that leads him to, rhetorically at least (see 2016b), largely dismiss the work of the California School. Sterelny (2012, 2017a, 2017b, 2020) presents a more realistic account of social learning, one that broadly endorses the Californian approach, while incorporating Morin-friendly extensions tuned to real world cognitive concerns.

For Sterelny automatic social learning heuristics are an important part of the human suite of social learning faculties, but not the only part. Recognising the constraints that emerge from the Californian modelling project, but not being bound by them himself, Sterelny proposes a more nuanced explanation. On his account social learning results not only from automatic model choice but also through individual agents paying attention to the information those models transmit and the learning context itself. Learning heuristics are varied and prompt different learning in different domains. Sometimes trusting other agents is suitable but in other contexts agents should incorporate assessments about outcomes and intentions. He

[18] Indeed, Sterelny in his (2021), credits Dennett for first proposing such an approach to him.

suggests that Henrich's flagship manioc example is an outlier, not representative of the typical, primarily procedural, information transmitted culturally by our Pleistocene cooperative foraging ancestors. On his account, the majority of skills early hominins had to acquire were less causally opaque than manioc processing and the chains of causal feedback were shorter. For example, Acheulean hand-axe manufacture does not present the same complex causal chain as a multi-step, temporally extended toxin reduction procedure. Once familiar with the basic technique, a knapper could assess the effectiveness of her strike technique on the fly, and engage in trial-and-error experiments aimed at improving her products.

Sterelny stresses that the understanding an agent brings to a practice is a matter of degree and is relative to the domain of action. In a small, relatively stable domain an individual can come to understand the causal interactions between the various components of some technique. Californian blind imitation may be an effective strategy in some cases, but in many instances agents do attempt to cognitively assess and determine the best course of action. For example, in a noisy environment it is optimal to copy the majority, but in less noisy settings a sufficiently sophisticated agent can innovate, engaging in hybrid learning, both copying the successful but also recreating and modifying the modelled behaviour. Alternatively certain types of knowledge and behaviours, in particular symbolic beliefs and actions, are undermined by individual innovation as this hinders their robust signalling role. Thus, the various social learning strategies employed by humans are arranged along a gradient from automatic heuristics to intelligent appraisal of goals and methods. In each case the strategy deployed is cued by the status of the learning domain and the type of knowledge being disseminated.

To explain how such a sophisticated suite of SLSs may have emerged, in his most recent work examining the emergence of human uniqueness (2021) Sterelny divides hominin evolution into two periods, those of early and late hominins. He thinks that these were, broadly speaking, two distinct and significantly different periods of human evolution. The shift between these two stages took several million years and involved a major increase in the volume of social learning required by an individual to survive. For early hominins, starting with the australopithecines around 3.4 million years ago, cultural repertoires were small, and, across individual human lifespans, relatively unvaried. The Parisian picture of (chimp-like) humans becoming discerning collectors of attractive cultural information thus mischaracterises early hominin lifeways. This is because the majority of information passed across generations was procedural, relating to specific skills like tool manufacture and hunting practices. In this setting a combination of pervasive niche construction and automatic learning heuristics was sufficient to reliably populate the manifest image of the agents it contained.

However, as this thin layer of culture began to accrete through time, through the repeated use of simplistic learning heuristics like imitation and conformism, coupled with strategic but automatic model choice, it began to generate more complex worlds, in which increasing discernment *was* required both to sort information and to be able to effectively signal one's status in complex social formations. By around 200,000 years ago human children had a lot to learn to become competent members of their group. These changes in hominin lifeways were multidimensional: group size, skill specialisations, hierarchies, and sexual division of labour. In contrast to the primarily procedural nature of the knowledge transmitted in early hominin life, later hominin life worlds required the transmission of

declarative and subtle symbolic knowledge. Additionally, as social complexity increased so too did the importance of fine-grained status signals for enabling coordination, and social learning that was responsive to such cues. Operating in this new learning domain required agents to be able to both access and more efficiently filter social signals and symbols.

The increased volume and complexity of knowledge and the emergence of more socially stratified hierarchies in the late hominin period created selection pressures for the more discerning learning explored by the Paris school; however, contrary to their understanding of these capacities as consciously deployed genetic endowments, they can instead be understood as products of cumulative cultural evolution of the sort modelled in California. These capacities, though strategic and learned, do not usually result in explicit, consciously represented social learning choices on the part of their users. Automatic social learning strategies, which are at the core of mindshaping, were, and indeed are, crucial to the formation and transmission of human culture. But once they enabled the processes of cultural accumulation to emerge (supported by favourable environmental circumstances) new tools, or cranes, evolved with allowed humans to then get better at social learning. Late hominins began to socially acquire tools for social learning itself, capacities like language, and theory of mind skills and selfhood (Sterelny 2012: 26; Clark 2008).

This general process is an example of Dennett's de-Darwinization of culture referenced above, whereby initially bottom-up automatic processes of selection gradually become more top-down and self-directed, more efficient at searching design space:

> Human culture started out profoundly Darwinian, with uncomprehending competences yielding various valuable structures in roughly the way termites

> build their castles, and then gradually de-Darwinized, becoming ever more comprehending, ever more top-down organising, ever more efficient in ways of searching design space. In short, as human culture evolved, it fed on the fruits of its own evolution, increasing its design powers in ever more powerful ways. (Dennett 2017: 150)

As a result, initially automatic, Darwinesqe social learning generates the tools required for the culturally fine-tuned direction of social learning. Cultural evolutionary processes gradually led to the emergence of a suite of higher-level social learning tools that are themselves learned socially. Eventually, advanced coordination enabling tools like selfhood emerge, which as we saw in Chapter 2, acts as a regulating narrative constructed using the feedstock of the broader group culture. The self is (from the perspective of strategic interaction) a tool that allows for agents capable of complex cognition that otherwise threatens to generate irresolvable coordination clashes, to remain inter-temporally consistent and intelligible to one another. Additionally, in limiting possible behaviours and self-assignment of PAs, selfhood assists agents in the efficient division of attention between various competing ephemeral cultural elements, of the sort that worry the Parisians. The story one tells oneself about oneself plays a role in guiding attention. An agent who is a self-described baseball fan will be likely to pay attention to baseball statistics but discard horse racing odds. This is a learned social learning strategy, but not one deployed consciously, the agent merely has an interest in some types of information over others due to their culturally derived self-understanding.

The proposal that humans make use of culturally evolved tools to direct social learning is echoed by Heyes (2016). She proposes that in addition to automatic, domain-general social learning strategies, humans are capable of

deploying what she calls *metacognitive social learning strategies* (2017: Chapter 5; 2016) [MSLS]. These "are reportable, domain-specific rules that represent "who knows", i.e. properties of the cognitive processes of the rule user and of other agents. Potential examples of metacognitive SLSs include *copy the boat builder with the largest fleet* and *copy digital natives*" (2016: 4, emphasis in original). Arguably many automatic social learning strategies are themselves examples of implicit metacognition, insofar as they track relative certainties about states of the world, for example "copy the majority when uncertain" involves some assessment of certainty, even if not explicit.

MSLSs are a species of discerning learning, but unlike the Parisian approach, during such learning the agent is not just assessing the specific skill or cultural entity for usefulness or attractiveness, instead they are applying a rule that is culturally inherited and domain relative, and as the idea of implicit MSLS reveals, is not necessarily consciously represented. Heyes's exposition of the concept is somewhat internalist, likely reflecting her own disciplinary home in psychology, but the basic idea is sound. Humans, unlike other species, don't just apply automatic and unchanging learning strategies or engage in individual innovation; they also make use of social learning rules that have evolved and are propagated culturally and tuned to specific domains and the potential cognitive properties of others. The idea is also intuitively plausible (for what that's worth): most people make metacognitive assessments of their own and others' knowledge when deciding whom to trust, and often follow specific socially sourced rules such as "consult a respected movie critic" in advance of choosing a film to watch.

This fact has led Hugo Mercier to describe humans as learners who practice open vigilance (2020). By this he means that we are not gullible dupes, copying

everything we see or find interesting. Rather, we assess the source of knowledge and the domain of learning when choosing whether to adopt a new piece of information or not. Furthermore, building on Mercier's work, Morin and colleagues have shown that humans are surprisingly unreceptive to social information in general, they are stubborn about preserving the beliefs they already have (Morin et al. 2021). Ultimately social learning is responsive to different learning domains, and specific cues, or signals received, and these strategies are, according to Heyes, adjusted to new domains as they emerge via cultural evolution.

3.3.3 Social Learning in New Domains

We have seen how the main processes of mindshaping encountered in the previous chapter, imitation, pedagogy and conformism, can be understood as different modes of social learning. Humans automatically filter the social world to determine good models for social learning from bad, useful models for mindshaping from unnecessary. The social learning literature attempts to unpack these processes and understand how specific social learning strategies are cued in specific domains. The California School models various automatic processes that can support cultural accumulation at the population level, showing that such processes are likely to dominate when individual learning is difficult or costly and when learners are uncertain (Boyd & Richerson 1988; Nakahashi et al. 2012). Sterelny uses rich archaeological and ethnographic work to support his account of hybrid social learning that is somewhat domain relative. Likewise, Heyes draws on various experiments to demonstrate the existence of metacognitive strategies for social learning (Shea et al. 2014; Bahrami et al. 2012) and draws our attention to their variation across cultural groups (Mesoudi et al. 2015).

The vast majority of human social learning is best characterised by the automatic processes modelled by the California School. For example, humans do not (typically) consciously choose the norms that structure their interactions with the world or indeed their "type" of self. As children, people are indoctrinated into the norms of their community by caregivers and peers, they do not (initially) choose to be a conservative or liberal, a Christian or a Hindu. Yet, as the inclusion of pedagogy in the processes of mindshaping demonstrates, some human social learning or mindshaping can be directed, either individually or collectively due to immersion in institutions that are the product of conscious design. And thus while individual agents have relatively fixed cognitive and belief profiles, they *are* malleable at the population level. Institutional pressures or specific incentives, be they cultural or material, can alter the belief profiles of groups, for better or worse.

At the core of social learning is the idea that individual agents engaged in social learning primarily seek to detect signals from their close cultural peers that demonstrate prestige or knowledge, they do not consciously choose, in the moment, their preferred learning strategy or the information they wish to learn. In a sense, this general strategy was chosen by cultural evolution which favoured groups that are better at coordinating. Thus, social learning is strategic without being consciously directed or guided by individual decision making: it is strategic to adopt, and update through time, the norms and mores of one's peer group, but people don't *consciously* go about copying prestigious others most of the time, they just automatically do so. But they do "choose" in the economic conception of choice as behaviour that is responsive to incentives. If some beliefs become too costly or cease to offer any signalling power they will cease to be propagated through a group, even though no individual may have consciously reasoned their way to changing their mind.

In day-to-day quotidian contexts individuals are stubborn learners as Mercier (2020) suggests, they do not incorporate information divergent from group consensus lightly. This is strategically beneficial, as shifting one's cultural positions too far from friends and family would lead to a reduction in coordination viability. But if their group as a whole faces significant disincentives linked to behaviours or beliefs then individual beliefs can shift slowly as the group as a whole adopts new signals. For example, if conservatives in the United States were to suffer sustained electoral defeats due to their stance on abortion, it is likely they would slowly shift their position on this issue. This shows us that to bring about change in cultural norms policymakers should aim interventions not at individual learners, but rather at modifying incentives at the population scale. We will return to the practical applications of this view of social learning in the next chapter.

Taken together, the rich account of model choice presented in the cultural evolution literature can be used to better trace causal pathways and potential issues that may arise when social learning/mindshaping takes place in new domains. It can be used to highlight the fact that social learning is not solely guided by strategies that require agents to rely on the epistemic trustworthiness of the interaction domain. Indeed, the social environment is not always trustworthy and our evolved strategies for social learning often price this in. We are wary learners, and thus usually favour models who can honestly, and reliably signal self-similarity or community relative prestige markers.

My referring to social learning as strategic, and as relying on domain-relative assessment, does not imply that the mindshaping enabled by these processes is consciously directed. People do not make conscious individual choices about what or whom to be shaped by, or learn from, rather they are receptive to subtle signals and

cues that direct their attention. The additional detail provided by the de-Darwinization argument is that these sensitivities are not necessarily hardwired, they are responsive to changing learning domains, who “knows” in one context may be signalled by entirely different cues to another and those cues can be “learned”. Crucially the domain, or context in which learning takes place determines the optimal strategy to deploy and the cultural affordances humans engage with modify their social learning strategies through time.

Interestingly, Heyes hits on this key fault line when she proposes the MSLS “copy digital natives” as a strategy that directs social learning in the contemporary world. New, more opaque, domains often require updated strategies to guide social learning. For Sterelny it is capacities like this that are especially important when the expertise passed on from the proceeding generation is rendered obsolete by fast changes in the environment. "In such fast-changing domains, our capacity to navigate novelty depends heavily on cognitive tools." (2012, 27). The digital sphere is a paradigmatic case of a fast-changing domain and the manner in which it alters the signalling environment in which social learning is cued and guided, especially with regards to the signals that guide automatic social learning is dramatic.

For example, in the online context cues that are often used to direct social learning, and implicit mindshaping in quotidian domains, are open to manipulation, and thus no longer trustworthy. Formerly useful (or at least relatively honest) learning cues, such as prestige or majority opinion, are rendered problematic as the backend technology that governs the generation of these metrics is opaque. Visibility on online platforms such as Facebook, Twitter or YouTube is controlled by deep learning driven recommender algorithms, that prioritise engagement with the platform above other goals (Covington et al. 2016). This interferes with the

trustworthiness of social learning cues. The opacity of the systems that govern the generation of these cues coupled with their potential untrustworthiness renders them increasingly inappropriate as guides to social learning. In this new environment, our existing cognitive tools may thus lead us astray, or undermine the crucial coordination enabling processes of social learning in this new context.

As a result, in response to these changes to the learning domain, agents may be incentivised to adopt more conservative learning cues or to develop new MSLSs. For example, in an opaque context information shared by close group members (friends) is more likely to be adopted as the perceived trustworthiness of the source is high. Social learning in general is often guided by perceptions of self-similarity, and thus honest signals of group membership or similarity may become more valuable in such a context. This potentially pushes agents to adopt strategies for social learning that are conservative and excessively geared towards discovering trustworthiness in interlocuters or to use new ways of discerning self-similarity, ones tuned to the specific informational affordances of the online domain. The exact manner in which the structure of this new informational domain interacts with social learning will be explored in greater detail in the following chapter, specifically in the context of the role played by interpersonal signalling in generating coordinated groups.

3.4 Conclusion

We have seen how mindshaping fits into the cultural evolutionary story, thereby demonstrating the epistemic virtue of consilience. Furthermore, in the sense that it relies on extant processes core to the theory of cultural evolution, thereby removing the need to posit an additional domain specific mindreading module,

mindshaping theory is also more parsimonious than the theory it replaces. Thus, at least by the lights of cultural evolution, which currently represents our best theory of how human cognition evolved, mindshaping is a plausible hypothesis. Furthermore, viewing mindshaping through the lens of cultural evolution draws our attention to the importance of model choice in these iterated processes that play out across massive timescales. This attention to model choice can enhance our understanding of mindshaping, and help us understand how the structure of the learning domain may interfere with or modify its processes.

Issues of empirical verification loom large in the cultural evolutionary literature. The omnipresent risk of manufacturing a "just so" story is a serious issue. Concerns like this drive the Californian modelling approach and hold the Parisian school back from more widespread influence. In Sterelny's view "the attempt to understand the lifeways of our ancestors and the ways those lifeways served as foundation and springboard to ours is challenging. But it is not hopeless." (2021: 6). It isn't hopeless because, at its best, models of cultural evolution do not just provide coherent narratives. They also make frequent contact with the data available across disciplines, weaving together evidence from the fossil record, anthropology, psychology, genetic analysis and cognitive science. The necessity to keep in touch with the data also animates Heyes's work. Together with a colleague, she rallies the California School and broader cultural evolutionary research community to

> get geekily specific about cognitive science; to stop relying on plausibility arguments—whether or not they are based on mathematical modelling—and to start testing hypotheses about the nature and origins of cultural learning against the full range of data from comparative and developmental

> psychology, human experimental psychology, and cognitive neuroscience. (Clark & Heyes 2017: 297)

Broadly speaking my efforts here aim to feed into such a task. However, in addition I take an explicitly normative focus – seeking not just to understand the phenomena at hand, but also to influence our attempts to grapple with the world and enable beneficial collective action. Sterelny's focus on the environmental niche and its role in constructing and populating human cognitive profiles reinforces the importance of this task. If the environment plays a significant role in canalising development, and generating a specific cognitive profile then we need to pay close attention to the types of environment contemporary humans occupy.

Overall the coherence, plausibility and empirical support for the cultural evolutionary project is strong. However, as of yet little use has been made of the work to inform policy. For example, Henrich in his recent monograph is relatively sanguine about the future of humanity, maintaining that as long as we can keep information transfer happening, i.e. avoid echo chambers, humanity will remain on a relatively egalitarian path. In part the reluctance to engage with the contemporary world reflects the scale of analysis at which cultural evolution operates. The massive timescales of evolutionary history, coupled with the use of population-level analysis push these theorists away from consideration of intermediate or micro scales. Yet undoubtedly these theories have implications for living humans. The lack of attention to the normative implications of cultural evolution models is a lacuna to be filled.

I close this chapter with an observation about where the implications of taking cultural evolution seriously leave humans and their relationship to knowledge. The story it tells us about the emergence of human uniqueness is ultimately one that

deeply undermines the idea that humans are individually special creatures. Instead, our uniqueness emerges in collective capacities, themselves products of cultural evolution. It also supports the proposal that the most widespread and influential understanding of the evolution of human social cognition – that of atomistic individuals selected for their competency at reading other minds – requires substantive revision. Instead, the order or emergence must be reversed: atomistic individuals arrive late on the scene, with selection pressures for mindshaping abilities that keep human groups cohesive and coordinated coming first. The stability offered by mindshaping-based group formation then eventually enabled the material basis for the emergence of atomistic selves. Thus mindshaping set the stage for the emergence of individuality, making individualism paradoxically, as Dewey recognised (1963), a social achievement.

The contingency inherent to an organism that relies on culturally evolved components grafted onto a genetic base for its survival can be dangerous. Human minds are open-source, and this leaves them potentially vulnerable to exploitation or pernicious or inadvertently sub-optimal modification. But this same contingency can also be a source of hope. Richard Rorty saw the fundamental contingency of human nature as a call to reinvent ourselves as a more humane species. In his words: "Nowadays, to say that we are clever animals is not to say something philosophical and pessimistic but something political and hopeful — namely, if we can work together, we can make ourselves into whatever we are clever and courageous enough to imagine ourselves becoming." (Rorty 1998: 175). A lot is hanging on Rorty's "if" here – a mere possibility may not ever be realised. Promoters of the project of bringing about a better world need to be fully cognizant of the adapted means of mind creation that underpin our uniqueness. Helping more humans to work together

requires supple knowledge and use of mindshaping processes to guide the design of epistemic environments, against a backdrop of competing, pernicious or perversely incentivised actors. By arming ourselves with a fuller understanding of how mindshaping generates coordinating groups, maybe Rorty's basis for hope can be strengthened.

Chapter 4 – Coordination Online

Across the previous three chapters I have developed a set of interlocking claims about the constitution of human social cognition:

1. Coordination plays a central, yet undertheorised, role in enabling efficient joint action in groups of humans. Additionally, despite its importance for human social life, the most widely accepted explanation of how coordination is achieved, implicit in game theoretic accounts, fails to adequately explain coordination dilemma resolution.

2. Specifically, efficient and accurate mindreading is faced with the fatal issue of computational intractability: any bout of behaviour is compatible with an infinite set of propositional attitudes. Hence, the ascription of an accurate PA, one driving the behaviour of another agent, is effectively impossible. Yet, despite the implausibility of mindreading-based accounts of human social cognition and, in particular, coordination, such accounts still dominate philosophy of mind, in turn underwriting their continued use in game theory.

3. In recent years a set of alternative proposals to mindreading have emerged, broadly termed interactionist accounts. The most empirically rigorous of these new approaches is known as the theory of mindshaping. On this account, any faculties humans may have developed for mindreading must rely on pre-existing mindshaping mechanisms. Thus, mindshaping is explanatorily prior to mindreading. Mindshaping acts to conform, or regulate the cognitive profiles of individual agents to match shared group models. As such, the purpose of folk psychology is not to describe or discover other minds, but rather to shape and regulate them by providing models for minds

to fit. Mindshaping can explain how humans can engage in fluid and tolerably efficient coordination because it resolves the intractability issue by suggesting that, rather than discovering internal mind states, interactants negotiate, on the fly, acceptable solutions to coordination dilemmas. These solutions then determine the contents of minds and the structure of social groups. Effective coordination is the outcome of mindshaping, and any appeals to pre-existing determinate mindstates are discarded.

4. The theory of mindshaping not only resolves issues in philosophy of mind, and supplements the game-theoretic account of coordination resolution, but it also slots well into the most promising theories explaining the evolution of human cognition. These theories show how cognitive plasticity coupled with hypersociality in humans can give rise to rich cumulative culture and explain how complex cognition can emerge in a broadly ape-like species. Their emphasis on plasticity and the use of environmental scaffolds to support cognition aligns with the mindshaping view of socio-cognitive competence. Indeed, the same processes that drive cultural evolution also underpin mindshaping. Thus, mindshaping emerges as an implicit assumption central to our leading theories of the evolution of both social and general cognition. Coordination is thus, in some respects, hardwired into our cognitive machinery, and it is fluid, domain-general coordination, aided by communication technologies, beginning with language, that enabled *Homo sapiens* to rise to its current position at the ecological apex. Furthermore, by utilising a cultural evolutionary account to support the mindshaping hypothesis, a key feature of mindshaping is brought to light, namely the central role played by model choice. If mindshaping is to work to solve

coordination dilemmas dynamically and efficiently, then targeted model choice must play a central role in choosing who and what to shape and be shaped by.

Taken together, these claims make up a persuasive story about human social cognition, coordination and ultimately the constitution of selves and groups. In addition to being a persuasive story, these claims are also consilient with recent research in cognitive science and can be seen as drawing together and giving a name to a nascent trajectory in the special sciences that rehabilitates the importance of the social realm in the explanation of human cognition (Ross 2022). However, despite the plausibility of this reconfiguration, little work has been carried out that utilises this new paradigm to engage in applied normative research. This is problematic because if this story is on the right track then it has clear ramifications for the human social world. As we have seen, coordination dynamics can generate scenarios in which the rational action, indeed often the only action that seems possible for the agents involved, is to destroy another group (Hardin 1997). Less dramatically, though more importantly, the outcomes of coordination dilemmas fundamentally determine the form of the quotidian social worlds we all inhabit, as individuals and groups. Coordination dilemmas lie at the root of social coalitions, determining the distribution of resources and the equilibrium a society comes to accept as fair. Because the logic of coordination means that to be part of a coordinated group trumps uncoordinated interaction, the equilibria reached are typically optimal for all participants, even if it does not maximise potential individual welfare. It is better to be a slave in a prosperous society than to be a wealthy person cast out into the desert.

Thus, the effective channelling of our collective productive forces to maximise social welfare depends on harnessing or guiding coordination. Where

coordination power is hijacked or stymied by actors unconcerned with broad-based social welfare gains, or where the structure of communication misdirects coordination power or generates noise that saps its potential, the paths to achieving beneficial outcomes are blocked. So in the interests of bringing about a more just world it is critical to understand how large-scale coordination dilemmas are resolved in the contemporary context, in particular given the recent upheavals in our core communication technologies, in the hope that such knowledge may equip us to better mitigate the destructive, sub-optimal or counter-productive tendencies to which the logic of coordination may give rise.

In what follows, I deploy the theoretical lens developed so far to examine online communication technologies. Specifically, my claim is that due to a series of problematic design features, the online realm interacts with human social learning and coordination processes in a novel and pernicious manner. At their most problematic, these features incentivise the use of signalling strategies that generate increasingly extreme ideologies, thereby making zero-sum conflicts more frequent. The manner in which these processes affect political institutions – creating deadlock and incommensurable demands, resulting in political misdirection and stasis – ultimately saps overall coordination power, makes such power vulnerable to manipulation and hijacking by selfish actors and has generated a highly connected yet paradoxically disaggregated public sphere. This state of affairs undermines the social welfare of all.

The model being proposed here, which builds on the concept of mindshaping to examine how social cognition interacts with communication technologies, is for now only a well-informed hypothesis. My central claim, that the role coordination plays in structuring our ontologies should be taken into consideration when we

assess communication technologies, must be further strengthened by direct empirical support. Such work is beyond my scope here, however, the theoretical framework I have developed can be used to suggest what such work could look like. Thus, I close this chapter with a set of methodological considerations that apply to the task of operationalising my central claims. In addition to providing support to my central claims here, these guidelines can act as a blueprint for work that aims to refine the mindshaping-as-lynchpin hypothesis in general.

Before we can examine digital communication technology in more detail and set out ways in which it may interact with coordination dynamics and belief formation, I will first describe how coordination interacts with our epistemological faculties in general. In particular, it has become increasingly clear that for most people in most circumstances knowledge of the world is pervasively structured not by the aim of discovering truth, but rather by the importance of signalling group membership status.

The chapter proceeds as follows: In Section 4.1 I examine the social turn in epistemology and the role played by signalling in belief formation. In Section 4.2 I examine the history of communication technologies and symbolic belief formation in human cultural evolution, showing how group size expansion relies on our ability to better transmit and centralise symbolic knowledge and thereby scale up mindshaping processes. Section 4.3 then examines digital communication technologies, laying out a series of potentially problematic areas where they may interact with the use of symbolic beliefs to signal coordination potential. Finally, Section 4.4 lays the groundwork for an empirical investigation of mindshaping processes in the digital domain.

4.1 Epistemology for Coordinators

Epistemology, or the study of how humans come to have knowledge of the world, has traditionally been conducted on largely individualistic terms. The early modern philosophers, thinkers like Descartes and Locke, placed the individual knower and their sense faculties at the root of all true knowledge of the world, and this general perspective has, up until surprisingly recently, determined the limits of epistemology. However, in keeping with the general intellectual trend that underlies the emergence of the mindshaping paradigm, the social realm, and the place of the knower in a social configuration, has increasingly begun to feature in epistemological enquiry. This sub-field has come to be known as social epistemology (Goldman & O'Connor 2021). Given the well-evidenced role of social learning in cumulative cultural evolution, and, on the mindshaping paradigm, in the literal construction of the mind, this broadening of perspectives in epistemology is long overdue.

Accordingly, an agent's knowledge of the world is no longer seen as solely the product of atomistic individual enquiry, a paradigm that also motivates the mindreading-first conception of human social cognition. Instead, to fully explain belief formation, we need to conceptualise individuals as components in social networks which structure the beliefs they hold. Social epistemologists draw our attention to a range of features that pertain to knowledge acquisition that were overlooked in traditional accounts. Central topics examined include the role of testimony in knowledge production (Coady 1992), how peer disagreements affect knowledge claims (Christensen 2007) and how social structures and the totality of concepts they accept can perpetuate epistemic injustices (Fricker 2007; Haslanger

2019). Each of these concerns turns on the fact that an individual's knowledge is influenced and constrained by that of others in the social realm.

Another, more empirically minded, approach developed in social epistemology uses agent-based models of social interaction networks to isolate and examine how specific features they exhibit may affect epistemic outcomes. Inspired by a modelling approach first developed in economics known as network epistemology (Bala & Goyal 1998) this work has been spearheaded by philosopher of science Kevin Zollmann (2007, 2013). The models that Zollman has developed demonstrate that the form of communication networks utilised by scientific communities play a significant role in enabling or hindering true belief formation. Building on Zollmann's approach, Emily Sullivan and colleagues (2020) apply these modelling resources to real-world social media interactions to show which specific communication structures can deliver or hinder the generation of wisdom-of-the-crowd derived epistemic benefits.

Fundamentally, the key insight prompted by this turn to the social is that understanding which beliefs are adopted requires paying close attention to how, which and in what contexts people interact. This shift in epistemology dovetails with the claims made by proponents of cultural evolution theory about social learning just outlined. And, as we will see, it aligns with the core idea driving this thesis, namely that people do not mainly have settled, fully distinct beliefs that others discover within their minds in advance of social interaction. Instead in social epistemic exchange individuals participate in ongoing dynamic interpersonal belief negotiations with the primary aim of enabling coordination. Indeed, on the mindshaping paradigm epistemology is necessarily a social affair, and though mindshaping is not referred to in the social epistemology literature (though Sally

Haslanger's (2019) work is a notable exception), it makes a good fit with the social turn in epistemology.

Fundamentally, as we have seen, social interaction and belief adoption, in general, is *strategic*. People are not gullible dupes who believe everything they are told, or copy everything they see, or exist as components in deterministic networks of social relations (Mercier 2020). The large and growing literature on social learning as strategic confirms this (Kendal et al. 2018). As we saw in Chapter 3, a range of complex strategies govern human social learning, indeed humans seem to socially learn to socially learn better (Heyes 2016).

However, "strategic" in this sense does not refer to the conscious strategizing of a Machiavellian mindreader, bent on figuring out the intentions and desires of conspecifics in order to further their own self-interest. Instead, the invocation of strategy here refers to non-random actions. People's choice of social learning models *is* guided by some sort of strategy, but it is *not* a strategy represented in the consciously accessible cognitive system of the learner, at least not typically[19]. Social learning is usually linked to the optimisation of some goal, but there is no requirement that this goal be the *accurate* representation of the environment. As the pragmatists have long reiterated, knowledge is for doing, for coping with our environment, not necessarily representing it faithfully. Thus the key question in any bout of social learning is: what does the strategy deployed aim to optimise? What aspects of our world is it helping participants cope better with? At the core, how is it enhancing individual or group-level fitness? In answering these questions we need to

[19] As we have seen, Heyes (2016) proposes that some social learning strategies may be metacognitive and as such are perhaps consciously accessible and modifiable.

pay close attention to the differing learning contexts, and the types of beliefs being propagated in social learning.

For example, as we saw in the previous chapter in non-opaque causal contexts it seems likely that strategic learning aims to optimise efficiency. So, for example, in cases of flint knapping, shelter building or spear throwing it seems straightforward that social learning was used by our ancestors to optimise the effectiveness of these activities given a specific goal, i.e. sharp hand axes, watertight roofs and accurate, long-distance throws. If I see that a certain strike angle makes a more precise knap then I am likely to adopt this into my practice of knapping, regardless of social conventions or the relative prestige status of the model. In such contexts, characterised by functional knowledge, social learning *is* often truth directed.

In other contexts with greater causal opacity, but nevertheless involving specific ends-directed action that could be optimised, it is harder to explain how social learning is guided. For example, as we saw, Henrich (2016) cites manioc processing as a process that could be made more functionally efficient if specific, causally opaque, steps were discarded, but would result in the end product being mildly poisonous. In such cases, the optimal strategy is to copy one's conspecifics closely and faithfully imitate the apparently useless steps, but it is unclear what cues this strategy or keeps it in use. However, worries about such examples of rote conformist copying are perhaps overblown as they are, at best, fringe cases (Sterelny 2017a). Usually, important/survival-relevant forms of culturally transmitted knowledge are procedural (particularly in less complex contexts that characterised early evolutionary environments) and have relatively transparent causal profiles

befitting straightforward efficiency-directed social learning as opposed to rote-conformism.[20]

However, in addition to more causally straightforward learning, since the advent of symbolic culture a large domain of human affairs consists of beliefs that have tangible effects on behaviour, but don't have clear efficiency outcomes or involve ends-directed procedural knowledge. For example, cultural artefacts like styles of dress or body ornamentation, relationships to the spirit realm, etiquette, the realm of social norms in general or beliefs about distant outside groups. The ways in which these cultural elements alter behaviour are not obviously more or less efficient and the knowledge they spread and the beliefs they authorise are not clearly linked to instrumental goals. Thus, their causal implications are not transparent, and in some cases, for example, folk cosmological beliefs, they have no tangible effects at all. Beyond brute cultural attraction as proposed by the Paris school, which likely only explains a small fraction of these sorts of non-instrumental beliefs (Sterelny 2017b), what could be the adaptive role of social learning in this domain?

A recent explanation that aligns well with the mindshaping paradigm suggests that many of these symbolic beliefs may serve a signalling function, acting as hard-to-fake and thus honest signals which can reliably indicate group membership and status (Williams 2022a; Funkhouser 2020; Jagiello et al. 2022). As societies became more complex and grew in size, the ability to robustly and honestly signal group membership and social standing became increasingly important for enabling the efficient division of labour and avoiding conflict with other group members. By adopting the beliefs (and behaviours) of certain conspecifics,

[20] Indeed in the case of manioc processing Mercier and Morin (2019) report (based on personal communication with the field researcher who first described the manioc process) that the apparently causally opaque steps, central to Henrich's account, both improve the taste of manioc and are less labour intensive than his account suggests.

individuals signal group membership and thus coordination suitability (Funkhouser 2020; Joshi 2022). Individuals stand to accrue large payoffs by signalling their status in a group as a good coordination partner. In this domain the open vigilance that Mercier (2020) claims characterises human epistemic faculties is not necessarily vigilance about the *veracity* of claims, but rather about their importance or unimportance in the group ontology.

However, the variety and quantity of such symbolic beliefs preclude conscious learning. Thus, rather than explicit, efficiency-directed social learning, it is likely that conformist mindshaping mechanisms of the sort examined in Chapter 2 play a key role in the propagation of such beliefs. As we saw, individuals automatically and pervasively conform their ontologies with interaction partners. Indeed, this type of conformism, the desire for belief consonance, is inherently enjoyable for humans – group camaraderie is a powerful force in social life (Golman et al. 2016).

From a fitness perspective it pays to conform to locally prevailing beliefs if this enhances one's status within a group or ensures one will be included in coordinated actions going forward. But this is *conformism as a signal*, not for conformity's sake. What this means is that this quasi-conformism is in fact quite labile, while appearing like rote conformism. As Gergely and Csibra (2003) have shown, from infancy humans are sensitive to intentions in deciding what actions to copy. In other words, overimitation is not rote, blind, high-fidelity copying, rather it is strategic and linked to perceived salience. We acquire most of our norms not via explicit instruction, but assimilate them automatically from our group, using subtle cues, in order to fit in better. These types of automatic processes are hard to counteract or modify but that is not to say their products are static or determined in

advance. Group membership requires the dynamic and ongoing renegotiation and adoption of updated symbolic signalling resources through time.

Thus natural selection has hardwired a set of social learning strategies into human brains that push people to strategically conform symbolic ontologies with cultural confederates as a means of indicating group membership. Such conformism is not conscious, but neither is it rote or blind. It is strategic but the agents engaging in it are largely uncomprehending of its regulative aim (Dennett 2017). Such strategy pays off on two levels: It is individually optimal to be seen as an upstanding group member, one who should be included in coordinated ventures; but, crucially, from the group-level perspective it is also beneficial if the individuals are effective coordinators, as it promotes ecological success. Together these benefits likely generated the selection pressures for mindshaping and the automatic use of symbolic beliefs as signals it enables. However, though coordination generates a net benefit for the group and individual, it is not necessarily equitable, and as we saw in Chapter 1, coordination can be individually sub-optimal but overall to be included in coordinated activity is usually better than to coordinate with no one.

As a result, in many cases, beliefs about symbolic elements of culture are linked to enabling efficient coordination within a group, not to truth or (within reason) even efficiency. This leads to the emergence of regional cultures that utilise cultural packages that combine high signalling power with relatively low instrumental efficiency. For example, strictly observed religious obligations may significantly reduce productivity but increase in-group cohesion. The upshot of using symbolic beliefs as signals is that closely interacting groups of humans universally develop and adopt various shared, restricted sets of beliefs about the world, which vary widely across geographically distant groups. Signals that are more difficult or

costly to fake, such as those related to appearance, accents or physical behaviours, or signals that unambiguously alienate other groups, are more effective as their costs make them more likely to be trustworthy (Sterelny 2012).

It is the processes of ongoing dynamic mindshaping that allows humans to regulate themselves and one another to adopt similar values and beliefs, and because these symbols pervasively structure and constrain interaction, they become real though virtual components of ontology. Interpersonal regulation of this sort is a key factor in making humans hard to influence in the moment. As Mercier (2020) emphasises and empirically demonstrates, it is hard to convince individuals to adopt new beliefs, as humans are wary learners. Indeed, as Morin et al. (2021) show in a metastudy of social learning tasks, people in experimental settings tend to discard most social knowledge even when such a strategy is sub-optimal.

These experiments, however, only test conscious efforts to gather efficient knowledge. Much of our ontology is informed by the passive absorption of implicit attitudes towards the world from close cultural group members, in particular when we are children, though this continues throughout life. The absorption is driven not by rational belief updating in the pursuit of true knowledge but is rather primarily responsive to group attitudes. The famous Asch line choosing experiments (1955), and their replications (Germar & Mojzisc 2019), described in Chapter 2, show how group influence can modify beliefs and even perceptions. In these experiments, participants are motivated primarily to coordinate, at the expense of objective factual information.

Thus, though useful in describing a specific context, the conception of social learners utilised by Mercier, Morin and colleagues (Mercier 2020; Morin et al. 2021; Trouche 2018) – i.e. learners as perpetually wary and in a sense "scientific" in their

acquisition of knowledge – is problematic, as it again runs the risk of the over intellectualising social learning (Sterelny 2017b). The impetus to do so grows, as I described in Chapter 3, out of the mindreading-first presuppositions that inform almost all work that deals with social cognition and cultural evolution. Putting people in isolated experimental contexts and having them attempt to refine the design of some object, or discern the intentions of other agents, and then seeing if they choose to use social information or not presupposes, incorrectly, that the way people interact in general is informed by conscious, active, strategizing.

The discounting of explicit social information can be better accounted for on the mindshaping paradigm: rigidity in belief sets is crucial for coordination through time, chaos would ensue if humans copied everything they encountered from their associates. In most cases of interest in which social information spreads there is little conscious strategizing being done, rather, mindshaping processes pervasively unfold, guiding information acquisition as people act naturally in the social world to keep one another converged around ontologies consisting of largely symbolic beliefs. Processes like these are better captured by Asch-type experiments than the sort typically used to assess cultural information propagation. Thus, viewed with the group conformity experiments in mind, we can reinterpret the conclusions of Morin et al. (2021) about the underuse of social information as instead showing that individual ontologies are not influenced mainly by piecemeal exposure to information or factual claims, but are rather primarily responsive to broader, ongoing group coordination goals.

Thus, the difficulty of altering, or incentivising the modification of, beliefs piecemeal emerges not only because people are wary learners, but also from the fact that belief sets are linked to, and embedded within, larger group ontologies. To

convince a believer in QAnon that the claims made by Q are spurious requires not individual case-by-case debunking, but for the affected agents to either lose interest in sustaining the ontology, perhaps when an alternative conspiracy captures their interest, or for the interpersonal networks of communication that sustain the claims to be shut down. In other words, attempts to modify the webs of belief that characterise specific cultural groups either require large shifts in incentives, for example making communication about QAnon costly, or for the group itself from within to organically shift the contours of their ontology.

Historically large shifts in incentives that forced entire groups to discard ontologies were usually the result of the conquest of one group by another. The contemporary geographical distribution of religions acts as a rough map of historic conquests of this sort and demonstrates how the linking of an imposed belief system to social success, coupled with the active repression of existing beliefs amounts to a powerful change in incentive structures that filters down to reform core group ontology. This is not to say a project aimed at changing individually problematic group beliefs from the outside (without using force) is futile, but rather that the means and methods required are not rational, individual-level persuasion, but the modification of the incentive environment surrounding specific belief sets, and changes to how group members interact, both within groups and with other groups.

Of course, believers in QAnon are not only conspiracists, they are also members of other communities, many of which likely repudiate the ontology of QAnon. Most people inhabit multiple overlapping cultural groups – economic, religious, racial, sexual, national, professional – each of which has a distinct reference network of beliefs attached to it. It is the managing of this proliferation of identities, a state of affairs that is a corollary of the emergence of complex society,

that the narrative construction of self, described in Chapter 2, mainly functions to enable. Humans must keep tabs on multiple intertwined and sometimes diverging narratives about their selves relative to different reference networks in their social environment. Membership in each group is linked to incentives, for example in a public social context an upstanding Texan might implicitly affirm, by not disavowing, claims about natural racial or gender hierarchies to signal group affiliation and reap social benefits while repudiating such claims in other contexts (though of course they may not, this will depend on the extent to which their various cultural reference groups are aligned). The implication here is not that agents are cutthroat social climbers, willing to go along with anything to get ahead, but rather that different social contexts place subtle pressure on beliefs affirmed and indeed believed. The mindshaping conception of interpersonal regulative processes shows how these pressures influence belief sets without conscious strategizing on the part of the agents involved.

However, shifting material fortunes or the emergence of new norm sets (though in reality the two are often linked (Gelfand & Lun 2019)) can also strain group affiliations by altering incentive structures, even in the face of coherence and coordination generating mindshaping processes. It is here that more explicit strategic considerations enter the frame. A. O. Hirschman's (1970) exit, voice and loyalty framework is a useful characterisation of the potential strategic options available to cultural group members confronted with changed incentives. The core of Hirschman's proposal is that if an agent grows dissatisfied with their primary cultural group they have several distinct strategic options: they can voice their dissatisfaction in the hope of altering the normative reference network of the group, choose to exit the group and join another, or keep quiet and accept the changed

structure. As we will see in the next section the effective management of these potential strategic moves to stop groups with less equitable resource distributions from fracturing by effectively altering the strategic landscape to make voice or exit costly was crucial to the expansion of groups throughout cultural evolution.

Fundamentally the framework being described proposes that agents placed in close proximity tend to automatically homogenise their belief sets in order to both enhance coordination ability and better signal their position within social groupings. The extensive and universal use of symbolic beliefs by human groups is one outcome of these processes, as symbolic beliefs, which are unmoored from efficiency or survival constraints, can more effectively and flexibly act as signals. An important upshot of such a dynamic is that the number, fidelity and efficiency of connections between group members determines the levels of homogeneity displayed concerning belief sets and the potential size of such groups. For example, in hunter-gatherer settings loosely linked sub-tribes in larger regional groupings likely had some shared overarching belief sets, which would diverge somewhat within individual tribes and radically with unlinked distant tribes. This divergence was important because it meant that welfare/fitness-destroying variants of beliefs in our early prehistory remained localised. Thus, local homogeneity coexisted with global diversity.

However, this relationship is not a constant, as the emergence of ever larger and more homogenised social groupings throughout history demonstrates. For groups to grow in size they must share networks of beliefs across larger scales, a process which has been driven by advances in communication technologies. Indeed, as we will see, the history of humanity and the social groupings it has supported, is tightly linked to changes in the scope, fidelity and efficiency of communication

technologies and the centralised symbolic signalling resources they enable. Advances in communication technologies systematically underpin and accompany changes in human group sizes and thus the potential coordination power such groups can generate, for better or worse.

4.2 Communication Technologies and Group Signalling

The developmental trajectory of human culture, beginning with the stabilisation of behaviourally modern traits, is one in which average coordinating group sizes have trended upwards. In explaining this many factors must be taken into account – the emergence of complex social life cannot be reduced to one single breakthrough, genetic or cultural – however, one necessary component of this expansion is the alignment of group ontologies across increasingly large numbers of unrelated individuals. In other words, though mindshaping processes likely emerged early in the developmental trajectory of hominins (Zawidzki 2013), what needs to be explained is how these processes could scale beyond groups of agents living in close proximity.

One crucial element required for such an expansion is the ability to advertise group membership to other members with whom one is personally unacquainted. This suggests that there should be a relationship between changes in information communication and recording technologies and group composition (Floridi 2014). Prima facie, evidence for the existence of such a relationship is provided by the observation that the major milestones on the journey towards ecological dominance roughly overlap with the development of increasingly powerful and high-fidelity means of communication. In what follows, I want to build on this observation while setting aside issues of causality. There is no final answer to a question such as: did

group sizes increase due to the development of writing or did increased group sizes lead to the development of writing? The causality runs in both directions – throughout human history increases in group sizes and civilisational complexity have been accompanied by evolutions in the fidelity, efficiency and scope of communication technologies. Changes in communication technologies appear to be necessary enablers of increases in group sizes, though of course they are not sufficient causes of it.

4.2.1 The First Communication Technologies

The first emerging dedicated communication technology is spoken language, which is thought to have slowly emerged between 3.5 and 1.7 million years ago. By the time of the transition from Oldowan to Acheulean stone tool manufacture around 1.7 million years ago our ancestors likely had relatively advanced linguistic capabilities, as the complexity of the hand axes that characterise this culture suggests that advanced pedagogy would have been required to reliably instruct initiates and preserve the techniques across generations (Morgan et al. 2015). Additionally, the fine motor skills required for their production utilise brain areas that overlap with those for complex language use (Stout & Chaminade 2012). Without advanced linguistic capabilities none of the various features that so concretely set *Homo sapiens* apart from other species would have developed. Though it is as of yet an open question as to whether humans are the sole possessors of language (Ross 2019), it is undoubted that we are the species that most effectively leverages our language faculties to engage in pervasive niche construction. Humans use external records of language to generate robust signposts that enable further niche construction across generations (Sterelny 2012). This ability lies at the root of our complex cognitive

faculties which are themselves products of language supported cultural evolution (Heyes 2017). Indeed, according to the mindshaping paradigm, without language there is no rich conscious mind to speak of at all, and thus no individual human selves as we know them today (Dennett 1991a). Once our early hominin ancestors had linguistic capabilities then mindshaping processes were also operative, and these agents were already more than biological individuals, possessing selves of some sort.

Language is thus the starting point for the emergence of complex human culture and cognition and is the medium that all subsequent advances in communication technology are fundamentally, though not solely, designed to transmit. Language lies at the root of complex coordination (Zawidzki 2013), and once it had emerged it allowed for the process of cumulative culture to begin and hominins to expand outside Africa circa 2.1 million years ago (Lordkipanidze 2017). Language would have allowed early hominin bands to efficiently divide labour and coordinate actions, and to develop and adopt tightly aligned ontologies that made this coordination largely automatic. The importance of pedagogy to the preservation of their cultural toolkit also suggests that these agents were using finely tuned social learning strategies, though the small group sizes would have made the efficient direction of attention relatively straightforward. Yet, despite the relatively early emergence of language, for the majority of the history of the *Homo* genus the evidence suggests that group sizes remained small, and the pace of cultural evolution glacial. Hominid social formations likely comprised daily interaction bands of 30-50 individuals, which may have been integrated into alliances or clans made up of around 150 members, constraints persuasively linked to neocortex size by Robin Dunbar (1993).

This general social structure persisted up until what has come to be known as the Upper Paleolithic Revolution beginning in earnest around 38,000 BCE (Bar-Yosef 2007; d'Errico 2007).[21] This period saw a series of cultural innovations which allowed for the development of larger and more complex group formations and the emergence of inter-tribal trade. A key and pervasive component of this new cultural package was the extensive use of symbolic markers, for example the use of animal bones to create ritual objects and the systematic usage of beads, pendants and ochre-based colourants as body decorations (Bar-Yosef 2007). Symbolic objects such as these fall into the category, described in the previous section, of signals of coordination potential that functionally dominate technical efficiency considerations. Body decoration and modification act as a signal to other agents indicating tribal allegiances, while the content of those signals themselves are arbitrary, bestowing no direct efficiency gains.

These early signalling devices effectively acted as low bandwidth, localised communication technologies, advertising tribal affiliations and generating common knowledge about a person's social standing. By recording, advertising and reinforcing established coordination equilibria these markers allowed interacting agents to specify their preference sets in advance of interaction, thereby establishing common priors, allowing for the efficient resolution of increasingly complex and frequent coordination dilemmas (Ross & Stirling 2021). Additionally, the explicit

[21] This date marks the point at which the archaeological evidence of symbolic behaviour becomes consistent and widely accepted. However, the discovery of intentional symbolic markings from 75,000 BCE in the Blombos caves in South Africa (Henshilwood et al. 2018), and recent evidence of long distance exchange of ochres circa 300,000 BCE (Brooks et al. 2018), suggest that elements of behavioural modernity in fact emerged much earlier than 38,000 BCE. These discoveries suggest that the origin of symbolic culture is likely to be pushed further into the past as more evidence is unearthed and verified. Additionally, this suggests that the use of revolution as a metaphor in this context is somewhat misleading. In reality, the shift to cultural modernity was a gradual, protracted process likely advancing and retreating as climactic and other environmental features influenced the fortunes of Paleolithic groups.

signalling of tribal affiliation was likely incentivised by the development of inter-tribal trade during this period (Ofek 2001; Ibáñez et al. 2015). Visible markers of tribal allegiance were needed to distinguish unfamiliar agents as friendly insiders, trading partners or potential enemies. Given what we now know about the role of symbolic signalling in enabling coordination (Funkhauser 2022; Williams 2022a), it is likely that the emergence of such communication devices was a crucial first step towards the development of larger cultural groups composed of agents who did not share kin or reside in close proximity. The use of symbolic signals to efficiently identify group members, coupled with the ontology conforming, common knowledge generating and strategy aligning effects of ritual practices, thus likely played a key role in the gradual increase in group sizes during the Upper Paleolithic (Powell et al. 2009).

This slow upwards trend eventually led to the emergence of the first small-scale urban civilisations beginning around 9,500 BCE in Mesopotamia (Hodder 2007; Clare 2020). The more favourable climate of the Holocene, coupled perhaps with wild resource depletion, is thought to have spurred the development of agriculture and permanent settlements (McMahon 2020; Gupta 2004). The archaeological record clearly demonstrates that these early urban cultures utilised extensive symbolic cultural elements, including temples, burial rituals and figurative art forms (Hodder 2007; Clare 2020). Additionally, these settlements engaged in extensive trade with nearby groups, likely leading to a degree of homogenisation in lifeways across larger territories as groups began to specialise in specific forms of production and exchange those products, thereby standardising objects and foodstuffs in localities (Ibáñez et al. 2015). The transition to agricultural lifeways also observed during this period required the ability to coordinate large numbers of

agents across larger territories. It is likely that the extensive symbolic components of these cultures facilitated the coordination and group cohesion required for the transition to sedentary lifeways. Thus, symbolic culture can be seen as a physical record of the expansion of mindshaping across larger territories. The relatively close occurrence of the first widespread evidence of symbolic cultural artefacts and the development of agricultural societies suggests that symbolic markers and beliefs are required to enable more sophisticated and complex social contracts, likely by enabling the recording, dissemination and reinforcement of established resource distribution equilibria.

The next significant development in the history of communication technologies is the advent of writing. The earliest evidence we have of script comes from the Sumer region in Mesopotamia c.3300 BCE, in the same general area of first sedentary settlements and villages, though writing systems are now understood to have evolved independently across at least four geographically isolated domains, each of which was experiencing unprecedented population densities for the time (Gnanadesikan 2008). This example of convergent evolution again demonstrates that advances in communication technologies are closely linked to societal growth. Across the archaeological record written script emerged almost simultaneously with the first proto-states, in ancient Mesopotamia, Egypt, Mesoamerica and China (Storey 2009; McMahon 2020).

In order for a population to expand beyond a certain size, it requires the ability to record and disseminate information for administrative and economic purposes, thereby enabling the productivity increases required to sustain a larger population. But the surviving evidence we have of early written script also shows that it was used to codify law and social hierarchies and to centralise symbolic

beliefs (Gnanadesikan 2008). This centralisation and standardisation allowed larger groups of agents personally unknown to one another to draw on established resources to reliably signal their group membership and participation in a common ontology – thus writing allowed for the first centralised high-fidelity mindshaping processes to commence. Further evidence for the importance of religion and elite hierarchy in early urban life is provided by the existence of elaborate temple and palace institutions (Seymour 2011). Visible religious adherence operates as a powerful and often costly signalling device and as such institutional religion is now understood to be a key component in the emergence of large-scale societies, acting to solve coordination problems, and encourage in-group pro-social behaviour (Vlerick 2020; Atran & Henrich 2010).

The development of writing played a central role in institutionalising and disseminating such belief systems, and many of the earliest surviving texts are religious in nature, such as the Egyptian Pyramid Texts (2400 BCE), the Gilgamesh (2100 BCE) and the Rigveda (1500 BCE) (Gnanadesikan 2008). Durable script allowed for standardisation and dissemination, across increasingly large geographic scales, of aligned symbolic beliefs that served to signal coordination potential, mindshaping agents to share relatively homogenised group ontologies. The role of religious orders in controlling access to writing and their close association with political elites served to restrict and centralise control over these symbolic ontologies, as a result concentrating coordination power in elite hands (Gnanadesikan 2008). Once established, political formations of this sort, centred around a ruling monarchy and a priestly caste who were the sole interpreters of divine scripture that doubled up as social law, persisted for the next several thousand years across the majority of large human population centres, and still control some

states to this day. Part of the explanation for this longevity is that precarious ecological conditions make centralised and tightly regulated symbolic packages of norms optimal as they enhance coordination in the face of threats (Gelfand & Lun 2019). This demonstrates the relationship between the specific governance institutions possible and the communication technologies available (Turchin et al. 2018).

This centralisation of symbolic authority and control of ritual practices, and thus of key common knowledge generation devices (Chwe 2001), has important implications for the viability of voice, exit and loyalty as strategic possibilities, as described in the previous section. Prior to the emergence of sedentary human settlements hunter-gatherer groups were radically egalitarian with low levels of hierarchy or resource inequality (Marlowe 2005).[22] One factor explaining this state of affairs likely relates to the potential viability of exit as a strategy for resolving interpersonal conflict. If one did not like the norms developing in one's group it was relatively simple to fracture off, given that survival skillsets were simple enough to be known to all adults and there was an abundance of unsettled territory. Additionally, voice, whereby individuals speak up to try and modify group norms, was also a viable strategy in the small face-to-face context of early hunter-gatherer lifeways. This meant that problematic or sub-optimal norms could either be mitigated by the use of voice or alternatively groups could quickly dissolve. As a result, groups that did prosper were relatively egalitarian.

However, once human lifeways became more sedentary the viability of exit and the power of voice was significantly reduced. For one, the changed material

[22] Recently this widely accepted claim has been problematised. Graeber and Wengrow (2021) provide extensive evidence that hunter-gatherer societies were more complex and fluid than previously thought – sometimes egalitarian and sometimes hierarchical in cycles linked to seasonal variation.

basis of existence made exit more difficult. Agriculture requires more complex skills and a robust division of labour and suitable arable land. Thus to successfully found a new settlement required more knowledge, skills and manpower than it did for hunter-gatherers. Nevertheless, groups could still in theory fracture as norms became intolerably unjust if enough agents could agree to simultaneously depart. However, a central strategic concern that arises in generating group-level action is that of creating common knowledge of discontent, of uniting private grumblings with public action (Kuran 1997; Chwe 2001). Once ritual practice and symbolic signalling technology became centralised, the common knowledge generation devices required to deploy exit, or to use collective voice to shift norms, were increasingly monopolised by religious and military elites. The shift in communication technologies enabled by the emergence of writing, whereby large groups came to be orientated around centralised administrative and spiritual hubs, thus had the function of making exit and voice more difficult to deploy as a strategy for resolving the emergence of unjust norms (Ross 1988). As a result, larger but more hierarchical and unequal cultural groups could emerge, orientated around palace and temple institutions that controlled common knowledge generating technologies. This marked the emergence of large-scale coordination power as a resource wielded by elites, a resource that was created by advances in communication technologies, technologies that closely interact with how that power is dispersed and controlled.

This brief overview of the links between the emergence of symbolic culture, the development of writing and the growth of human groups has served to demonstrate two points: 1. New communication technologies enable human communities to scale up in size, partly via the dissemination of shared scripts of action and symbolic resources which allow for mindshaping to operate across larger,

impersonal dimensions. 2. Shifts in communication technologies alter the strategic options available to agents and thus the distribution of power in a society.

4.2.2 Modern Communication Technologies

The next significant development in communication technology which signals the beginning of the modern era as we understand it was the invention of the printing press around 1440. The communication forms it made possible by printing fundamentally modified the political constitution of human groups and the scale and distribution of coordination power. The printing press had the effect of further homogenising and aligning human ontologies across increasingly large geographical domains. As Elisabeth Eisenstein emphasises in *The Printing Press as an Agent of Change* (1980) one crucial effect of the printing press was to make written sources identical, thereby placing everyone on the same page, so to speak, compared to older handwritten scripts which had to be laboriously and often erroneously transcribed. One effect of this standardisation of knowledge was to allow for the sharing and comparison of observations of the natural world, a key factor in the rapid progress in the natural sciences during the 15th and 16th centuries.

In addition to spurring the development of the natural sciences, the printing press also played a central role in the emergence of democratic nation-states as distinct political entities in the 18th and 19th centuries. At their core nation states are what Benedict Anderson (1983) calls imagined communities. These communities are not found, but made, and Anderson proposed that the development of printing directly contributed to this genesis. He describes how once the markets for books printed in the religious languages, primarily the preserve of social elites, were saturated, the printing presses began to produce books in the vernacular languages in

search of further profits. This inadvertently generated a shared discourse across a once disparate population now united by the use of a common language. Where once an inhabitant of north-west France was from Brittany, or “around here”, now they were made aware of their existence in a bounded geographic space known as France, sharing a common linguistic heritage. The agents engaged in this new discourse came to reconceptualise themselves as components of a distinct community rooted in a concrete geographical location, with distinct customs and mannerisms, a conception that replaced prior diffuse religious conceptions of belonging. The formation of a national discourse created a new political coalition, one made up of a multitude of agents individually unknown to one another but united around a symbol, that of the nation, and provided with a range of signalling resources with which to make those communities real and reinforce their connectedness, and advertise their coordination potential. The accessibility of printing, and the removal of religious-based control over the creation and dissemination of knowledge, also had the effect of restoring voice as a crucial and powerful strategy in the modification of group norms, and the challenging of elite power.

The shift in self-conceptualisations and redistribution of communication power enabled by the printing press allowed the nation-state to become the preeminent coordination vector in the modern world. Nation building is the most powerful means of amassing and stabilising large-scale coordination power and centralising mindshaping processes yet devised, power that is both generative and destructive – disputes over state boundaries lie at the root of many of the most brutal conflicts in recent history. The role of the printing press in providing a set of symbols and signals that individuals in a nation could share shows how the genesis of these entities is directly related to advances in communication technology, just as

the original cities that emerged in early history were enabled by advances in writing technology. In both cases changes to how information is disseminated shifted the balance of power in the existing society.

In the case of early symbolic cultures centred around palace and temple complexes, coordination power was controlled by a combination of religious and political elites who minimised the power of voice and exit as potential social contract altering strategies. The printing press shifted this equilibrium by making access to information and the means of dissemination more egalitarian, despite various moves by the Catholic church and extant monarchies to regulate the nascent print industry. As a result, voice was reanimated as a powerful strategic move with which agents and groups could challenge the power of established elites. This set the stage for the decline of monarchy and religious rule and the emergence of a new ruling bloc made up of an alliance between economic elites and a new form of nominally representative political elites (Mann 1986).

As nation-states and their associated institutions became the dominant political forms they too sought to exert control over the various means of mass communication, first restricting the print industry, and then the later electronic forms of communication such as radio, telephone and television. These technologies were initially closely regulated by most states, which controlled broadcasting licences, and in many cases funded state broadcasting agencies. Indeed for most of the modern era state regulation of mass communication was the norm. Furthermore, the professionalisation of journalism and publishing, coupled with hierarchical gatekeeper structures, and the existence of close links with political elites, ensured that mass media was anchored to a relatively coherent, centralised and primarily conservative ontology (Klaehn 2010).

By exerting (explicitly or implicitly) control over the means and content of communication nation states thus monopolised the coordination power unleashed by increasingly powerful forms of information communication technology, just as in the early proto-states religious authorities controlled access to the skills of reading and writing. Communication technologies thus both enable the generation of coordination power and also constrain and structure that power, how it can be deployed and controlled. Importantly, the manner in which this control is exercised is not by brainwashing citizens using propaganda. Controlling the means of communication does not provide states with the means to bend individual agents to fit whatever cognitive profile is conducive to maintaining power. Influencing group-level mindshaping is not equivalent to shaping minds to fit some desired profile.

Instead by controlling communication, authorities can limit the range of strategies available to agents. At core a political formation is a group of agents individually seeking to optimise their fortunes in a specific communication environment, using signalling to advertise coordination potential and mindshaping one another to adhere to the developed equilibria. In an authoritarian state system the ability to generate common knowledge of discontent is severely curtailed, which makes it more difficult for alternative coordination vectors to emerge or amass coordination power (Kuran 1997). Thus the form or structure of a communication environment interacts with the range of strategies available to agents.

For example, contrary to folk wisdom, the Nazi state garnered support not by brainwashing a nation of rabid antisemites into being, but by using a combination of generous social policy for racially non-suspect Germans (creating individual incentives to acquiesce), with common knowledge of the lack of exit as a potential strategy in a hostile geopolitical environment and of the potentially severe

repercussions of using voice (Aly 2007). Fundamentally they amassed and maintained coordination power by limiting the strategic options available to the agents who made up the Third Reich, not by using propaganda to brainwash its citizens. This understanding of the Nazi regime can explain the rapid "denazification" that occurred following the defeat of the Third Reich, once the incentive context changed, so too did the behaviour and beliefs of the populace. Indeed in most cases, propaganda acts primarily as a signal that modifies the coordination dynamics, not as a means to inculcate beliefs. By perpetuating state affirming narratives the authorities broadcast a strong signal of power and generate common knowledge: everyone knows that everyone else is aware of the power wielded by the state and aware of the limited space for dissent. The majority of individuals in China or North Korea may not believe the content of state propaganda but they do recognise that the pervasiveness of propaganda signals the power of the state and the futility of attempting to support alternate coordination vectors (Huang 2015).

In non-authoritarian states the outcomes of state influence on the media are more subtle, instead the media acts to propagate a specific narrative about the world, and, by making such a narrative prevalent, signals the minimal space for socially acceptable dissent, a process that has been referred to as the manufacturing of consent (Herman & Chomsky 1988). Of course, from the perspective of social welfare the centralising of narrative control, and the signalling of state power in order to minimise alternate coordination vectors, may in some contexts be beneficial. In general, ecological or material instability favours tight coordination norms, thereby empowering centralised authorities. Though the media environments of the mid-20th century were relatively socially conservative and minimised space for

dissent from the status quo, they also respected the norms of coherence and truth seeking as a goal of enquiry and generated a cohesive public sphere which likely contributed to general social welfare during *Les Trente Glorieuses* (Piketty 2014).

Furthermore, this set of circumstances was not accidental, media theory in the first half of the 20th century explicitly understood the power that mass communication could generate and the resulting importance of careful design and regulation of its affordances (Lippmann 1922). This effort was linked to the core tenets of liberal ideology, at least in the Western world, media was seen as a means of enhancing individual freedom of choice, and effort was made to represent relatively diverse viewpoints. Indeed the emergence and rise of liberal political philosophy was intimately tied up with the development of mass communication technologies. The freedom of expression enabled by the printing press in its earliest incarnation was key in enabling the development of liberalism, a form of governance that rejects the authority of scripture and monarch.

However, beginning in the 1970s, state control and guidance of mass communication technologies, extending from the advent of writing to the tight regulation of mass media in the 20th century, began to diminish, particularly in the Western world. In theory, the opening up of discourse to a wider variety of viewpoints provided citizens greater scope to challenge the status quo, thereby revitalising the power of voice as a strategic option in group formation. However, in practice, this new media landscape had the effect of undermining public discourse and entrenching, rather than challenging, the power of established elites. As Yochai Benkler (2020) examines in detail, these changes owe their origins to the neoliberal shift in political thought beginning in the 1970s. This new ideology promoted the mass deregulation of electronic broadcasting, first in America and then in Europe.

Beginning with the repeal of the fairness doctrine in 1987, which required broadcast media to report on public affairs and provide impartial treatment of alternative viewpoints, broadcasters in America began to explicitly cater to segmented audience profiles, and prioritised advertising revenue above other goals. This change had, by the late 1990s, generated a deeply fractured media landscape with different cultural groups often inhabiting separate informational spheres. This model has been replicated across the Western world, with the proliferation of highly segmented, niche interest media networks both on television and radio, often owned by transnational media corporations (Gershon 2013).

As will be examined in more detail in Chapter 5, by denigrating all forms of state regulation of industry as both inefficient and inherently biased, and placing ostensibly non-partisan market structures in their stead, neoliberalism enabled economic elites to engineer markets to produce outcomes that favour their interests, regardless of broader externalities. In the case of media this dynamic resulted in the fragmentation of the public sphere and the downgrading of truth or coherence as journalistic norms, in favour of maximising audience engagement to attract advertising revenue and an increase in private control of public discourse.

In a sense, the contemporary media landscape maximises the *perceived* availability of voice and exit as strategies, yet paradoxically disassociates these strategic options from meaningful political engagement. More accurately, the neoliberal media landscape generates simulacra of voice and exit, marketing a mirage of a dynamic public sphere engineered to maximise consumer engagement while generating political stasis and increasing inequality. These outcomes follow from the fact that communication technologies are not politically inert: controlling

the means of communication allows private entities to control the questions that enter the public domain, and the answers that are provided (Floridi 2015).

So far we have seen how the trajectory of human history has tended towards larger networks of coordinating agents, enabled by advances in the means of communication that allow for broader, higher fidelity and more centralised mindshaping and the dissemination of signalling strategies. Our evolved faculties of social cognition initially allowed for efficient coordination in tight-knit hunter-gatherer groups, wherein members shaped one another to adopt shared ontologies that facilitated fast and efficient coordination. More advanced communication technologies build upon this initial endowment, acting to expand the size of the groups sharing ontologies. Throughout history the enhanced coordination power generated by new communication technologies was typically monopolised by royal, religious or state entities, with new technologies altering the balance of power between the various elite factions.

The rise of digital, internet-based communication technologies as the primary means of communication is the latest development in this story. As a communication environment, it displays both continuities with legacy forms of mass communication, but it also has novel features – communication takes place on an entirely new substrate, one that is global, instant, virtual, high-fidelity, open access and non-hierarchical, and yet is dominated by a handful of private providers. In the next section, we will examine how the mindshaping-based conception of human social cognition can be used to shed light on how this novel communication medium interacts with the use of beliefs as signals in enabling coordination.

4.3 Signalling and Coordination Online

That the internet, and the use of social media in particular, is having problematic effects on public discourse is of course, by now commonly accepted, and the subject of a voluminous literature both popular and academic. This literature is particularly concerned with the status of truth or facts in the online sphere, often beginning with the presupposition that the existence and apparent popularity of false information on social media dupes or brainwashes individuals into making poor decisions, supporting demagogues or believing in conspiracy theories (for an overview of this literature see: Bennett & Livingstone (2020b)). As a result, the majority of proposed interventions are aimed at the level of individual users, for example increasing the proportion of facts agents are exposed to or teaching individual critical thinking skills (e.g., Kim et al. 2019; Pennycook et al. 2020; Bago et al. 2020). This preoccupation with the truth-status of claims in the public domain and with modifying information appraisal skills is prompted by the implicit influence of an atomistic, mindreading-friendly, conception of human social cognition and information acquisition – on such a picture humans are understood first and foremost as atomistic truth seekers. And, as such, the existence of false information in the public domain causes their well-intentioned efforts to go awry. Thus, resolving any issues with pernicious discourse online requires correction at the level of the individual agent and the information they are exposed to.

However, I argue that this conceptualisation leads to a misdiagnosis of the cause of dysfunctions in the online communication realm and to misguided and ineffective solutions. In particular, the focus on truth and facts, what I call “truth-chauvinism”, conceals the importance of coordination in generating ontologies, and the reality that individual agents are not particularly responsive to individual facts, but rather to the

status of knowledge claims in relation to their group. By taking the mindshaping-based conception of social cognition seriously, we can better understand how and why contemporary communication technologies, and in particular social media as they are currently designed, may generate pernicious outcomes. As we will see, they interact with the use of symbolic beliefs as signals for indicating in-group membership, generating significant noise, undermining large-scale coordination and dragging down general social welfare. On the reconceptualization presented here, the main issue with online communication relates not to the veracity of claims or the rationality of the agents involved, but to how the systems of communication utilised incentivise the use of symbolic resources that sow discord and out-group animosity, thereby generating significant coordination noise. This form of noise, in turn, renders coordination power more vulnerable to capture or misdirection by bad actors. What follows is a set of points summarising in detail how the mindshaping-based conception of ontology formation may shed light on the manner in which this new communication environment generates pernicious outcomes.

- **Interface design incentivises visible, explicit and non-ambiguous signalling.**

  The design of social media platforms, by encouraging salient public interactions in the form of comments, likes, retweets and posts, incentivises users to engage in frequent non-ambiguous signalling of group allegiances. This real-time interactive form of communication online is on a scale unprecedented in the history of humanity. Metrics of interaction are prominently displayed and often linked to individual profiles. The effect of these features is to incentivise at least a subset of users to engage in highly

visible herding around the signalling of symbolic, usually group-affiliated, beliefs. Damon Centola refers to this as strategic complementarity (Centola & Macy 2007). The idea is that the more people in your immediate group express certain types of opinions the greater strategic benefit you derive from aligning beliefs with them. In general, this logic lies at the heart of coordination, but the online context incentivises explicit, highly visible and non-ambiguous belief signalling in a manner that modifies the general signalling equilibrium. This may enable a vocal minority to wield outsized power in determining the truth status of specific claims within groups.

- **Lack of physical context and links to the offline world devalues signalling in general.**

  Participating in communication online is unlike most other public communication contexts insofar as participation is entirely virtual. In such a context, affirmations of controversial beliefs are unlikely to generate immediate physical danger, unlike for example shouting racial epithets in the street. Additionally, there is only a weak relationship between stating a belief online and one's reputation in the offline world. Usually, the self presented online is relatively isolated from the person's offline life. These features operate to significantly lower the value of signals in general. In other words, if speech acts carry less risk, then fake, dishonest or overblown signalling becomes more likely. Thus, online, existing signals are devalued, and a larger set of beliefs fall into the symbolic domain as they need not cohere with prior affirmations or communication norms. Given the importance of honest signalling for coordination, this represents a fundamental shift in the

economics of signalling. Effectively, by making voice more accessible to all agents, the value of any individual instance as a signal of genuine commitment is reduced.

- **Lowering the value of signals incentivises the use of more extreme signals.**

  By lowering the cost of signalling *in general* online communication incentivises the use of more extreme signals. This is because when signals become easy to fake and thus less trustworthy, agents are pushed to use more extreme signals as their use still carries some cost, making them more likely to be honest, for example, conspiratorial or fringe beliefs. The affirming of extreme beliefs to appear honest to a group is an example of what Hugo Mercier calls bridge burning (2020). By making a public statement endorsing a belief that offends a majority, or rejects widely held common sense, one signals willingness to bear a cost to join a group. For example, only a truly committed group member would proclaim that the earth is flat or the Holocaust never happened or that Dianetics is legitimate science. Mercier suggests that this sets up a dynamic whereby group beliefs are pushed in increasingly extreme directions: "For beliefs to work in the burning-bridges scenario, they have to be extreme. This creates an incentive for new recruits, or even for members who wish to improve their status in the group, to push the limit of what the group already finds acceptable." (2020: 195). This may also lead to more frequent use of moralised talk, because moralisation, i.e. making beliefs non-negotiable, is inherently costly as it involves limiting decision sets. Thus one significant outcome of the discounting of signalling

power online is the ratcheting up of the use of extreme and moralised rhetoric by groups across the political spectrum.

- **Low-level conformism enables the spread of extreme signals**

  The low-level behavioural matching processes described in Chapter 2 as core components of conformist mindshaping likely play a role in propagating more extreme ideologies within groups. Agents placed in groups tend to automatically mimic one another, synchronising behaviours and autonomic processes (Morgan et al. 2017). Furthermore, this phenomenon has been demonstrated to occur in the online context (Wikström et al. 2022). Thus, it seems possible that once specific signals are incentivised online, even for just a subset of group members, the tendency for in-group members to conform their self-presentations and behaviours allow such signals to spread rapidly even without agents consciously seeking to adopt them. The power of extreme beliefs to reliably signal group membership gives them selective advantage in the context of such unconscious matching processes – salient signals are more likely to be adopted and spread.

- **Homophily generates echo chambers, reducing intergroup signal alignment.**

  The online context, by enabling real-time, non-geographically restricted communication allows individuals greater choice in determining their interaction partners. This enables people to choose to mainly interact with others who share similar interests, dispositions and backgrounds, a phenomenon known as homophily (Acemoglu, Ozdaglar & Siderius 2021).

This is a well-documented phenomenon, one that leads to the emergence of echo chambers and filter bubbles in which individuals form groups isolated from divergent opinions (Nguyen 2020). This acts to entrench group ontologies, components of which may have been challenged in traditional communication contexts, which may have included a more diverse range of interaction partners. Additionally, this new context incentivises the use of more extreme signals due to the need to demonstrate group fealty in a homogenous relatively undifferentiated group. Echo chambers may thus generate an escalating spiral of extreme beliefs being used as signals which can develop unchecked by intergroup communicative norms.

- **Content bias activating information is given a selective advantage, modifying the signalling landscape.**

  As global, open access and instant communication forums structured by dynamic algorithms optimised to prioritise engaging content, social media function as a highly efficient selection environment. These features make them extremely effective at discovering "high quality information", where quality refers to having cross sample cognitive appeal (Acerbi 2019). Partly due to the explicit design of content selection algorithms, but also thanks to hardwired informational preferences, the determinants of informational fitness are linked to innate cognitive content biases, for example selecting for threat-related information, gossip or sexualised content (Youngblood et al. 2021; Acerbi 2019). For example, information communicating out-group animosity has been shown to be more likely to go viral online (Rathje et al. 2021) as has information that may be understood as false, but is considered

"interesting-if-true" (Altay et al. 2022). The prominence of content bias activating information online significantly changes the signalling landscape in which agents interact, making the sharing of content of this nature more likely. Importantly the information selected for may actually be true, but creates a biased representation of the informational sphere which modifies signalling strategies.

- **Threats are prominent online, tightening in-group norms, and increasing the importance of honest signalling.**

  One upshot of the efficiency of content selection online is that threat-related content is often "more fit" as it activates robust cognitive-behavioural biases (Blaine & Boyer 2018; Bebbington 2017). In general, hypervigilance about threats is adaptive – it is better to be extra vigilant for predators than to get attacked. Additionally reporting threats may improve an agent's reputation in a group as it signals competence (Boyer 2015). This has likely played a key part in the increasing prevalence of false polarisation, whereby individuals think opposing groups are significantly more extreme and monolithic than they are (Enders 2019). Content selection biases, in particular for threats to one's group, likely contribute to this perception. False polarisation modifies the signalling landscape by increasing the perception of threats from out-groups. Ecological threat perception of this sort has the effect of tightening in-group norms (Gelfand et al. 2011) which may create pressure to robustly signal group allegiance and appears to increase support for dominant leaders (Petersen 2020) and likely increases the moralisation of norms. Taken together, these factors suggest that the salience of threatening content online

may tighten in-group norms, further encouraging the signalling of allegiance via visible markers, and increasing support for dominant leaders which in turn increases threat perception for other groups, in a vicious escalatory cycle.

- **Efficiency and threat perception empowers extremist groups, who can supply honest signals.**

  The efficiency of the online communication environment enables agents with fringe or extreme beliefs to find one another and form coalitions, thereby significantly increasing their power and ability to influence discourse. In the pre-internet environment, the primary limitation to forming fringe or extremist organisations was one of communication. The coordination of malicious actors made possible by internet communication can provide such groups with leverage to cause disruption. However, the structure of signalling online also means that the beliefs of such groups, insofar as they exist along established political continuums, may become increasingly valuable, as honest signals become scarce. Groups lying at the extreme of ideological spectrums provide honest signalling resources with which to indicate allegiance to a given ideology, a function further incentivised by ecological threat perception. Daniel Williams describes a related phenomenon he calls rationalisation markets, whereby signal producers respond to demands from ideological groups for rationalisations of group-affirming positions (2022b). Fringe groups may gather support in a similar manner, whereby demand for honest signalling resources by more moderate agents leads to the production and spread of more extreme beliefs. The demand for honest signals created

by the general online debasement of beliefs as signals may thus create enhanced demand for extreme beliefs that still carry some signalling value, and thereby empower formerly fringe political formations.

- **Context collapse online mixes factual content with phatic communication and entertainment**

  The majority of agents using the internet to communicate are not primarily seeking factual information about the world, but rather use it for entertainment or commercial purposes and phatic communication (Allen 2020). However, the design of social media actively mixes the factual and the entertaining, with the valence of the factual often subservient to the entertainment value. This is continuous with the informational context developed by cable television (and indeed was likely a feature of the majority of pre-written communication as oral culture relies on memorability for information preservation), but it also incorporates an enhanced level of user interaction. This may increase the entertainment value of the online domain, but it also acts to make it an environment imbued with significantly more subjectively important signals – hearing a group member affirm a position appears as more trustworthy and is a more important signal than if provided by a news anchor or newspaper report. The context collapse of informational domains also means that political and factual information is often selected for its entertainment value. This may have the effect of incentivising the selective adoption of beliefs both for signalling and entertainment value as opposed to factual content. These effects emerge because beliefs hold utility for agents that is divorced from the truth status, and the formation of beliefs

online aims to maximise utility across a variety of domains (Bénabou & Tirole, 2016).

- **Emphasis on novelty causes signalling to trump coherence, generating noise.**

  A central design feature of social media platforms is the creation of a fast-moving, temporally immediate informational environment. Information arrives in real-time and is constantly updated, a design feature first developed by 24-hour-news-cycle broadcasting (Benkler 2020). This appears to drive engagement, as participants are repeatedly primed to engage pattern matching learning faculties[23]. However, this feature may also act to disincentivise epistemic accountability through time by rapidly shifting attention to the new. For example, it has been shown that when agents are exposed to fake news in a fast-paced setting it is more likely to be believed (Bago et al. 2020). In such a context, consistent signalling of group allegiance may take precedence over epistemic coherence. In general, incoherence of this form, whereby individuals adopt group signals regardless of their links to other claims, can generate coordination noise on a large scale. The emergence of large-scale, complex societal formations that require high-level coordination means that increasing the levels of sub-group ontology misalignment is highly problematic.

[23] Prima facie this bears interesting, largely unexplored, parallels with the design of gambling slot machines, as detailed by Natasha Dow Schüll (2014), insofar as such design is an example of utilising novelty to prime learning mechanisms and foster engagement. Indeed, it has been shown that individuals susceptible to problematic gambling are also likely to engage in problematic social media use (Bhargava & Velasquez, 2021; Akbari et al. 2023).

- **A lack of gatekeepers gives signalling priority over intersubjective agreement.**

  By providing an alternative to information sources controlled by institutional gatekeepers, such as professional editors, publishers and broadcasters, who were incentivised by professional norms to strive towards intersubjective agreement and coherency, the online information context is one in which signalling power often operates as the foremost criterion of value. For most, though not all communities, signalling group membership, as opposed to accurately representing reality, drives belief formation. As a result of supplementing the professionalised informational model with open access to publication, truth as a goal of enquiry is superseded by signalling efficacy in many domains. This further contributes to the generation of coordination noise.

- **Accurate information generates little economic value online.**

  The extreme abundance of information available in the online realm devalues information as a commodity in general. To understand how this is a novel development we can look at how publishers derived value from information in the past. For example, in the 17th, 18th and 19th centuries the yearly almanac was very popular in Britain and America, second only to the Bible in total sales (Tomlin 2014). This was largely due to the knowledge it contained – tidal charts, planting tables, weather records etc. – information which was at the time a scarce commodity. As such, for consumers, its signalling function was secondary (but not absent: the almanac was primarily popular with Protestants, and played a role in forging their distinct

subjectivity (Tomlin 2014)). As a result, almanac publishers were strongly incentivised to print useful and factual information. However, in the online context information, and, in particular, information of the sort formerly collated by almanacs, is so abundant that it is nearly worthless. Additionally, the institutional structures of the modern nation-state ensure that in many important contexts minimum standards are enforced by law. In other words, the affordances of the modern world (food safety standards, rules of the road, minimum schooling requirements etc.) ensure that individual beliefs often do not significantly, or immediately, affect basic welfare. However, the removal of basic survival concerns has the effect of further debasing the value of true information about the world in general. As a result, in many contexts the primary economic value that can be extracted from publishing information is via the collection of user data in order to better target advertising (Floridi 2015). Social media platforms gain little direct economic value from the truth value or general social utility of the information they host. Instead, their primary incentive is to use information to ensure people will stay engaged and reveal their consumption preferences[24]. This means that online publishers have few incentives to reign in signalling for the sake of social welfare as their revenue streams are unmoored from such concerns.

If operative, the dynamics outlined above paint a grim picture of contemporary public discourse. The efficiency of the online communication environment, coupled with its specific engagement-orientated design and its fundamental devaluation of

[24] However, as Floridi (2015) emphasises, controlling the means of information dissemination comes with necessary political power, and to say that these platforms have *no* stake in what their networks propagate is false.

accuracy or inter-subjective agreement as norms of enquiry or generators of value, interacts with the way beliefs are used as signals to indicate coordination potential. The disconnection between online speech acts and physical outcomes devalues the signalling power of beliefs in general. These features incentivise some individuals to adopt increasingly extreme beliefs on topics that have no immediate welfare effects for themselves and shift the discursive space shared with their interlocutors in increasingly extreme directions.

The removal of gatekeepers in structuring public discourse, coupled with the siloing of cultural groups, allows beliefs as signals to become decoupled from requirements to mesh with broader societal norms. As a strategy, adopting beliefs that align with fellow group members is individually optimal as it advertises and enhances coordination suitability. However, it has significant and pernicious effects in the aggregate. Specifically, these features of online discourse contribute to intergroup conflict, generate pervasive coordination noise and result in broad political stasis. At the core of these problems lies coordination power and the manner in which online discourse as currently configured misdirects and stymies the exercise of such power, sapping its effectiveness and, at worst, rendering it vulnerable to hijacking for pernicious ends.

Central to the above claims is the idea that what a group takes to be true, or its fundamental ontology, is primarily determined by signalling pressures, and the utility of different signals given broader structural features of the interaction environment, not by scientistic assessment of the informational environment to determine what is true. Thus, the modification of ontologies cannot proceed by the piecemeal correction of false claims but via the modification of the communication environment within which the agents are embedded. Like in the case of

denazification, to change beliefs and behaviours the broader incentive structure faced by the pertinent agents must change. An understanding of how communication contexts can be engineered to modify the incentives they generate is the key to increasing coordination and social welfare, not an obsession with truth, or facts, or narratives.

From an individual perspective, agents strategically optimise their social learning to either maximise efficiency in causally transparent contexts or, more often, to signal group membership and coordination potential. However, though individuals pursue optimal strategies, the aggregate outcomes of social learning in mass communication contexts are complex, non-intuitive and can often generate sub-optimal outcomes from the perspective of overall welfare. For example, coordination of ontology in an environment characterised by perceived threats or zero-sum interactions can lead agents to adopt strategies that require the elimination of, or support for the elimination of, other outsider groups. Mass communication technologies make the propagation of apparent evidence of threats and the exploitation of such information by political actors significantly easier.

At the limit, the features outlined above potentially set the stage for a formerly relatively cohesive polity to degenerate into a state of polarised intergroup conflict. This outcome can emerge from optimal individual strategy choices that serve to exacerbate out-group animosity to the point of sparking physical violence, a phenomenon described by Russel Hardin in his work on deadly coordination breakdown (1997). This set of circumstances underlies many of the recent instances of genocide, for example in the former Yugoslavia, Rwanda and Indonesia. In each case, modern mass communication media played a role in enabling specific political groups to disseminate signals to specific ethnic or political groups about outside

groups (Kiper 2022; Thompson 2007; Robinson 2017). This is one reason that mass communication technologies require careful scrutiny and transparent regulation, especially our contemporary high-bandwidth, large scale and instantaneous forms.

The dynamics proposed above show how the informational ecology created by the shift to online discourse may incentivise the adoption of increasingly extreme positions in any given discursive domain. In a free-for-all signalling environment where moderate signals are devalued, it becomes optimal to advertise group allegiance by adopting increasingly polarised positions vis-a-vis the "other" group. The prominence of threat-related information in the online informational ecosystem also has the effect of further tightening group norms, leading groups to tolerate less internal dissent and require more frequent displays of allegiance. Fundamentally this results in the misdirection of coordination power towards destructive ends. Indeed, recent increase in political polarisation seen in America (Abramowitz & McCoy 2019), Brazil (Gethin & Morgan 2018) and Western Europe (Borbáth et al. 2023), coupled with the global rise in instances of hate crime (Schweppe & Walters 2016) suggests that a dynamic of this sort may be playing out at large scale. Furthermore, the norm tightening dynamic described by Gelfand (2019), whereby ecological threats tighten in-group norms, may act to catalyse emerging political polarisation, both as a result of likely future material instability generated by climate change and also as a result of the perception of threats from out-groups that polarization itself generates.

The perceived availability of voice as a strategic option online exacerbates these polarization-amplifying tendencies. The disconnect between speech acts and real-world outcomes means that the range of potential expression is larger online. If there are few potential repercussions it is easier to speak up if one is unhappy with group

norms. Additionally, it is simpler in a virtual context to deploy exit and join a new group of like-minded agents. These moves need not have effects in the physical offline world to be desirable to agents. Indeed, the lack of offline outcomes on the individual scale may make such moves more attractive, or likely. The ease of such strategic moves online, however, belies their effects in the aggregate, acting again to increase polarisation and the disaggregation of the public sphere, while simultaneously undermining the likelihood or possibility of effective political action.

Fundamentally, the shift to the online communication context has resulted in the generation of significant coordination noise. Even if groups do not descend into open conflict, the effectiveness of a state to bring about increases in welfare relies on large-scale coordination. Insofar as communication online, in particular using social media, undermines large-scale coordination it directly affects large-scale social welfare. The factors outlined above, if operative, generate significant coordination noise. Fundamentally this results in the diminishing of overall coordination power available to states as a result of signalling dynamics in the online context. For example, the global response to climate change, relative to the by now well-established scale of immediate action required to avoid catastrophic outcomes, can be seen in part as an example of coordination noise hindering political action (though in this case aided by deliberate noise generation by corporate interests).

The upshot of pervasive political polarisation and widespread coordination noise is to generate deep political stasis. By making large-scale coordination more difficult and individual group ontologies increasingly incompatible the shift to online discourse appears to have made the development of a coherent and progressive social contract that can gather majority assent nearly impossible. Across the developed world, democratic states are increasingly incapable of developing

coherent and far-sighted policy (McCoy et al. 2018). Of course, the causes of this state of affairs are myriad, but the disaggregation of the public driven by the shift to online communication is a significant contributing factor. Fundamentally, political polarisation and the noise it generates saps the power that effective coordination has the potential to generate for human projects. This undermines the potential for polities to confront challenges, resulting in a greater likelihood of material scarcity, which itself further contributes to polarization.

The claim that public discourse has been negatively affected by the use of digital technologies for communication is now commonplace. However, the various proposed solutions to these crises, initiatives like fact checking, digital education or content moderation, assume the agents involved are at core truth seekers responsive to best evidence, agents who are merely slightly confused or misled by this new informational realm (Muhammed & Mathew 2022). On such a view all that is required to improve the quality of public discourse is the atomistic, piecemeal correction of factual claims or the inculcation of individual vigilance. Solutions of this sort emerge from the general implicit influence of mindreading-based conceptions of human social cognition.

The mindshaping-first understanding of human social cognition rejects this individualistic conceptualisation of human epistemological faculties, replacing it with a picture of individuals who are, first and foremost, hardwired to adopt the beliefs of close conspecifics as a signalling strategy. Behaviour like this is optimal, and thus the issue lies not with individuals, who are primarily responsive to concerns about coordination with fellow group members, but rather with the incentive structures within which they are embedded. Therefore, to resolve these issues we need to look *not* at changing individuals, or the individual facts they are exposed to,

but instead at how the *structure* of communication online interacts with group formation. Better infrastructure design, informed by a more accurate conception of social cognition and signalling incentives, may allow us to incentivise healthy cross-group coordination instead of exacerbating the formation of increasingly extreme and incommensurable group-level beliefs.

At core, my proposal is that the current design of online social media platforms emphasises epistemically regressive features of human social cognition due both to the form of these technologies and a design philosophy that prioritises increased user engagement. This reduces social welfare by hindering collective action. Redesigning this infrastructure using a mindshaping informed theory of social cognition can, I believe, help remedy some of these pernicious effects. Fundamentally, if we are to take seriously the idea of using a mindshaping-based framework to redesign online communication, then the descriptive and predictive power of the claims in underwrites must be empirically assessed. As an initial step towards providing such support the following section describes how the mindshaping-based approach to online communication makes different predictions from the mindreading-centric approach. I then provide some general methodological guidelines for future empirical work aimed at assessing the claims of mindshaping theory.

4.4 Re-Engineering Mindshaping Online

As we have seen, the mindshaping reconceptualization of agents and social cognition makes significantly different claims to an account informed by mindreading about how and why information is propagated and groups are formed. If mindshaping is on the right track then the claims it makes will stand up to empirical investigation and ultimately real-world implementation. In order to make

the contours of such a project clear, it will be useful to delineate some specific presuppositions and predictions the mindshaping paradigm makes as compared to the existing understanding of online communication which I suggest derives from a mindreading-centric understanding of social cognition. To this end, Table 1 lays out the ways in which the mindshaping paradigm differs from the existing dominant understanding of knowledge acquisition and social influence. Table 2 then builds on these claims to show how the two paradigms differ in their proposed solutions to the problem of dysfunctional information propagation and pernicious group formation online. Later, I will use the concrete claims emerging from the mindshaping paradigm to provide some preliminary methodological considerations that could be used to operationalise it for empirical study.

Current Paradigm: **Agent as Scientist**	**New Approach:** **Agent as Social Animal**
Social cognition is a purely epistemic capacity for accurately reading other minds. Agents aim primarily to adopt true beliefs, both about other minds, and the world.	Social cognition and coordination is enabled by interpersonal belief regulation. Agents aim primarily to signal in-group coordination suitability, not discover "true" information. Beliefs adopted emerge from processes of signalling and coalition formation that are difficult to control and often unmoored from accuracy considerations.

Facts are assessed piecemeal in a neutral knowledge domain.	Facts exist in webs of belief generated by one's larger cultural group. Beliefs are assessed based on suitability and consilience with prior, group derived, ontologies and for signalling utility.
Expertise is neutral and is used to guide agents to accurate data.	Expertise is community sanctioned, thus normative and contested: your expert cannot overrule mine.
Accuracy or efficiency of action are the primary utility derived from beliefs.	Utility derived from information is linked to goals other than accuracy or efficiency, particularly in modern affordance context. Specifically problematic are the signalling of in-group status and honesty/commitment.
Information gathering is a dispassionate and unbiased process of updating guided by instrumental efficiency cues.	Information acquisition is guided by social learning strategies which are sensitive to social cues and aims not only at adopting at accurate information, but also at enhancing status.
Agents adopt isolated instances of misinformation if presented in a believable manner.	Agents are "wary learners". Misinformation adopted usually aligns with pre-existing group derived belief sets. It preaches to the choir. The main issue is not with the specific information

	itself but “choir-level” belief sets and their signalling requirements.
Individual agents possess discrete, self-standing, *sui generis* selves, which guide and constrain their preferences in advance of social interaction. The informational priors updated by social interaction and knowledge acquisition are derived from these internal preferences.	Individual selves supervene on the processes of social interaction, as such their content is directly responsive to the broader incentive structures which constrain groups and emerge from interpersonal negotiation in such a decision field. For example, average Nazis were products of social conditions, not inherently evil agents. Beliefs are directly tied to broader incentive structures.

Table 1. Differing mindreading and mindshaping conceptions of information acquisition and function.

Given this divergence in basic presuppositions, a mindshaping-based explanation of social cognition also suggests that engineers and developers engaged in interface and system design in the online context must take a different approach. Specifically, mindshaping theory makes different suggestions about how to optimally design these affordances to more effectively minimise the emergence of pernicious zero-sum group conceptualisations and beliefs and to facilitate more civil cross-group dialogue. It suggests that the coordination noise generated by dysfunctional communication online must be addressed from the level of system design, aiming to re-engineer group-level interaction and incentive structures, as opposed to interventions aimed at the individual level such as re-education or fact

checking. To draw out how the two paradigms differ in their claims and proposals, Table 2 outlines some common solutions to the issues of dysfunctional online discourse that are broadly informed by a mindreading-centric conceptualisation of human social cognition, and outlines how a mindshaping-based conception can potentially correct and update these proposals.

Existing Proposal	**Mindshaping-Informed Correction**	**New Proposal**
Focus on correcting the truth status of claims in the public domain. i.e. factchecking.	Truth on its own is an unimportant criterion for most individuals in quotidian contexts.	Focus on why partisan or uncivil information is interesting or provides utility for groups in a signalling context. A central issue is a polluted/noisy signalling context providing fringe/extreme signals with enhanced influence.
Use education to teach agents to be more rational and cultivate epistemic virtues in assessing claims.	Agents are unmoored from so-called rational standards in most contexts and responsive first and foremost to social incentives.	Use education to create a link between social status and civil communication. Educate social groups to shame fascists, not to fact check the fascist's claims.
Sanction or flag elites who spread lies.	In-group reputation is delinked from external standards of accuracy. Information uptake is guided by group loyalty and group	Modify algorithms to minimise visibility of elites who flout transparent and democratically accepted civility standards (shadowbanning). Delink partisan

	driven entertainment value (a key driver of the popularity of Trump, Johnson, Bolsonaro).	speech from entertainment by developing clear expression rules that govern public speech by political entities.
Ban individual agents who flout established epistemic norms or stoke up partisan animosity.	Issues with uncivil discourse and out-group animosity relate more to signalling context and broader material or ecological factors than to individual malign actors. Extreme signals become more valuable when threat perceptions are raised.	Deescalate threat perception by reducing false polarisation, potentially by developing visible metrics of partisan agreement. Use algorithms to make partisan speech less visible.
Change algorithms to deprioritise falsehoods.	Outright lies play a limited role in generating group-strife, most partisan and uncivil group behaviour exists inside the realm of "truth". Selective presentation of facts is used to serve entertainment and group signalling purposes.	Reduce incentives or ability to use partisan animosity for entertainment or signalling value. Make overtly political speech carry less weight in algorithmic sorting on social media. Revalue signals in general by linking online behaviour to offline social reputation.
Make it easier for people to seek out	News consumption and fact seeking forms only a small proportion of most	Enhance capacity for rich phatic communication as a means to accrue status. More clearly

reliable information online.	individuals' use of the internet. Instead interpersonal mindshaping and group status reinforcing forms the bulk of social media usage.	delineate informational contexts, and change algorithmic ranking weights to discourage the mixing of political information with group status signalling.
Inform agents that specific claims are not reputable or are disputed.	Guidance may check individual instances of misinformation sharing but cannot modify incentives to adopt in-group congruent beliefs regardless of truth status. In some cases, officially sanctioned fact checking can mutate into a signal of out-group hostility.	Focus on signalling function that popular misinformation serves by placing it into a framework of honest signalling. Decreasing visibility of partisan discourse and general ecological threat perception should naturally decrease demand for such signals.

Table 2. Differing solutions to countering malign information propagation online.

Given the strong theoretical support for mindshaping outlined in the previous chapters, it is likely that at least some of the effects outlined and solutions proposed are on the right track. However, to fully demonstrate the utility of such an approach, the next step required will be empirical work that investigates these predictions in a controlled context. Such work will provide the mindshaping paradigm with valuable empirical support, and allow for the careful delineation of the various potential interventions that the model can help inform. Though such empirical investigation is

beyond the scope of this work, what follows are some general methodological considerations that can be used to inform the design of experiments aiming to operationalise a mindshaping-based understanding of communication in the online context.

4.5 Experimental Resources

The majority of extant research into online information propagation reflects the continued influence of the mindreading paradigm in explaining social cognition. For example, in investigating how misinformation is propagated, and in particular the influence of fake news on an agent's beliefs, the typical experimental approach is to present agents with instances of fake news and measure their level of credence in the information or their likelihood of sharing it after exposure (e.g. Porter & Wood 2022; Guess et al. 2020). In these experiments subjects act alone, the experiments are brief and don't usually measure the persistence of the effects reported. By minimising the role of interpersonal belief regulation, the implicit use of a mindreading-based paradigm encourages investigators to see agents as atomistic, and thus not particularly responsive to social pressures in forming beliefs.

A mindshaping-informed approach suggests that experimental formats such as these are deeply undermined by their lack of a social component: information assessment is not conducted in a social vacuum, and significant pressure to coordinate viewpoints emerges due to the ongoing interpersonal regulation of beliefs that takes place when agents communicate (Pickering & Garrod 2004). Thus, to effectively investigate information propagation in general, it is crucial to examine how information is shared in an experimental context that incorporates group/peer feedback processes. Human social life is necessarily public, and in the online context

with large, global audiences and highly salient status cues, it is public in a new and unprecedented manner.

4.5.1 Minimal Group Paradigm

Given the central role that coordination and specifically the pressure to signal membership of a coordinated group plays in generating many of the effects outlined above, it will be crucial for experimenters to be able to generate coordinated groups, or at least a simulacrum of such coordination on the fly. In order to do this – take agents unknown to one another in advance and create groups who feel pressure to regulate their beliefs and behaviours to maintain or signal coordination potential – experimenters can draw upon and adapt existing experimental protocols that investigate group coordination and its effects. One influential set of directly applicable experiments investigate what is known as the Minimal Group Paradigm (Otten 2016). This work demonstrates that the generation of biased in-group feelings on the fly is surprisingly easy. For example, in one of the first experiments to demonstrate the effect (Tajfel et al. 1971), the experimenters asked participants if they preferred paintings by Kandinsky or Klee and placed them into groups ostensibly (though not actually) based on their answers. This minimal categorisation was shown to induce significant in-group favouritism in the post-treatment distribution of monetary rewards, even when acting in the total common good required only a small reduction in the in-group payouts.

These findings, showing that in-group feeling can be relatively easily created in the lab using arbitrary and often random assignments of group membership have been robustly replicated in subsequent experiments and remain an important component of experimental work on social psychology (Otten 2016). Pinter &

Greenwald (2010) assess various methods for inducing minimal group membership, demonstrating that a highly effective manner to do so is to have participants memorise the names of a set of group members. The use of such techniques for generating an illusion of in-group solidarity in an online context is relatively straightforward, for example, participants can be randomly assigned to groups, asked to memorise the names of fellow group members and be provided with arbitrary visible symbols that demonstrate their group affiliation. Such treatments will likely generate a significant feeling of in-group membership and thus pressure to coordinate via mindshaping-based regulation, though, of course, these effects will have to be demonstrated empirically.

The use of this type of cueing to generate coordinated groups can also be supplemented with synchrony tasks, another approach that has been shown to generate in-group favouritism and enhanced coordination (Mogan et al. 2017; Wiltermuth & Heath 2009). Specifically, groups of agents who act in sync to achieve some task, for example walking in step or listening to music and moving together to receive some joint reward, have been shown to cooperate more in subsequent economic games, even if this requires personal sacrifice (Wiltermuth & Heath 2009; Tunçgenç & Cohen, 2016). Synchrony tasks of this sort can easily be adapted for the online context, perhaps with participants clicking on or manipulating objects in sync to achieve a joint task. The extensive literature on both the Minimal Group Paradigm and the effects of synchronicity on group construction provide valuable resources for the generation of coordinated groups.

4.5.2 Confederates

An additional experimental approach that will be useful for the empirical investigation of mindshaping processes is the deployment of confederates in experimental interaction. The use of confederates to influence individual judgement in social contexts was popularised by the now well-known Asch conformity experiments (1951), and this protocol has subsequently been adapted for online use (Wijenayake et al. 2020). For example, by presenting individuals with manipulated or fabricated representations of group judgements experimenters demonstrated that 78% of participants conformed with the majority opinion at least once, though this varied in proportion to information type, with conformism more likely in objective tasks with correct answers (Wijenayake et al. 2020). Confederates will be useful both in aiding the generation of apparently coordinated groups and for manipulating information flows and signals to attempt to influence agents to conform to different signalling strategies. Given the virtual context these confederates need not be actual agents, making their use and deployment significantly lower cost and simpler than in traditional experimental contexts. The deployment of confederates in this manner can allow experimenters to vary what individuals perceive as being dominant or minority group beliefs, and investigate the effects that varying the proportion of such beliefs have on agents primed to coordinate with co-participants.[25]

4.5.3 Monetary Incentives

Another experimental technique that will be of use in investigating how group membership may alter an agent's ontology is the use of monetary incentives to

[25] The use of deception is considered to be unethical in experimental economics, ruling out the use of confederates if such work is conducted in an economics department.

elicit accurate belief distributions from subjects (Harrison et al. 2022). This can be done by presenting subjects with real-time lotteries, with payments in their local currency, which allow them to make bets about their subjective beliefs about various answers obtaining, with pay-outs for correct answers. By allowing for probabilities to be discovered, more subtle changes in belief distributions pre- and post-treatment can be ascertained. As we saw in the Minimal Group Paradigm, economic games, usually centred around resource distribution are often used to reveal agents' in-group biases, however, these are usually hypothetical distributions and, as such, may not accurately reflect an agent's preferences. The addition of monetary incentives to the experimental toolkit provides experimenters with a powerful means of discovering actual belief distributions, and to partially correct for confounds emerging from the necessarily artificial context of the laboratory.

Specifically, when testing for the effects of mindshaping as a result of intergroup alignment, it will be useful to incentivise agents to reveal private beliefs, in addition to beliefs they reveal in the group context. Testing for divergence in these two states can perhaps reveal if mindshaping and group signal alignment alters or shifts agents' belief sets due to group influence, and if those changes are enduring or the ephemeral products of in the moment social pressure. Additionally, belief discovery in this manner may make it possible to measure subtle shifts in belief alignments that may occur as a result of participating in coordination enhancing exercises with partisan confederates. If incentivised belief distributions are determined before and after treatment, and linked to ideological stances, then it may be possible to see a shift in baseline beliefs in more extreme directions as a result of the artificial signalling pressures generated during the treatment.

4.5.4 Online Sample Recruitment

Experimental protocols designed to investigate the claims made above about online communication should closely mimic the online experience. Thus, experiments should utilise computer interfaces with agents communicating using the types of affordances common to social media platforms. Conducting experiments using services such as Amazon's Mechanical Turk (MTurk) to recruit and conduct experiments online has become increasingly popular in the social sciences and has various features to recommend it (Almaatouq et al. 2021; Bentley 2020). For example, it allows experimenters to use a much larger sample without incurring significant cost increases and may enhance the experimenter's ability to recruit a more representative/diverse sample. Furthermore, conducting experiments online can allow for much longer experimental timescales to be used, with repeat participation being easier to incentivise compared to requiring physical presence in the laboratory (Salganik 2018). These features will be particularly useful in investigating mindshaping processes, allowing for the monitoring over time of how beliefs change due to specific types of interaction, and the persistence of such changes.

However, despite the potential advantages of online participant pools for social science research, the use of services like MTurk to conduct online experiments is somewhat controversial and presents specific issues relating to sample quality and participant attentiveness that must be taken into account if such services are used (Hauser et al. 2019). For example, experimenters must carry out rigorous participant screening to identify problematic or inattentive MTurk participants (Pyo & Maxfield 2021; Bentley 2020). However, it has been demonstrated that once adequate screening has been conducted that MTurk samples are at least as reliable as those recruited on campus (Kees et al. 2017).

Researchers attempting to operationalise the mindshaping paradigm in the context of online communication should make use of the empirically proven resources just described. In particular, they should test methods for generating coordinated groups on the fly, using trivial or ideologically-based assignments, group synchrony tasks and/or the attribution of visible affiliation markers. These coordinated agents can then be exposed to information in contexts with motivated confederate agents seeking to alter their belief sets. Decisions and belief assertions should be made in a context that they understand as being public and having reputational implications. And experimenters should make use of monetary incentives in assessing belief distributions pre- and post-treatments to tease out the power of group-based effects on preference formation.

4.6 Conclusion

A central presupposition of much mainstream discourse about the ills of the online informational context is that filling the internet with facts would go some way towards resolving the problem of incompatible ontologies and help fix the widespread political polarisation that is a feature of many western democracies. However, adopting a mindshaping-based explanation of social cognition reveals that facts alone are not the issue, and instead the prevailing view obscures the reality that facts are themselves sites of contestation influenced by pre-existing power differentials. In many contexts, choice of emphasis and selective reporting can allow for the same source material to be used to support radically different ideologies. Thus, removing all obvious falsehoods online wouldn't solve the problem of polarisation, incivility or hostile in-group ideology formation. Instead, structural reforms are required to reshape how individuals and groups relate online to enhance

impartial, transparent and civil communication, and reduce incentives to engage in extreme signalling.

Thus, the central issue that must be kept in focus when thinking about how to improve online discourse is not false information, but rather the fundamental malleability of human ontology. From this perspective, the widespread truth-chauvinism that characterises contemporary liberal society is incorrect. It belies the fact that ontologies are contingent and change through time: many things that would have been fact checked as true in 1823 are now false. In general, an outsized focus on facts and truth inevitably ends up reifying a specific ontology. That this is the case is forcefully demonstrated by the fact that, by our current lights, severe and abhorrent injustices were naturalised components of everyday ontologies in the very recent past: this reality, which can be described as moral progress, is not however unidirectional, it relies on upon a set of factors, not least the enhanced material abundance made possible by industrialisation, but also upon the spread of a specific ideology, liberalism, an ontology that explicitly promotes equality. In the online context we must focus on generating inter-group solidarity, not on fixing truth, an approach that tries to convert today's contingent attempts to cope with reality into eternal dogma.

Indeed, the outsized focus on fake news and misinformation in mainstream discourse about online communication inadvertently serves to shore up the dysfunctional aspects of the current design model. This mistaken diagnoses leads social network providers to make visible, yet fundamentally ineffective, interventions which do not address underlying incentives for users to engage in partisan signalling and develop increasingly extreme in-group ideologies.

Arguably the various dynamics outlined above are still to reveal their worst effects. Given the role played by material scarcity in causing in-group norm tightening, if the coming climate emergency imposes increased material constraints on populations, then it is likely that the underlying problematic features of contemporary communication environments just described will begin to make their full force felt. Thus, the redesign of these crucial public affordances making use of a more accurate conception of social cognition and the role of interpersonal mindshaping in constructing, constraining and naturalising ontologies is urgent. To better understand how we can incentivise such redesign we will need to examine the political and economic context in which contemporary online communication, and in particular social media, was developed. It is to this task we now turn.

Chapter 5 – Politics, Economy and Spectrums of Possibility

"…although the conclusions of the social disciplines were about man, they were treated as if they were of the same nature as the conclusions of physical science about remote galaxies of stars. Social and historical inquiry is in fact a part of the social process itself, not something outside of it. The consequence of not perceiving this fact was that the conclusions of the social sciences were not made (and still are not made in any large measure) integral members of a program of social action. When the conclusions of inquiries that deal with man are left outside the program of social action, social policies are necessarily left without the guidance that knowledge of man can provide, and that it must provide if social action is not to be directed either by mere precedent and custom or else by the happy intuitions of individual minds"

John Dewey, 1963 [1935], Liberalism and Social Action

To make the conclusions of social sciences integral members of social action, as Dewey urges in the quote that opens this chapter, requires social scientists to apply their theories in the world with the aim of improving it. In this spirit, the previous chapter set out a range of recommendations informed by mindshaping theory aimed at reforming digital communication infrastructure to bring about beneficial social outcomes – i.e. the reduction of coordination noise, the promotion of civil inter-group discourse and the minimising of incentives for agents to adopt extreme beliefs as signals. However, in addition to the theoretical application of conceptual apparatus to a given set of problems, efforts which aim at realising social action also raise a set of concerns that are fundamentally political in nature.

For one, such social action requires would-be reformers to make their own political outlook explicit. In other words, to lay out their vision of a better world such reforms aim to bring about. My own position is informed by the work of John Dewey and Richard Rorty, and is liberal democratic in bent. According to Dewey social action should be aimed at the realisation of *effective* liberty for all (1963). This contrasts with an alternative understanding of liberty as a primarily legal construct, achieved through the recognition of individual equal standing in the eyes of the law.

Effective liberty, as Dewey construes it, is more demanding than this, requiring in addition the maximal reduction of material insecurity for all agents in a liberal democratic society. This is because Dewey understands liberty as a *social achievement*, not merely an inborn quality that can be unleashed by suitable institutional arrangements. Its achievement requires agents to be given the time and necessary resources to develop individuality. In this sense Dewey's conception of individuality dovetails with that of mindshaping – agents with individual selves are products of social structures, not pre-existing entities to be set free. Thus social reform must be aimed at the minimisation of material insecurity, itself deeply dependent on our ability as a species to maximally coordinate our energies and minimise welfare destroying intergroup strife. This political position, and the mindshaping-compatible manner in which it has been further developed by Rorty, will be outlined in greater detail in the conclusion of this chapter.

In addition to laying one's own political cards on the table, and indeed more importantly, any attempts to realise social reform must also take stock of the extant political terrain within which such efforts take place. In other words, to implement effective reforms requires an awareness the broader political-economic realities constraining the actions of the various actors involved. The specific governance and economic regime which has structured the bounds of the possible, at least in the political context, over the last several decades is often described as neoliberalism. As we will see, though it is increasingly under strain (Gerstle 2022), this particular economic and political philosophy continues to structure governance across the globe. Furthermore, and of particular interest given the focus of this work, the neoliberal worldview has played an outsized role in the evolution of digital communication technology. Ideas drawn from neoliberal ideology are widely

appealed to by the leading corporate entities in order to justify their contemporary dominance and specific form. As a result, any attempts to influence the providers of these technologies to better align their design goals with those of welfare-oriented policymakers requires close attention to be paid to neoliberalism. Thus, the first task of this chapter is to provide a detailed description of neoliberalism and to show how it has influenced the development of digital communication infrastructure.

If this broader incentive structure can be modified however, either through targeted government regulation or the development of alternative business models or infrastructure forms, then digital communication holds out great promise for the future of human society and Dewey's vision of liberal democracy. It can potentially enable highly efficient global coordination, radically democratise the public sphere and expand the possibilities for creative self-expression. Yet, simultaneously, given their complexity, pervasiveness and malleability these technologies can also, if incorrectly or perniciously deployed, pose significant threats to general welfare and individual freedom, far beyond the baleful effects currently being generated. Thus, in an effort to make clear the stakes at play, and trace the realistic possibilities that social action should aim at both realising and avoiding, I examine a set of positive and negative potentialities digital information communication technology may generate.

In what follows, in Section 5.1 I first outline and problematise neoliberalism, the specific political economy regnant over the previous few decades. Then, in Section 5.2 I show how this governance regime has deeply influenced the development and deployment of digital communication technologies. In Section 5.3 I turn to an examination of possible future states that digital communication infrastructure can help realise if reformed, for better or worse. Finally, Section 5.4

closes this chapter, and the thesis itself with a summary of the main points raised across the work, drawing out the crucial threads and claims that make up its core content.

5.1 Neoliberal Utopias

Neoliberalism, as we will see, is a species of utopian thinking that correctly highlights the epistemic capacity of markets, but by adopting a faulty commitment to their supposed apolitical nature, renders these structures open to capture by specific interest groups. This vulnerability to market capture has affected the development of digital mass communication technologies in particular, as neoliberal ideas guide their design philosophy and provide a justifying ideology for their contemporary form. As a result the effects of neoliberalism on the design of communication technology, and indeed the internet more broadly, may ultimately outlast the political dominance of the ideology itself.

Neoliberalism promotes a particular conception of states, markets and individuals, one that shares close affinities with the individualist mindreading conception of social cognition[26]. Specifically, neoliberalism repudiates the state as a competent agent for promoting societal good, places market structures in its stead, and understands individuals in atomistic terms – the upshot of these claims has been to promulgate a lax regulatory regime and atomistic social ontology that has, among other deleterious outcomes, enabled the rise of powerful and opaque digital technology corporations. The hands-off neoliberal approach to market design has

[26] The simultaneous rise to prominence of neoliberalism and atomistic, spectatorial conceptions of self-formation in philosophy can be understood as two faces of a general intellectual trend emerging during the latter half of the 20th century, one implicitly driven by the apparent threat of Soviet state socialism and the rejection of collectivist conceptions of human achievements this entailed (Amadae 2003, 2016).

allowed these corporations to amass unaccountable power and created incentives for them to design their products in a manner that inadvertently undermines social welfare.

Neoliberalism is a notoriously contested concept (Venugopal 2015; Vallier 2021), and in popular usage it is often used to refer to a vague yet ominous formation at the root of the many dysfunctions of the modern world (i.e. Monbiot 2016). This malleability has led many to reject the nomenclature as hopelessly indeterminate and thus analytically useless (Dunn 2016; Grzanka et al. 2016). However, notwithstanding these concerns, there exists a persuasive and illuminating strand of research that takes the concept seriously and uses it to shine a light on a particular set of governance techniques that were regnant in the latter decades of the 20th century. On this account neoliberalism can be understood as a coherent political philosophy, one with a distinct history, core aims and key intellectual protagonists (Mirowski 2014, 2019; Brown 2015, 2019; Slobodian 2018; Davies 2016; Biebricher 2018). In my interpretation of this work neoliberalism emerges as a deeply utopian, but fundamentally flawed project, one motivated by legitimate concerns about liberty and the challenges to individual freedom presented by state entities.[27]

The early neoliberals were responding to a widespread loss of faith in liberalism following the dislocations of the two World Wars, the blow to capitalism generated by the Great Depression, the apparent successes of state-run war

[27] In what follows I refer to neoliberal political philosophy as a monolithic concept with a core set of central theses. Naturally this idealises the intellectual terrain and the various works and thinkers collected under the rubric of neoliberalism differ in their emphasis and indeed often contradict one another. Neoliberalism is perhaps best understood as a family concept with a core emphasis on markets as preeminent epistemic devices usually shared by all parties. In what follows Hayek's work is used when characterising core, or orthodox neoliberalism. However, contemporary neoliberalism is a type of "pop" or folk neoliberalism that differs in key respects from the proposals of Hayek and other early neoliberals. This formation emphasises an atomistic conception of human social life and places individual responsibility at the root of societal dysfunctions, and is used to argue for a "small state", in the sense of a state that plays a minimal role in wealth redistribution.

economies and the rise of National Socialism and the Soviet Union (Gerstle 2022; Biebricher 2018). It is this context, Wendy Brown argues, that gave purpose to the nascent neoliberals: "Forged in the crucible of European fascism, neoliberalism aimed at permanent inoculation of market liberal orders against the regrowth of fascistic sentiments and totalitarian powers." (2019: 9). However, the proposed nostrum contained oversights and strategic elisions that have paradoxically allowed neoliberalism to sometimes align itself with authoritarianism and which have generated a deep complacency about the total power which digital technology corporations wield over contemporary communication structures.

Neoliberalism, according to Philip Mirowski, one of the ideology's most historically astute biographers, is fundamentally a political philosophy of market society (2014: 8). On his telling neoliberalism is not only, or even primarily, an economic doctrine. Instead, it is a proposed description of knowledge generation devices from which an ideal form of governance emerges. The core of this political philosophy is a belief in "the epistemic superiority of the market in all things" (Mirowski 2019: 46). This conception of the market – as an information processor as opposed to merely a resource allocation device – owes its popular origins to Friedrich Hayek, a canonical figure in neoliberal thought. However, for neoliberals the market is epistemically superior not just because it is an efficient information processor – aligning supply and demand across complex domains and efficiently coordinating consumption strategies – but also because it is *apolitical*, processing information transparently without top-down oversight (Davies 2016). These special properties mean that the primary role of government must be to help market structures operate without interference. This role does not necessitate a small state (as many critiques of neoliberalism incorrectly claim, e.g. Harvey (2005)) but rather

a state that is clear about its role, and committed to expanding and protecting market mechanisms by whatever means necessary (Mirowski 2019).

For neoliberals, the crucial importance of such apolitical information processing emerges from a general suspicion of value-laden political discourse, understood as beholden to charismatic rhetoric and plagued by individual deficiencies in rationality (Davies 2016). Using the market to answer disputed political questions about goods and ends removes such value judgements from the hands of potentially corrupt, and likely incompetent political processes and functionaries. As they see it, the market, in compiling the limited and private information of unlimited numbers of agents, information that cannot be otherwise discovered, is the only structure in a position to efficiently guide social policy without subjecting individuals to the will of an arbitrary majority. Any other type of conscious or directed planning presupposes, in Hayek's estimation, "the existence of a complete ethical code in which all the different human values are allotted their due place" (1944: 42-43), an obviously impossible proposition for a liberal social order.

By using the market as an unbiased oracle, one that can answer political questions without invoking inherently relative value judgements, neoliberals in effect pursue "*the disenchantment of politics by economics*." (Davies 2016: 19, original emphasis). Neoliberalism thus consciously aims at the depoliticization of public discourse. This is the utopian core of the neoliberal project: it envisions a future post-political idyll, in which disputes about ends and means are seamlessly, impartially, and efficiently resolved through the power of the market. This is a seductive vision since corruption is a seemingly inevitable by-product of power and if it could be eliminated, by outsourcing all important decision making to an incorruptible oracle of truth, that would undoubtedly be a good thing.

However, unfortunately, the neoliberal conception of the market as an impartial information processing device suitable to replace political deliberation is inherently problematic. Fundamentally this is because markets are *not* in and of themselves apolitical. Markets must be designed and require standards of valuation, themselves the product of normative deliberation. Thus neoliberalism is fundamentally conflicted "in its relationship to its own prerequisite ethos: a wholly calculable, measurable world is only possible on the basis of particular non-calculable, immeasurable values or vocations." (Davies 2016: 22). Neoliberal market-based valuation thus smuggles into supposedly apolitical market-based judgements "an implicitly moral agenda, which makes certain presuppositions about *how and what to value*" (Davies 2016: 8, original emphasis). Indeed, in practice what neoliberal governance often amounts to is a vindication of markets *as they are now*, replete with existing, often discriminatory, value systems. This then entrenches status quo power relations, though now reconceptualised as the just outcomes of efficient market processes (Brown 2019).

By positioning the market as value-free neoliberals merely "shift questions of normativity elsewhere, into spheres of expert procedure and methodology, while often ignoring the irredeemably normative constitution of socio-economic life." (Davies 2016: 26). Thus the central problem for the entire neoliberal construct, diagnosed repeatedly by Mirowski, is that "The Market" as an all-seeing, all-powerful, apolitical information generating device is a fiction, and instead contemporary market society decomposes into "a collection of diverse boutique markets operating with differential effects for the clients of the business of market design" (Mirowski & Nik-Khah 2017: 241). Rather than acting as oracles that give us access to the true or the real, markets are in reality themselves the sites of

struggles where "truth" is produced. If markets incorporate inescapable value judgements and are fundamentally the products of design, design aimed to bring about ends for specific agents, they are patently *not* apolitical.

Given the utopian core of neoliberal thought, perhaps it is best understood as an experiment designed to test the hypothesis that market forces can provide a tolerable replacement for potentially corrupt and inefficient political deliberation in determining the structure of society. However we now have some conclusive results from this multi-decade experiment and the data doesn't look good, at least when we place it in the light of the post-war political settlement that obtained up until the 1970s: national inequality has shot up in the decades since neoliberal governance norms became dominant (Milanovic 2016; Piketty 2014), life expectancy in America and the United Kingdom (the most orthodox neoliberal nations) has fallen (Chetty et al. 2016; Wise 2022) and increasing numbers of people live in lethal despair (Case & Deaton 2020). Additionally, the much-lauded global reductions in extreme poverty supposedly achieved during the peak neoliberal years turn out in many cases to not reflect reality on the ground (Alston 2020). These facts cast doubt on the proposition that relying on the market alone to determine resource distribution, insofar as this is the true aim of neoliberal political philosophy[28], is conducive to the broad-based maximisation of welfare.

[28] Itself a contested question. For example, David Harvey (2005) develops a Marxist critique of neoliberalism describing it as "a political project to re-establish the conditions for capital accumulation and to restore the power of economic elites" (19). On his view the insistence on supposedly apolitical market structures merely provides cover for capital accumulation. The overlap between the rise in inequality and the ascendency of neoliberal governance structures suggests that his thesis carries some weight, though his use of Marxist categories is ill suited to the contemporary structure of capitalism and his imposition of intention on the part of an elite class vastly overestimates the competence of such formations. Indeed, in my view, it is not that neoliberalism is itself an elite conspiracy, and more that its specific mistaken conception of the market has enabled and indeed incentivised the capture of political processes by economic elites who then rationally use this power to entrench market forms that further benefit them.

This is not to imply that the widespread adoption of neoliberalism was the result of a consciously apprehended conspiracy. Nor did the originators of neoliberalism see their interventions as a move in a class war. However, as the statistics on national inequality show us, a class war (or rout) is what they nevertheless brought about, one waged from the top. It was not their aim, but crucially, thanks to their understanding of markets, it is not necessarily an outcome incompatible with their preferences either (Biebricher 2020). If applying market principles to society results in unequal resource distribution then that is a regrettable cost of achieving freedom from state coercion and reflects a justified distribution of resources to actors all privy to the same information and subject to the same forces. And if allowing technology corporations a free hand in determining their products generates a fractured public sphere, well then again this is the regrettable, but unavoidable and ultimately apolitical, outcome of allowing a free market in such services. Thus the neoliberal approach to governance has directly contributed to the baleful state of contemporary mass communication, giving technology platforms full control of the designing of their products. However, the pernicious influence of neoliberalism on contemporary communication extends beyond merely the justification of a lax regulatory regime. It also provides a problematic ideology of selfhood that is used by technology companies to justify the design of their communication platforms.

In tandem with its veneration of the epistemic capacity of markets, neoliberalism also promotes a specific conception of the agents interacting with these market structures, a view that forms a now influential social ontology (Foucault 2008; Brown 2015; Mirowski & Plehwe 2009; Davies 2016). Neoliberalism understands society (at its best, given the right markets, protected in

the right ways) as composed of individual entrepreneurs relying on feedback from impartial and efficient market mechanisms to generate profits, reinvest this capital and drive innovation forward. This entrepreneurial spirit is not, however, restricted to the sphere of business: human selves are also a type of capital, capital that too requires investment and innovation to remain profitable (Feher 2009).

Each individual is in turn responsible for their own stock of capital and should be encouraged, through their embedding in suitable market structures, to strive to be effective entrepreneurs of the self (Foucault 2008), appreciating their value through wise investments in health, education, self-presentation and interpersonal networks. Thus, for neoliberalism, self-conceptions are ultimately responsive to the market, and communication structures should efficiently transmit demand signals for various self-conceptions, enabling individuals to maximise their capital value. This conception of selfhood licences the encroachment of market amenable valuation into domains formerly structured by more intangible, and uncommodified (though still strategic) value structures, such as religious or national identities, or kin loyalties. The current form of social media – wherein pervasive metrification and easily curated public records of engagements and beliefs encourage self-sculpting to match trends, capture the attention of others and thereby increase the market value (however defined) of the self – can be seen as an example of this neoliberal conception of self-formation put into practice.

In understanding social outcomes as the result of individual entrepreneurial agents operating within value-free and maximally efficient markets, mainstream neoliberalism is committed to methodological (and ontological) individualism, a claim that both underwrites and entails its rejection of socialism, neoliberalism's enduring political foil (Amadae 2003). This is because, on its conception, agents

who have access to the necessary information required to prosper are individually responsible for choosing correct actions to maximise their fortunes. Outcomes can be reduced and explained solely by reference to individual action. This results in a deep scepticism about structural explanations of social ills, and in the depoliticization, and individualisation, of social facts like racism or economic power imbalances (Brown 2015).[29]

However, despite this misguided commitment to methodological individualism (Ross 2005; cf. Udehn 2001), we can nevertheless acknowledge neoliberalism, at least in its classic Hayekian form, as recognising the role played by social forces (in their case market forces) in shaping individuals. Social forces *do* play a constitutive role in determining the content of individual selves, and the resource distributions generated by specific markets are now, and have been since the emergence of intergroup trade, crucial components in these processes. That selfhood is responsive to social forces, and that agents placed into specific contexts invariably adopt self-presentations that benefit them, is also a core premise of mindshaping. However, as stressed throughout this work, from the perspective of mindshaping this fact renders the specific institutional environments within which agents are embedded with enhanced ethical, and thus political importance.

Thus we return to the fatal flaw of the neoliberal utopian fantasy: despite claims to the contrary, market structures are not generally apolitical or value-free, they encode specific conceptions of value or ways of valuing (Davies 2016). Due to the malleability of selves, placing agents into specific neoliberal market structures inculcates them to adopt or be responsive to the specific values encoded within those

[29] This approach bears interesting similarities with the mindreading approach to explaining social cognition – the agent of mindreading, who engages in semi-paranoid anticipation of the actions of opposing atomistic and isolated actors, whilst embedded in a competitive and often zero-sum social world is the same agent at the centre of the neoliberal free-market fantasy.

structures. By structuring institutional and communication infrastructures as if agents *are* first and foremost responsive to solely *market-derived* pressures in forming their self-conceptions, such agents are brought *into being*. In other words, by enacting specific institutional reforms which restructure the incentive field within which agents are embedded, neoliberal social ontology becomes a self-fulfilling prophecy. What this demonstrates is the central role the design of institutions, including, but not limited to, communication infrastructure, plays in partially determining the types of agents which populate a given human society. Yet, for neoliberals, who truly believe in the utopian vision of markets as ur-epistemic devices, intentional institutional design is at best sub-standard, and instead the optimal approach to enhancing broad-based welfare is for flawed human actors to defer to design guided by the "free market".

Effectively the widespread acceptance of neoliberalism's claim that market structures are fundamentally apolitical acts to stifle discussions about the values and ends necessarily embedded within these structures, which in turn play a role in constituting the agents interacting with them. As a result, the neoliberal understanding of governance relinquishes crucial processes that guide self formation to markets that are in reality saturated with values, which, cloaked in the rhetoric of spontaneous emergence guided by market forces, are then left to be determined by entities that act to further their own interests in an unaccountable or uncontrolled fashion. If markets play a central role in making individual human selves then markets and the institutions that support them require careful, transparent, and democratically accountable design. The fact that human self-conceptions are responsive to market forces, as neoliberalism recognises, is the very reason the neoliberal laissez-faire approach to market design must be tempered, and why the

design of communication infrastructures *cannot* be left solely in the hands of private, profit-motivated actors.

The historical concurrence of neoliberal governance and the development of digital communication technology has meant that these affordances have largely been developed in a context characterised by a laissez-faire approach to private enterprise. Neoliberal ideology has thus played a significant enabling role in generating contemporary digital communication technology while furnishing its providers with a powerful justifying ideology.

5.2 Neoliberal Digital Communication

In the 1990s and early 2000s, just as digital technology was emerging as a powerful new tool for communication and information dissemination, neoliberal conceptions of governance and social ontology had reached their apogee. By the 1990s there appeared to be no alternative to the neoliberal agenda. This was (and still largely is) a world described by Mark Fisher as governed by capitalist realism (2009), a context in which capitalist modes of production (and in particular neoliberal variants of capitalism) are taken as natural facts and thus structure the bounds of the possible. This stasis in turn owes its origins in part to the specific and highly effective manner in which early neoliberals, led by Hayek, made strategic use of knowledge production institutions, like universities, think tanks and prize ceremonies, to promulgate and validate their alternative to socialism (Mirowski & Plehwe 2009).

As a result, for several decades across the developed world, parties of the left and right, along with global institutions like the WTO, the IMF and the World Bank, were in agreement about the optimal governance model. This was structured around

a universally applicable package of neoliberal reforms – privatisation of the public sector, deregulation of utility provision, labour casualisation, reduced tax burdens, free international capital flows and globalised supply chains – each component designed to protect markets from democratic interference and allow them to operate as the sole resource allocation devices (Brown 2015; Davies 2016; Mirowski & Plehwe 2009; Biebricher 2018).

Simultaneous with the development of this economic-political consensus, the internet was emerging as an important means of information communication. The architects of the now dominant information communication platforms, people like Mark Zuckerberg (Facebook, b. 1984), Larry Page (Google, b. 1973), and Jack Dorsey (Twitter, b. 1976), came of age in an era in which neoliberal ideology had become the orthodoxy. As a rapidly growing, profitable and almost entirely deregulated industry operating on a global scale, the early trajectory of the internet seemed to vindicate (and perhaps in part *did* vindicate) neoliberal ideas about entrepreneurship and the growth potential of deregulated markets. The early days of innovation in what became known as Silicon Valley were led by self-styled hackers, a subculture that was firmly libertarian in bent. However, regardless of their unorthodox views of society (at least by the standards of the day), these developers required investment to bankroll their innovations, and this came in the form of venture capital. Thus from the origin of digital technologies there existed a productive alliance between personal freedom-oriented libertarians, and profit seeking investment funds **(**Gerstle 2022**)**.

As a result of this overlap and mutual reinforcement, neoliberal political philosophy has played a significant and privileged role in how digital communication companies conceive of and design their products, while

simultaneously structuring the broader political-economic framework within which they operate. The circumscribed (or highly selective) regulatory regime that goes hand in hand with neoliberal governance directly enables the outsized and largely unaccountable role technology corporations play in public life. The combination of neoliberal hostility to market regulation and it's view of market structures as applicable to all realms of human experience has played a key role in the emergence of an online communication ecosystem that is dominated by a handful of corporations that systematically subordinate epistemic values to commercial aims and seek to monetise information dissemination across all spheres of interpersonal interaction.

As of 2022, a handful of corporations, Meta, Alphabet and Twitter, control platforms that carry a majority of online communication. Facebook has almost 3 billion active users monthly (Meta Platforms 2022), Google is responsible for 85% of global online search traffic (StatCounter 2022) and Twitter has become the primary source of political news for social media users in the US (TechCrunch 2022). This consolidation of power and centralisation of control over the means of information exchange has multiple causes, but at root lies the ideology of neoliberalism.

One key factor is the existence of powerful network effects in the online realm. The importance of utilising shared communication networks means that specific digital services often become dominant, particularly within regional blocs. The reasons for this are clear: if a communication platform is to be useful it should connect users with the majority of people they may need to coordinate with, and once a single platform emerges with a slight majority of users it will tend to become the preferred service. As a result online communication platforms in their current

form possess features of a natural monopoly, insofar as the majority of users will generally tend to prefer to all use one service over multiple disconnected platforms.

That digital communication services tend towards monopolistic positions in some domains is acknowledged as a cause for concern (OECD 2020), but for now, barring radical redesign[30], this appears to be an inherent feature of such services. Additionally, these entities actively and openly pursue a strategy of buying out nascent competitors to consolidate their market positions. However, though widely acknowledged, the fact that network effects and corporate strategy have enabled these crucial service providers to assume effective monopoly positions has not resulted in concrete regulatory action directed at the communication structures they provide.[31] As a result, at least in the Western world, these corporations largely are allowed to design their services without oversight. This lack of action too emerges directly from neoliberal conceptions of the market as an impartial information processor.

Given their advocacy of well-functioning markets for efficient resource allocation and the role they give to governments in safeguarding such markets, one would expect neoliberals to be wary of a small number of players dominating service provision and actively quashing competition in an industry. This seems to be an obvious example of market failure in need of regulation. However, for neoliberals, this is not the case. Influenced by the Chicago school variant of neoliberalism, and in

[30] As it stands the main communication platforms prohibit rivals from providing services for, or accessing, users on their networks. One proposal aimed at undermining monopolies that arise from network effects is to require interoperability standards between various social networks and allow users to communicate across platforms (Bailey & Misra 2022). Of course, from the perspective of neoliberalism, enforcing interoperability standards of this sort would be an example of problematic state overreach.

[31] Big tech as a whole has been the subject of some recent scrutiny both in America and the EU, however in both cases the focus was on individual privacy and data protection, an approach that "is compatible with corporate values such as profitability, setting legal compliance as a low bar for Big Tech. This sidelines values beyond profit, especially where these entail calls for social justice" (Birch & Bronson 2022: 4).

particular the work of economist James Buchanan, oligopolies, monopolies and dominated markets have been reconceptualised as the valid outcomes of market processes (Kiely 2018; Van Horn 2015). Indeed, for contemporary neoliberals such outcomes do not reflect a market failure but rather the justified rewards of successful business practice, a natural and beneficial outcome of competition.

Underpinning this approach to antitrust regulation is a specific controversial definition of consumer welfare, popularised by Robert Bork (1978), functionally interpreted as equivalent to lower prices for consumers. In other words, if a company comes to dominate a market but this does not lead to price increases for consumers in the medium term, then its domination is not problematic. As a result, contemporary neoliberals generally reject antitrust regulation in the digital sector because most digital communication services are provided free of charge, nominally benefitting consumer welfare in the Borkian sense, despite the existence of entities that have amassed high levels of market power (Bourne 2019). This same reasoning is applied to the aggressive buyout and merger strategies pursued by the leading technology corporations – so long as consumers benefit from lower prices these consolidations are not problematic. Thus, neoliberal approaches to antitrust, by relying on a circumscribed notion of the public interest, have set the stage for a handful of corporations to dominate the provision of online communication with minimal regulatory pushback.

However, neoliberal conceptions of consumer welfare, solely understood in pecuniary terms, do not apply seamlessly to the online communication sector. Specific issues arise when public communication is dominated by a handful of providers, issues that do not relate to the cost of these services. By implicitly understanding information as merely another commodity to be exchanged and the

facilitation of communication as a service to be provided by actors incentivised to offer the lowest prices, the neoliberal understanding of consumer welfare ignores the crucial role played by information exchange in human affairs.

According to the mindshaping conception of social cognition, as we have seen, public information exchange – how and what is said, by whom and when – structures social life, determining belief sets, optimal signalling strategies and the structure of group coordination. As a result, when a means of information dissemination becomes dominant it accrues significant power. Concern about such power is all the more pertinent in the online domain because the network effects that generate natural monopolies appear to be a structural feature. One dominant centralised communication service may fall, but another will rise to take its place. The crucial role of information communication structures in public life implies that once the providers of online communication services become dominant in their sector, they cease to be merely privately interested entities and become social actors with specific responsibilities. Indeed, the unique features of the online communication sector have generated proposals that it should be regulated as a public utility (Rahman 2018).

While a state monopoly on informational infrastructure would itself be problematic, there are grounds to argue that the online sphere requires a public broadcasting equivalent, considering its indispensable function for information dissemination and the role it plays in structuring quotidian interaction (Pickard 2020). However, regulation of the sort commonly utilised in relation to public utilities, like water, electricity and transport, have been severely eroded during neoliberalism's tenure as a guiding ideology. This atmosphere of hostility to regulation in general has allowed the corporations controlling the infrastructure that

facilitates much contemporary communication, commerce and ultimately coordination to remain opaque and democratically unaccountable (Taylor 2021).

Given this hostility to regulation the prevailing model of governance for digital entities has been one of internal self-regulation (Floridi 2021). In addition to a favourable political regime, the ongoing stability of this state of affairs owes no small part to the lobbying efforts of technology corporations, which operate one of the largest industry lobbies in the United States (Zakrewski 2022; Popiel 2018), a fact likely replicated across all states in which these entities operate and fear regulation. However, as Floridi reports, the self-governance model is increasingly acknowledged as a failure. At least in the EU, regulators are beginning to understand that the social externalities associated with unregulated digital media require law-based rules (Floridi 2021).

This is one example of the sort of structural change required to begin to address the issues generated by unregulated and highly mediated mass communication as provided by technology corporations. However, the magnitude of the lobbying efforts that can be marshalled by these large corporations suggests that the path to effective, wide-ranging and welfare improving regulation will be rocky at best. Whatever the future holds for neoliberal modes of governance the vast economic power that its regime has allowed technology companies to amass may emerge as one of its most enduring legacies. This is because market power itself generates political power, and as such, the dominant technology corporations have become political forces. This reality, coupled with the inherent power that emerges from controlling the infrastructure of mass communication (as the legacy media barons, like Rupert Murdoch, realised in the 1980s (Benkler 2020)) does not bode well for future attempts to regulate the sector.

In addition to promulgating and validating a lax approach to antitrust law and public regulation, neoliberal claims about markets are also used by the dominant corporations to argue that digital communication, and in particular social media, doesn't *require* oversight. One approach in this vein involves appealing to the market as an optimally efficient epistemic device and then arguing that in merely providing a substrate on which a "market in ideas" can operate the providers remain neutral parties (Pasquale 2016). Social media companies thus deny that they are publishers of content, and claim instead that they merely provide the infrastructure, or platform, on which individual agents exchange information, creating spontaneous market structures through which information efficiently flows to users (Gillespie 2010; Vadde 2021).

On this account, a platform is solely a form of infrastructure, something quite different to a traditional publisher. Social media companies can then argue that decisions about what information to publish and promote are resolved by a market of ideas they facilitate, which organically and efficiently sorts information in a transparent, fair and unbiased manner (Starr 2019). This platform/publisher distinction is enshrined in law in the United States by Section 230 of the Communications Decency Act of 1996 and is consistently used by social media companies to reject calls for greater responsibility for the content they host (Cusumano 2021). This legislation, emerging in the heyday of bipartisan support for the broad structure of neoliberalism, states that: "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider" (United States 2021). Thus "interactive computer services" are explicitly conceptualised in apolitical terms, playing a technical facilitator role.

Of course, there are limits to this stance and communication platforms do engage in some explicit moderation, particularly of abusive or illegal content. Nevertheless this appeal to the market as an apolitical and efficient resource allocation device is used to prop up a state of affairs that is profitable for the corporations concerned, insofar as it enables them to use the information they host to further economic goals. This specific provision in law, allowing what are publishers to be seen as mere market facilitators, illustrates how deeply neoliberal logic has influenced contemporary understandings of the digital sphere. Regulation or not, the idea of a free market in ideas where the best or most desired information naturally rises to the top, with innovative corporations playing an apolitical technical overseer role, is high neoliberal fantasy.

That such a stance is a fantasy, and the existing law outdated, is revealed by the fact that social media is deeply mediated, with the information it prioritises selected to enhance specific goals. It is by now common knowledge that content selection algorithms structure what information users are exposed to online. In the case of contemporary social media these algorithms are optimised to maximise engagement with the explicit aim of reselling this user attention to advertisers (Hendricks & Vestergaard 2019; Wu 2017). As such these platforms *instrumentalise* information and use it to attract users and keep them engaged while simultaneously amassing vast trails of information about their interests, needs, socioeconomic status, etc. Of course, due to the sheer volume of information produced online it is a domain that requires some means of sorting. The primary, and perhaps only viable, means for doing so is with automated algorithmic content moderation, as is currently the case. Thus, the question is not *whether* social media providers should interfere with the content they host or not, but rather *how* they should interfere.

As services that are usually free for users (which as we have seen also allows them to evade competition law scrutiny), online communication platforms face incentives to monetise their users. To do this they seek to guarantee advertising clients a large, engaged and finely segmented set of potential consumers (Wu 2017). Thus they are incentivised to design content selection algorithms to discover and prioritise content that serves these goals. As we saw in the previous chapter, this information selection environment often promotes information that activates inbuilt human content preferences, for example, threat-related information, moralised discourse or content of a sexual or disgusting nature (Acerbi 2019). For example, it has been shown that out-group animosity predicts the likelihood of information being shared on Twitter, thereby increasing engagement (Rathje 2021) and thus the likelihood of content selection algorithms exposing users to information of this sort, and that YouTube recommender algorithms consistently promote increasingly extreme and radicalised content in a bid to engage users (Alfano 2021). As we have seen, one problematic side effect of these attention capturing techniques is to generate a widespread perception of threat, and to promote outrage in the face of perceived threats (O'Callaghan 2020).

Thus, as a result of the economic basis of social media profitability, namely the monetising of attention, platforms are incentivised to use their services to promote a highly moralised, extreme, often politicised discourse that emphasises threats and out-group transgressions of group norms (Atari et al. 2021; Parsell 2008). This lies at the core of the dysfunctional signalling dynamics outlined in the previous chapter. Agents placed in such an information environment face pressures to openly declare their group allegiance, and do so using the signalling resources provided by their perceived fellow group members.

Essentially, by engaging users with novelty, threats and outrage social media diverts user energy into supporting ad hoc in-group ideological campaigns, simultaneously obfuscating structural and material issues that lie at the root of popular discontent. Among other deleterious outcomes, this mode of information curation serves to undermine and defuse the emergence of coherent coordinated polity that could effectively act as a check on existing political power (Brady & Crockett 2019) whilst simultaneously and paradoxically leading to the increased politicisation of discourse.

As a result, while the migration of public discourse online into a more interactive context has indisputably increased the quantitative levels of public political, or at least moralised, speech in general, concrete outcomes, in the form of challenges to the established and often inequitable status quo, have been largely ephemeral. On a fundamental level, social media as developed under the auspices of neoliberal political ideology serves to undermine the emergence of broad-based political coalitions, by dividing the populace into competing factions primarily focussed on signalling their stance on subjective cultural issues. A wide-ranging review of global empirical work examining the effects of digital media on democracy (Lorenz-Spreen et al. 2023) has shown that the penetration of digital communication technology has generally increased levels of polarisation, led to declining levels of trust, and created opportunities for populist political actors to gather support.

In theory, if the broader political incentive structure was to shift, either due to an exogenous shock or perhaps endogenous pressure from civil groups it may be possible to bring about a pivot in the digital ecosystem. In its place could emerge a decentralised, user welfare focussed and transparent model of online discourse with

publicly accountable governance regimes. However, beyond the obstructions that these now highly profitable corporations exercise via lobbying efforts, there is a further, complementary reason to be pessimistic about the prospects for such a transition: the phenomenon of technological momentum.

This idea, first proposed by Thomas Hughes (1994), is an attempt to describe the trajectory of large technological systems. The idea is that when a specific technology, for example, the automobile, is in its early stages of development its form is fluid and responsive to social forces. However, later when the technology becomes widely adopted it gathers momentum, making changes that fundamentally alter it in response to social pressures less likely. Once they are established, widespread technologies begin to shape society to fit their requirements. For example, contemporary landscapes are altered to suit the form of the automobile. If Hughes is right, and certainly the structure of many technologies appears to follow such a trajectory, then perhaps neoliberalism's most enduring contribution to human society will be the ways in which it has influenced the initial design of digital communication, and the internet more broadly. Thanks to the momentum now inherent to the vast technical system that is the contemporary internet, neoliberalism may live on past its potential demise as a governing ideology in the structures and design choices that characterise the online realm.

The fact that a handful of private corporations have been allowed to dominate the provision of public communication infrastructure, controlling both the form of those infrastructure and their content, is extremely problematic from the perspective of liberal democracy. The technology industry, in its explicit interventions, for now, seeks to appear relatively neutral politically, but we have yet to see how the main players may react to measures that could impact the profitability of their current

business model, for example, attempts to reform or temper the network effect driven dominance of individual firms or to implement regulations targeted at the form of the interaction structures requiring providers to prioritise welfare-prioritising goals.

That these powerful corporations will remain politically neutral in such a turn of events is unlikely. Nevertheless, the reform of this sector is a crucial component in any project aimed at addressing contemporary political malaise, and to confronting the crises of polarisation and out-group animosity faced by the global polity. Digital communication technology, and the types of information it incentivises users to adopt and the types of groups and selves these incentive structures favour, will play a central role in determining the form of future social and state formations. Understanding the specific political ideology that has allowed contemporary online communication infrastructure to take the form is has will be crucial for any project of reform. Furthermore, a clear understanding of the flaws inherent to the neoliberal conception of market outcomes is necessary to refute the faulty arguments these platforms use in order to justify their dominance and design. Fundamentally, any effective modification of the incentive structures faced by the providers of digital communications technology will require an understanding of the existing, market orientated, incentives that contributed to their current form.

5.3 Spectrums of Potentiality

When examining digital communication technology my main focus thus far has been on how these affordances may be generating problematic effects for large-scale coordination, thereby potentially undermining social welfare. However, like the printing press which, as we saw, set in motion a broad restructuring of society that culminated in the emergence of modern liberal democratic states, digital mass

communication technology too has the potential to help further democratise power structures, radically enhance global productivity, and generate a more diverse and representative social world. Thus, in this section I want to supplement my negative outlook with some observations about the potential positive effects these technologies can have for coordination, self-formation and large-scale welfare. These potentialities notwithstanding, this promise is also counterposed by significant threats. Achieving broad-based, global, social welfare gains will require a necessary balancing of these potentialities.

Digital mass communication technology has long been recognised as holding great promise, and until recently, a general optimism accompanied its rapid global penetration (e.g. Johnson 2013; Jarvis 2011). Many commentators, "techno-optimists" as they came to be known, believed the global interconnections enabled by digital technologies and the general democratization of communication they enabled would usher in a new era of prosperity and peace, with authoritarian states crumbling once citizens recognised what they were missing in the free world, one conspicuously led by the free-market polities orbiting around the United States (for an overview and powerful critique of this trend see: (Morozov 2013)). However, such a state of affairs has evidently not materialised, and instead, this techno-optimism has curdled, with the zeitgeist decisively shifting towards "techno-pessimism". Interestingly, the era of broadly exuberant popular accounts of digital technology ran more or less concurrently with the unchallenged dominance of neoliberalism as a governance technique – the shift to pessimism occurs around 2016 with the global rise of authoritarian and populist rejections of the status quo[32]. These

[32] However the extent to which these political formations, which undoubtedly make a break with the status quo in terms of political style and rhetoric, have in fact rejected neoliberalism is up for debate. For example, Trump's Republican Party (Wraight 2019), Bolsonaro's PSL (Iamamoto et al. 2021), Salvini's Lega (Monaco 2022) and Orbán's Fidesz (Geva 2021) all combine authoritarian, racialised

newly pessimistic accounts tend to emphasise actually occurring negative externalities associated with these technologies, often ones determined by the nature of the technologies themselves. For example, the manner in which they can and often do enable surveillance both by state and private actors, how they may perpetuate or reify oppressive social formations and how they allow perniciously motivated actors to spread disinformation (e.g. Mullaney et al. 2021; Zuboff 2019; Frischmann & Selinger 2018).

Both pessimistic and optimistic outcomes are possible, but neither is as of yet fully upon us, nor are these outcomes determined by the technologies themselves – to realise net positive outcomes it is crucial to keep a close eye on the structure of these affordances and the broader political and economic incentives faced by their designers and users. In what follows, I set out in broad brush strokes the kind of end states I believe we should aim to both usher in or avoid via the manipulation of these structures and incentives. As we will see, depending on their design, these technologies generate both promises and threats for the central goals of Deweyan liberal democracy – the reduction of suffering and enabling of free imaginative self formation.

In particular, several promising features of digital communication can be discerned: in making communication more efficient and lower cost, it potentially enables the type of mass coordination required to confront global challenges and

rhetoric with a broadly neoliberal economic agenda – privatisation of state industries, deregulation of business, austerity, welfare reforms and shrinking tax burdens. Though each have implemented some policies rejected by pure neoliberalism, particularly state support for indigenous industry and protectionist measures, the broad direction of their economic agenda is akin to geographically circumscribed neoliberalism. This is because actually existing neoliberalism, referred to above as pop neoliberalism, allows for the elite factions who support these new political formations to continue to enrich themselves. The apparent compatibility of elite enrichment, outright racism, and extreme nationalist sentiment with the continuation of neoliberal political policy demonstrates that the specific form of governance that neoliberals provide justification for need not necessarily be socially liberal. It also shows us that the claims of the early techno-optimists, insofar as their claims were premised on an expansion of the dominant political economy of their era, were deeply flawed from the start.

thereby enhance global welfare. It has in many cases eroded barriers to participation in public discourse – barriers that acted to marginalise minority voices and maintain status quo cultural formations. Partly as a result of this, the shift to mass digital communication also broadens the set of potential selves individuals can adopt. This is because increased global communication, in theory at least, widens the set of mindshaping models in the public domain, enabling greater nuance and free self-expression in the formation of individual selves. If realised, these potential outcomes can enhance general social welfare, reduce material insecurity and allow for rich imaginative self-formation.

However, each of these potentially positive outcomes – enabling efficient coordination, democratising access to public discourse, and expanding the set of potential selves – is matched by an opposing pole: digital communication can generate significant coordination noise; its specific technologies can enable pervasive gatekeeping; and it has the potential to restrict, or homogenise, as opposed to expanding, the set of mindshaping models potentially available for adoption.

Each potentiality can be understood as a pole on a spectrum, and though presented in terms of positives and negatives it is unclear that fully realising either end of each spectrum would generate an unambiguous positive or negative, respectively, for humanity. Rather, in each case, there is a sweet spot that lies somewhere in between the poles, though in all cases it lies somewhat closer to the "positive" end than the "negative". Thus we will see that there are some benefits to coordination noise, just as efficient global coordination can generate potential negatives for self-expression. Democratising discourse empowers not just oppressed minorities but also fascists and bigots. And maximum freedom in self formation can sometimes generate negative welfare outcomes.

5.3.1 Efficient Coordination/Coordination Noise

Internet-based mass communication can enable extremely efficient coordination. It allows agents located almost anywhere in the world to communicate instantaneously or broadcast their beliefs on a mass scale. It allows high-fidelity information to be stored indefinitely and transferred instantly with no loss of quality. This is an unprecedented development in the form and nature of information communication, fundamentally altering the patterns and processes of global knowledge transfer and information preservation. Beyond the benefits to individual humans, who can tap into vast stores of knowledge and communicate with one another with great ease, this radical increase in the efficiency of communication dramatically enhances the potential efficiency of large-scale coordination.

This is because increases in the efficiency of communication generated by the transition to digital communication technologies, through gains in speed, bandwidth and reach, can spread common knowledge to larger groups with greater ease. In theory, effective common knowledge generation via the unambiguous advertisement of preferences using these global information dissemination devices can help larger groups of agents align their strategies to better coordinate action (Chwe 2001). Such a capacity is increasingly crucial in the contemporary world where societies are pervasively interlinked via global trade and thus require common coordination schemes to efficiently interact.

For example, the global efforts to contain the COVID-19 pandemic serve as a recent instructive test case, demonstrating the power of digital communication to drive effective coordinated action, though in this case for a subset of motivated actors, particularly scientists and public policymakers. The global response was

rapid, with the first instances of human-to-human transmission officially identified in January 2020 and the majority of national lockdowns beginning at the end of March 2020. This was then followed by the development and mass production of effective vaccines within a year of the initial identification of the pathogen. This rapid response was directly enabled by online information communication technologies, for example in developing and sharing the parameters of national lockdowns, tracking the movements and close contacts of agents in some countries, and coordinating the efforts of researchers to determine the most effective mitigation responses and vaccine development strategies (Wilson & Jumbert 2018). The (relatively) orderly and rapid global response to the COVID-19 pandemic offers a clear example of the value of efficient digital communication driven coordination in confronting global existential threats.

The power of digital mass communication to enhance coordination on a global scale is likely to become increasingly indispensable in the coming years. Supra-national coordination, of the sort seen during the COVID-19 pandemic, will be required to mitigate the now inevitable disruptions that will be generated by anthropogenic climate change. In particular, effective global governance will be required to respond to planet-scale tragedy of the commons issues of the sort that necessarily emerge when the carrying capacity of the global commons comes under increasing strain. Effective measures to mitigate issues that take this form will require effective oversight and monitoring of resource usage across a diverse range of institutional settings (Ostrom et al. 1999). Digital communication technologies can play a central role in ensuring accountability and compliance with measures like emission reduction mandates and in mitigating carbon tax evasion. Enhancing the ability of states to resolve global coordination dilemmas, and to better respond to or

avoid catastrophes is a clear potential positive outcome of the adoption of digital communication technologies. In this sense, if correctly utilised digital communication technologies can, and indeed do, directly contribute to global welfare enhancement.

However, the prospects for international coordination and effective global governance are themselves partly reliant on the emergence of a relatively unified polity that assents to expert guidance in assessing and predicting outcomes. Unfortunately, as we saw in the previous chapter, the current structure of online communication significantly undermines the efforts to achieve consensus of this sort. This can again be illustrated using the example of the global response to the COVID-19 pandemic.

Specifically, the emergence of a large, and highly coordinated, cohort of "anti-vaxxers" in many nations was directly enabled by digital communication technologies (Germani & Biller-Andorno 2021; DiRusso & Stansberry 2022; Benoit & Mauldin 2021). Online information dissemination, particularly via social media, provides an accessible and low-cost means for various coordination vectors to compete in the battle to shape the minds of individual participants in public debate. However, as we have seen, the contemporary online economy is driven by a problematic incentive structure, one that leads to the construction of groups of agents that are highly coordinated around what are perceived as existential threats and thus remain maximally engaged. Thus, as described in Chapter 4, the widespread use of digital communication technologies often drives in-group overcoordination and out-group coordination fracture, in particular through the generation of moralised echo chambers (Atari et al. 2021; Parsell 2008).

When the structures of our communication technologies operate in this manner it systematically alters the equilibrium dynamics for all agents seeking to coordinate. This can have the effect of incentivising groups to adopt increasingly extreme or fringe belief sets to reliably signal group membership, as arguably seen in the surprisingly wide appeal of anti-vaccination conspiracies. The net effect of these processes is to generate groups committed to diverging opinions, primarily adopted for signalling purposes, which can be zero-sum, and often extreme. Thus the same efficiency of communication that underwrites global coordination in some domains can also generate significant coordination noise, thereby potentially undermining global responses to pressing crises.

Fundamentally the issue is that effective and efficient coordination is neutral concerning the types of action or belief adoption it enables. Effective coordination is necessary, but not sufficient for large-scale welfare enhancement. Coordination can underwrite both negative and positive outcomes, and the design of signalling incentives can, intentionally or inadvertently, underwrite significant harm – the Nazi regime created a powerfully coordinated mass by manipulating such incentives. Thus enhancing coordination efficiency alone is as likely to generate negative outcomes as it is positive. As it stands, the highly efficient nature of these technologies for communicating information, coupled with the perverse incentives faced by the platforms that design them, creates a fertile environment for conflict between cultural, ethnic and racial groups.

The fact that quotidian communication currently takes place on platforms that structurally incentivise animosity to out-groups, means that for now the global noise generating effects of efficient in-group coordination are more likely to dominate. However, though efficient coordination may power effective responses to

global crises, the end-state of a fully efficient globally coordinated polity may itself require a problematic reduction in cultural diversity. This is because to maximise coordination agents must adopt closely aligned cognitive profiles. Such an end state would potentially undermine the prospects of diverse experimentation in organisation and self-formation required to generate innovation. Diversity of background and experience provides crucial and otherwise unattainable epistemic benefits for enquiry (Page 2007). Thus, we need enough coordination to counter existential threats, but not so much that global society becomes monolithic. Fundamentally this spectrum of outcomes – fully efficient global coordination or mass coordination noise – requires careful navigation.

5.3.2 Democratising Discourse/Empowering Gatekeepers

An additional positive effect of the widespread adoption of digital communication technology has been the removal of publication barriers and the empowerment of minority voices this entails. By making the publication of information effectively costless and widely accessible the internet has directly enabled a general democratisation of discourse. This is a radical, indeed unprecedented, departure from pre-digital era publication systems which were highly professionalised with editors, publishing houses and broadcasting networks largely controlling what information entered the public domain and reached a mass audience. This epistemic democratisation has had tangible effects on social discourse in recent years. There have been a series of movements that have explicitly aimed to increase public awareness of the experiences and injustices faced by minority or oppressed groups. These are narratives of oppression that went largely overlooked or ignored in previous eras. For example, the Black Lives Matter movement has

generated mass awareness of the disproportionately high proportion of deadly police violence faced by the Black community in the United States and indeed worldwide, and the growing awareness of discrimination against transgender and gay people documented online has pushed some states to legislate in protection of their rights.

This democratisation of the public sphere benefits individuals insofar as it may lead to beneficial policy or social changes which can reduce individual suffering (e.g. Wong, Garza & Robbins 2021; Levy & Mattsson 2023). But also, as mentioned above, from a macro perspective greater diversity of participants in public discourse generates general epistemic benefits. As Scott Page (2007) argues, diversity of background and opinion in a group of enquirers is a more important component in problem-solving that the specific intellectual powers of any one agent. By democratising the public sphere, digital communications technology in theory enables the participation of a more diverse set of actors in public discourse, thus enabling the generation of more effective solutions to the various challenges faced by society.

Yet a corollary of the general removal of publication barriers is the empowering of voices that spread ideas that run directly counter to the ideals of a democratic and equal public sphere, in particular promoting the repression of minorities. The recent mainstream visibility of far-right political discourse has partly been enabled by this same democratisation of discourse online. The issue then is not so much that gatekeepers are always and only bad, but rather public discourse requires the right sort of gatekeepers, ones that transparently apply democratically developed principles that align with the goals of liberalism.

At the other end of this spectrum is the fact that the shift to digital, internet-based technology for communication also has simultaneously radically increased the

potential control it is possible to exert over information. Unlike a book or a pamphlet, information publishing online necessarily relies on complex and spatially distributed backend technology, systems that are beyond the control of individual users. Information can be blocked or modified and its users recorded and monitored in an unprecedented manner. This problematic feature of online communication is well illustrated by the relationship between authoritarian regimes and online communication. These states use the affordances of digital communication to quash dissent, and monitor their citizens (Niaki 2020). China in particular tightly controls the digital realm, explicitly engineering the processes of information production and dissemination in order to prohibit alternative coordination vectors from emerging (Griffiths 2019).

Indeed, the rich and interactive nature of the online communication sphere allows this control of the informational landscape to extend beyond brute repression. It has been shown that the Chinese authorities routinely inundate online information services and social networks with content following specific events that could spark collective anti-state action (King, Pan & Roberts 2017). The aim appears to be to flood the information environment to distract users, and thus avoid potential competing coordination vectors from emerging – in order words the explicit and active generation of coordination noise. What this demonstrates is how the mediated nature of the online information sphere can allow motivated agents to systematically interfere with the collective responses to politically salient events like injustice or political incompetence and how strategic use of the online informational ecosystem can undermine group coordination.

Undoubtedly, particularly in the Western democracies, the adoption of digital communication technology has contributed to the significant democratisation of

discourse, fuelling diversity in opinion and the empowerment of minority voices. Yet, simultaneously, the particular form of these technologies means that they can be, and often are, weaponised to control discourse, and shut down or drown out potentially competing coordination vectors. Furthermore through removing *all* publication barriers digital communication has also enabled the spread of ideologies that encourage the repression of out-groups. That these technologies can both enable free speech *and* more thoroughly repress it, give voice to the oppressed *and* to bigots, means that they must be carefully designed and regulated if they are to contribute to the enhancing of broad-based welfare.

5.3.3 Mindshaping Model Diversity/Cultural Homogenisation

A third set of potentialities generated by the global adoption of digital communication technologies relates to a central goal of Deweyan liberalism, namely facilitating imaginative self formation (Rorty 1989). Digital communication platforms, by reducing publication barriers and allowing agents who are spatially and linguistically isolated to communicate, can radically expand the potential repertoire of mindshaping models available to any individual. In theory, a global shared communication platform allows the potential set of selves to become larger. Given that a central goal of liberal polities is the maximisation of individual freedom such an expansion in resources for imaginative self formation is to be celebrated. The normative individualism that guides contemporary policy in the Western democracies seeks to maximally enable experiments in self formation, and it is such experiments that fuel the generation of new vocabularies, or ways of seeing the world, and with them the possibility for moral progress, both individual and societal (Rorty 1989). Allowing selves, as culturally installed virtual entities without inherent

content, to draw upon the maximum range of resources possible for their formation is thus central to cultural and moral progress.[33]

In this sense, even if in most cases agents choose to remain exemplars of their given cultural group, expanding potential access to understandings of how to be human is a beneficial outcome, as it makes it more likely that through recombination our collective ways of understanding what it is to be human will continue to evolve. Just as the printing press decisively shifted how agents came to conceptualise themselves in relation to one another, their rulers and the spiritual realm, bringing about a reduction in arbitrary cruelty and authoritarian rule, so too the internet can fuel similar human progress. Providing it remains a shared, global and relatively democratic communication affordance, while hopefully evolving a structure that minimises the generation of inter-group conflict, it can potentially lead to the emergence of new more inclusive conceptualisations of selfhood and of the place of *Homo sapiens* in the world. Such new vocabularies can potentially help us reconceive the relationship between our species and the other inhabitants of the world, an increasingly urgent requirement in the face of anthropogenic environmental destruction.

However, simultaneous with expanding *potential* access to a greater range of resources for self creation digital communication technologies can also act to *narrow* the set of potential selves that individuals may adopt. As we have seen coordination is enabled by the adoption of shared signalling resources which both advertise in-group status and conform ontologies making coordination easier. As a result of the efficiency of communication online and the large volume of participants and

[33] Progress in the sense not of a teleological understanding of history as inevitably approaching some end state of moral justice, but rather the demonstrable expansion of the moral circle over the last century to include more agents deemed worthy of protection from avoidable suffering.

potential signalling resources, coupled with the general pervasive sense of threat or zero-sum out-group competition that the specific communication structure fosters to enhance engagement, agents face incentives to adopt unambiguous signals advertising their group status. Signalling dynamics of this sort fundamentally restrict the set of self-shaping resources available within groups (Gelfand 2019).

Thus, on the one hand, digital communication technology, at least potentially, enlarges the sum total set of possible mindshaping models that can be used in personal self-formation, yet simultaneously it can also, via coordination pressures operating in a given structure, narrow the possible selves that are in practice accepted in a community. However, this is not to say that maximum freedom in self creation is a univocal, unambiguous good which policymakers should always strive to encourage. There is research that suggests that societies that get too close to *either* end of this spectrum – unlimited scope for self-formation or almost no scope – both suffer costs to general well-being (Harrington et al. 2015). Thus to maximise social welfare these two forces pulling in opposite directions need to be balanced.

5.3.4 Liberal Democratic Trade-offs

If we posit an ideal liberal democratic regime of the sort envisioned by Dewey (1963), one which aims to minimise suffering whilst maximising personal freedom, we can see that such political units must try to navigate a careful path between the various poles presented here. This is possible because unlike a governance regime guided by a classical conception of liberty, one which values negative freedom above all (Berlin 1969), a liberal democracy can intervene in public affairs to secure positive liberty.

Indeed, in some senses liberal democracies are more coercive than other potential or actual political regimes. They are coercive in the sense that such states seek to monitor and control their populations, to ensure that they comply with the various laws required to minimise suffering. So by one measure liberal democratic subjects are less free than in other political formations, a perspective which animated Foucault's work tracing the emergence of the modern state as an institution of control (2008). However, this is only one side of the coin: *public* unfreedom is balanced out by *private* freedom, itself directly enabled by the stability and the conflict-resolving institutions of the liberal democratic state. Many laws in liberal democratic societies restrict freedom, for example, speeding laws, wealth redistribution or food safety standards, but these same restrictions provide a stable basis from which to engage in imaginative self-creation whilst minimising potential harm to others.

Nation states must necessarily grapple with the trade-offs inherent in balancing liberty, fairness and coercion. The descriptive anti-individualism revealed by mindshaping and a liberal commitment to normative individualism pull in opposite directions in this context. Anti-individualism acknowledges that humans are collective, that structures beyond the individual agent determine values and that "free" choices are heavily influenced by social forces. Normative individualism says that although ontological individualism is false, because humans have selves we should treat individuals as if they were in practice free and thus maximise their freedom to create distinct selves and fully respect their autonomy. Neoliberal conceptions of liberal democracy tend to emphasise the freedom side of the trade-off just described; a position often implicitly informed by a commitment to ontological individualism that is rejected in the formulation of liberal democracy I use – a

rejection that is a direct outcome of accepting the mindshaping as lynchpin thesis. In the real world political decisions decide which aspect is given more weight in determining policy.

Because none of the potential outcomes described above are necessitated by the form of the technologies themselves they too require political scrutiny. The extent to which they are realised is the result of both intentional and unintentional design choices. Further complicating this picture is the fact that the outcomes not only relate on bipolar spectrums but also interact with one another. The realisation of an outcome on one scale may influence other spectrums. For example, the Chinese state maximises efficient coordination at the expense of diversity and free self-creation, whereas Western polities that are nominally subscribed to the ideals of liberal democracy sacrifice some degree of efficient coordination for the sake of other political virtues like individual liberty.

However, a significant danger faced by those polities with institutionally looser norm structures is that exogenous existential threats can make the coordination undermining trade-offs they make too costly. Effective coordination, even if driven by excessive and undemocratic coercion, can provide inarguable benefits through generating group level action, regardless of the costs to individual liberty (Gelfand et al. 2021). A key danger for existing (though flawed) liberal democratic states is that exogenous factors may destabilise their delicate balancing act, between maximising freedom and reducing suffering, in the medium to long term. In advance of such a state of affairs, emerging perhaps from geopolitical conflict or environmental catastrophe, reform of our communication infrastructure to make these polities more robust, through maximising internal coordination whilst remaining relatively open, is crucial.

We have seen how the structural design of our communications infrastructure may contribute to either the further flourishing or the potential dissolution of currently existing liberal democracies. However, reform in this domain represents relatively low-hanging fruit, in the sense of being directly amenable to policy intervention. Given that these affordances are necessarily mediated, and already deploy pervasive intervention to structure content to fulfil certain goals, their design provides clear scope for measures aiming to foster coordination while minimising inter-group strife. Furthermore, such redesign is crucial if we are to avoid pushing society into a state of large-scale, action undermining, coordination deadlock which may incentivise agents to abandon liberal democratic political systems in favour of more tightly coordinated, but repressive political ideologies of the type that have characterised much of human history.

5.4 Conclusion

This final section reiterates and crystallises the key points argued for across this thesis. Though the work covers a lot of terrain it is unified by a set of core arguments.

5.4.1 Politicising Philosophy of Mind

A central claim that implicitly informs my work here is that philosophy of mind and cognitive science are not neutral bodies of theory in the manner of the natural sciences[34]. This is because, as Dewey notes in the quote that opens this

[34] Maise and Hanna in *The Mind-Body Politic* (2019) propose a similar politicised approach to the philosophy of mind, and though their own emancipatory goals are aligned with my own, their explicitly phenomenological perspective, which focusses on the individual experiences of agents embedded in our particular political and economic juncture, requires their work to adopt a much richer normative framework than I believe is justified.

chapter, a description of how humans engage with the world and with others carries political ramifications. For one, any attempts to implement political reform ought to take into account the claims of our best theories of human cognition – hence my attempt to apply mindshaping theory to the reform of key public communication infrastructure. But additionally, our conceptions of how humans interact with, and in, the world also implies a specific picture of what ideal welfare-maximising social structures *should* look like.

For example, the previously dominant conception of human cognition, the functionalist computational paradigm examined in Chapter 2, which proposed a mindreading-based explanation of social cognition, incorporates clear, if unspoken, claims both about the ideal form of the political community and how to achieve such an end state. By conceptualising agents as atomistic observers embedded in fundamentally adversarial engagements aimed at discovering hidden mind states in others (Amadae 2016), it implicitly supports a social ontology and a particular governance regime that would most effectively allow such agents to prosper – one which values negative freedom over positive and places the locus of responsibility for outcomes on individual agents rather than social groups. That this was the general shape of the political orthodoxy during functional computationalism's apogee is unsurprising if we accept that philosophy of mind can have political implications or uses.

Thus, we can see the more or less simultaneous emergence and adoption of neoliberal governance techniques, the popularisation of the computationalist, language of thought paradigm, and the claim that game theory cashed out in wholly individualistic terms provided an accurate description of social interaction, as a set of broadly related phenomena (Amadae 2003, 2016). And though this conceptualisation

of human interaction has, as we have seen, often generated regressive political impulses, the valuable gains made for individual freedom across the 20th century, chiefly expanding suffrage and institutionalising the legal liberty of all agents in liberal democracies, are also arguably partly owed to the popularisation of atomistic understandings of mind, descending from Descartes, of which computational functionalism is a species.

However, though in philosophy this paradigm has been largely superseded by more interactionist, socially focused explanations, the political implications of this shift remain largely unexplored. This thesis has sought to begin such a task, taking a promising body of work in philosophy of mind and examining how its reconfiguration of our understanding of agents in the world can be used to directly inform the design of affordances those agents make use of in order to further broadly political goals. Specifically, mindshaping theory, as I have shown, makes concrete claims about how agents interact – forming aligned selves and coordinated groups – claims that I then show have implications for the design of communication infrastructure, if those structures are to facilitate more harmonious intergroup interaction.

Beyond political *uses*, the mindshaping framework, which understands human selves as socially constructed and historically contingent, also I believe, aligns with a specific set of political values and a conception of what an ideal polity should look like. Just as functional computationalism and mindreading aligned well with neoliberal forms of governance, the claims of mindshaping square up with, and indeed in my view imply, a particular political philosophy.

Specifically, the claims that mindshaping theory makes about the nature of social cognition and the formation of minds and groups dovetail with a political

position that has been called "Liberal Ironism" (Rorty 1989). This is a form of liberalism that is wedded to a deeply historicised conception of human selfhood. Specifically, it takes seriously the legacy of the Enlightenment as channelled through Nietzsche, namely that the content of cultural formations, knowledge claims and human self-descriptions are the products of contingent and ever-changing social forces. Such an understanding of the self and of human ontology as malleable and socially sourced shares important overlaps with mindshaping theory. This is unsurprising as Rorty's understanding of cognition is largely informed by the work of Dennett (Knobe 1995), which as we saw in Chapter 2, was also central to the development of mindshaping. To be clear, it is not that that the theory of mindshaping necessarily entails Rorty's species of Liberal Ironism, or vice-versa, but rather that these two stories share important common ground, and because Rorty's philosophy is explicitly politically engaged, it can serve as a useful basis for theorists attempting to apply mindshaping theory to political terrain.[35]

For Rorty, adopting a historicist lens reveals the absence of anything approaching a human political nature, as well as the impossibility of any ahistorical, context-free principle out there to be "discovered" that could guide or eternally normatively anchor human practices or fairness equilibria. Such a recognition of the fundamental contingency of human nature is not new, however it has generally

[35] Rorty is more strident about this link, indeed for him politics comes *before* philosophy. Thus a commitment to liberal democracy entails a certain philosophy. As he puts it, when describing his broadly mindshaping aligned claims in Contingency Irony and Solidarity: "the only argument I could give for the views about language and about selfhood put forward…was that these views seemed to cohere better with the institutions of a liberal democracy than the available alternatives do" (1989: 197). Additionally he interprets John Rawls's defence of liberalism as adopting a similar strategy: "…Rawls puts democratic politics first, and philosophy second. He retains the Socratic commitment to free exchange of views without the Platonic commitment to the possibility of universal agreement…He disengages the question of whether we ought to be tolerant and Socratic from the question of whether this strategy will lead to truth. He is content that it should lead to whatever intersubjective reflective equilibrium may be obtainable, given the contingent make-up of the subjects in question." (1991b: 191)

tended to be associated with illiberal or nihilistic political positions. For example, to Nietzsche this contingency was a cause for deep pessimism: there is no truth to be found in the world, only the will to power, and the subjugation of others that this inescapable condition entails. This same pessimism led Foucault, in his darkest moments, to conclude that the fundamental contingency of human social affairs means "that every social institution is equally unjustifiable, that all of them are on a par." (Rorty 1991a: 197). As a result liberal thinkers often attempt to sidestep the implications of this sort of thoroughgoing historicization, by appeals to universal or inalienable human rights, or transcendental moral principles in the Kantian tradition. In a similar fashion the core claims of mindshaping theory – that human ontology is pervasively structured by coordination pressures and selfhood is a product of contingent social forces – appear to rule out the possibility of discovering ahistorical truths about what makes a social structure better or worse. Thus, the theory of mindshaping can potentially undermine liberalism as a political position.

However, according to Rorty, rather than leading to Nietzschean nihilism or a relativist free-for-all, the absence of ahistorical or transcendental foundations is instead something for liberals to celebrate. This is because the lack of foundations revealed by a historicised perspective shows us that human communities are beholden to no external authority in determining their ideal state (Rorty 2021: 16). Contingency provides the grounds for ultimate liberty: human groups can determine *for themselves* what is of value. This malleability means that the search space for possible future fairness equilibria is maximally unconstrained. We can, in Ken Binmore's (2005) terms, move society to new, more egalitarian equilibria, realising a contingent understanding of fairness that can better meet the material needs of all.

Thus the claim that animates my work: a better understanding of the mechanisms and processes that guide and constrain mindshaping – the causal processes that make minds – can better equip humanity to make the most of its contingency, to democratically direct its future and with luck arm us with tools that can be used to block the (re)emergence of repressive, authoritarian and deeply illiberal social equilibria and instead bring about more egalitarian life-worlds.

Furthermore, the contingency of self revealed by a historicised, mindshaping-based, conception of human self-formation suggests that liberal democracy in its current incarnation is an unfinished project. As we have seen, contemporary liberal democracies have secured the negative liberty of their citizens – freedom from coercive authority (Berlin 1969). However, the legal recognition of individuals, and a solely negative conception of liberty, elides the necessarily *cultural basis* of the full realisation of liberty: individuality is an achievement, not a quality that is merely unleashed once the state recognises it in law (Dewey 1963). Such a view is central to mindshaping: individuality is a skill we learn through culture, individual human agents are evolved entities, responsive to cultural pressures – the liberal subject is shaped into being. Thus, the *formal* or *legal* liberty, which was the primary achievement of classic liberalism and is the main virtue of contemporary liberal democracies (at least for their naturalised citizens), must be matched by the realisation of *effective* liberty of thought and action for all, the realisation of which requires the reduction of material insecurity.[36] Material stability is a key requirement for human agents to be capable of fully expressing their individuality. In economic

[36] Dewey makes multiple references to socialising the means of production as the path to eliminating material insecurity (1963), perhaps placing his viewpoint closer to what we could call liberal-socialism. In this he reflects the thought current to his era: before the fall of the Soviet Union a fully socialised, centrally planned economy was widely thought to be a viable means of economic organisation, and indeed debates surrounding the viability of such a system have had a resurgence of late (for an overview of these trends, on both the left and right see Morozov (2019)).

terms, we can understand individuality as a type of luxury good that can only be widely produced once other more basic needs are adequately met.

One further crucial lesson provided by a historicised perspective on social progress, which understands human ontologies as ultimately contingent products of mindshaping, relates to the necessity of defending the gains made by liberalism. In other words, we must remember that it "*just happened* that rule in Europe passed into the hands of people who pitied the humiliated and dreamed of human equality, and that it may *just happen* that the world will wind up being ruled by people who lack any such sentiments or ideas. Socialization…goes all the way down, and who gets to do the socializing is often a matter of who manages to kill whom first." (Rorty 1989: 185-185, emphasis in original). Pointing to some authority or transcendental truth beyond intersubjective agreement, separate from the web of relations humans are embedded within, that proves that we should regard the poor as victims, not shirkers, or that torture is wrong, now and forever, is to miss the actually important task of enabling and encouraging individuals to expand their circle of concern to encompass larger and larger sets of agents whilst guarding against the encroachment of ideologies that dispute or reject such expansion.[37]

The fact that it just so happened that a given society fell into humane democratic liberalism is a call to arms. Liberals can never rest their belief in liberal democracy as the best system for governance yet invented on some necessary truth: liberalism is an achievement and an ongoing struggle, it must be protected, renewed, and developed. Thus, the crucial importance of foregrounding the political in social

[37] Binmore makes the same point in *Natural Justice*: "We retain what rights we have only because enough of us keep sufficient power in our collective hands that authoritarians are unable to take them away. Any propaganda that conceals this harsh reality is a danger to those of us who don't wish to live under oppressive regimes. I believe that we would do better to abandon all the rhetoric about inalienable natural rights—however effective it may be in the short run—lest we succeed in convincing our children that the price of freedom isn't eternal vigilance." (2005: 94)

science as Dewey urges, the gains in freedom made thus far may appear robust, yet as mindshaping theory reveals, they rest on no necessary foundations at all, and thus require explicit reaffirmation and ongoing defence if they are to be preserved. As we saw in the previous chapter, digital technologies, by pervasively structuring communication and thus partly determining how agents coordinate, may have increased the chances that such a regression could occur not due to pernicious individual actors or states, but rather as an inadvertent by-product of the specific structure of our communication environments. Deploying the theoretical apparatus of mindshaping to better explain how contingent human minds emerge and are shaped by such an environment can potentially help us avoid such a development.

5.4.2 Mindshaping and Coordinating

The central motivating claim of this work is that human beings have come to evolve rich, morally important and yet wholly contingent selves which function primarily as devices that aid interpersonal coordination. Their mindedness is partly a function of their social relations: without sociality and the cognitive tools installed by social processes, there would be no mature cognitive capacities of the sort that set humans apart from most other animals. The mindreading conception of social cognition gets the direction of this relationship backwards: humans agents don't come with minds installed and engage in attempts to discover the contents of other minds; instead human agents come to have minds through interpersonal mindshaping, which allows them to engage in ongoing dynamic interpersonal negotiation around mindstates. Humans don't read, but rather they co-create minds. This instrumental understanding of selfhood does not reduce selves to mere bookkeeping and behavioural control devices. Instead, its key contribution is to

better explain their evolution, and make clear their deeply social origin – regardless of their original evolutionary function, now that humans have selves they represent a significant locus of value.

It is the ongoing pervasive processes of mindshaping that resolve the mystery of how human communities can carry out joint action in the world, and successfully resolve coordination dilemmas without computationally intractable bouts of mindreading. Simultaneously, mindshaping explains why and, in providing specific processes, how, human communities tend to conform their actions to group-derived scripts, and as such provides a powerful window on the processes that can lead to inter-group conflict. Rational mindshapers can be placed in situations where they progressively shape one another to engage in out-group repression, just in order to reliably signal their in-group status.

The role of the social in structuring and guiding self-conceptions also reveals the deep contingency of individual selfhood. If the social makes the self, the specific character of one's social milieu will make a specific type of self. This both explains intercultural diversity in cognitive profiles and reinforces the claims made above about the political import of the theory of mindshaping. If socialisation goes all the way down, then who gets to do, or to guide, that socialization is a question of crucial importance. If our legitimate, albeit still insufficient, gains in reducing global suffering are to be maintained and extended, then the structures of socialisation need to be carefully designed, lest they end up incentivising the emergence of the types of selves that are less alive to the suffering of others. Fundamentally, mindshaping reveals the deeply political nature of our social structures and institutions, their form and affordances.

5.4.3 Evolving to Learn Strategically at Scale

The fact that mindshaping theory slots neatly into a range of mature approaches to explaining cultural and cognitive evolution provides strong support for its broader claims. Furthermore, the processes mindshaping relies upon, imitation, pedagogy and conformism turn out to be crucial components of cultural evolution. Viewing mindshaping through the lens of cultural evolution also brings into focus an element of this form of social cognition that is underexplored by Zawidzki and McGeer. Specifically, it reveals that mindshaping is guided by strategic goals. As such it is a type of strategic social learning, it does not operate blindly, copying the myriad models available in any social environment.

The key implication of this understanding of mindshaping is that our epistemological faculties are primarily social. People are not spectatorial scientists, sorting truth from fiction on the basis of sense data and trial and error experimentation, but rather expert readers of the social, coming in most non-functional cases to adopt beliefs that function as signals of their group status. What this reconceptualization of our epistemological faculties shows us is that the majority of our beliefs are strategic, and which specific beliefs humans come to adopt and espouse are determined in most cases by the nature of the group they understand themselves as part of. In particular, symbolic beliefs, beliefs about elements of the lifeworld which individuals have no direct knowledge of, or are fundamentally unknowable, come to serve a powerful signalling role in human affairs.

However, despite this invocation of strategy the agents involved are not making choices driven by consciously apprehended strategic concerns, but rather are making choices in the economic sense: their behaviour is responsive to incentives. Being a good group member provides benefits that adopting the beliefs of another

group would make unavailable, if one is to remain in the initial group. But this is not to say that a group member strategically calculates at all times the benefits of remaining in one group as opposed to joining another. Most of the time they are just behaving as a good group member. Indeed it is this lack of high-level strategic awareness of the outcomes of one's specific actions that makes one an effective and valued group member, or coordination partner.

However if the incentives change for the group as a whole, say some specific component of their ontology becomes very costly, as, for example, polygamy did for Mormon communities in the 1890s, then the group can rapidly abandon such beliefs, as the Mormons did en masse. What this way of understanding mindshaping reveals is the importance of paying attention to incentives in intergroup and individual-group interaction, incentives that make specific beliefs, or types of beliefs more or less appealing.

Secondly, this focus on the role of strategic social belief formation serves to highlight the role played by communication technologies in mindshaping, and in particular the role played by the structure of those technologies in making certain types and scales of groups possible. Effectively the scale of human communities, or the scale at which mindshaping can operate is constrained by technologies of communication. Furthermore, the specific form of those technologies makes different types of communities possible, determining in part the topology of those communities and the types of incentives they face in belief formation. As a result, the various major transitions in social organisation across the evolutionary history of *Homo sapiens* have been closely tied to developments in communication technologies of this sort, and likewise, the specific political structures that were most prevalent in those different epochs in turn relate to those technologies. Thus the

invention of writing played a crucial role in the emergence of early large-scale urban settlements due to the manner in which it could encode, centralise and disseminate symbolic signals to increasingly large groups, whilst ensuring a specific type of hierarchical social organisation.

5.4.4 Digital Communication Technology: Putting it All Together

Given this evolutionary trajectory, linking the form of communication technologies to possible human self-conceptions and groups, we can see how digital communication technology presents itself as a crucial new component of our life world. It is for this reason, coupled with the explicitly political and applied stance of my work, that I have sought to use the expanded theory of mindshaping presented above to shed light on contemporary online communication and in particular the crisis in public discourse generated by this new communication domain. A key claim of Chapter 4 is that current, implicitly atomistic and mindreading-friendly, approaches to resolving these dysfunctions are misguided, and that an understanding of human social cognition and epistemology informed by mindshaping theory can be used to provide a more promising set of recommendations for the redesign of these affordances.

Specifically, the role of in-group signalling in determining the content of beliefs, and in particular the crucial role of honest signals in forming and maintaining groups, can be used to understand how this new communication environment generates problematic incentives for agents. In particular, I have shown how specific communication incentives present on social media in their current form may push public discourse to assume an increasingly uncivil and zero-sum form, as agents rationally adopt more extreme, and out-group alienating, positions on a range

of topics. This core issue is exacerbated by a range of structural features present in contemporary online communication environments. Echo chambers, content biases, the prevalence of threat-related information online, the general collapse in context or loss of clearly delineated informational domains, and the perverse incentives faced by the communication platforms themselves are all components in the generation of a toxic communication environment, one that mindshapes agents to adopt pernicious belief sets.

However though the contemporary structure of these technologies generates problematic outcomes, the very fact that these technologies have such a significant effect on discourse is not itself necessarily a cause for pessimism. The baleful effects of contemporary communication arise largely as the result of inadvertent or perversely incentivised design. The necessarily complex and mediated form of these technologies means that they can be structured in new ways, ways more conducive to enabling civil and constructive intergroup dialogue, and potentially an expansion of more benign and inclusive cultural formations. Given this potentiality, I presented several recommendations in Chapter 4 for how the mindshaping-based approach to explaining belief formation can replace the prior and still dominant mindreading-based conception and provide a promising, and empirically assessable, means to better design these infrastructures to minimise the more problematic outcomes they currently give rise to.

However a potential opportunity is not a necessary outcome, and, to actually effect a change in the design of digital communication, the incentives faced by the providers of these services too must be changed. Problematically, however, the political and economic environment in which these services emerged makes such reform difficult. The free market ethos that animates the dominant technology

companies, and their trenchant resistance to anti-trust law, or even the status of content publishers means that any attempts to regulate the sector face an uphill battle. Furthermore, due to technological momentum, the influence of neoliberal ideology on the early design of the internet, and specifically communication platforms, may have locked in a particularly problematic form of these affordances for the foreseeable future.

Nevertheless, if the incentives faced by the providers can be modified through regulation, or if the existing forms are superseded by new, better designed, technologies, then digital communication holds out great promise for humanity, making Dewey's liberal democratic vision, in which suffering is maximally reduced and personal freedom maximised, potentially more attainable.

However, this broader liberal vision notwithstanding, the imminent material scarcity that is likely to impact a significant portion of the world's citizens in the near future due to the effects of anthropogenic climate change makes the reform of our communication infrastructure particularly urgent. Even if the major world powers can agree to binding emission reductions targets in the near future, itself an unlikely prospect for now, as it stands the already existing concentrations of greenhouse gases in the atmosphere will warm the climate enough (Marotzke et al. 2022) to make some currently densely populated areas of the planet more or less uninhabitable, and place severe stresses on global food production and supply (Masson-Delmotte et al. 2018). If these factors are not to lead to large-scale violence then, at the very least, the technologies people use to communicate must maximally strive to promote inter-group dialogue and minimise the use of extreme out-group discourse as a means of signalling group affiliation.

References

Abramowitz, A., & McCoy, J. (2019). United States: Racial resentment, negative partisanship, and polarization in Trump's America. *The ANNALS of the American Academy of Political and Social Science, 681*(1), 137–156. https://doi.org/10.1177/0002716218811309

Acemoglu, D., Ozdaglar, A., & Siderius, J. (2021). Misinformation: Strategic sharing, homophily, and endogenous echo chambers. *NBER Working Paper 28884.*

Acerbi, A. (2019). Cognitive attraction and online misinformation. *Palgrave Communications*, *5*(1), 1–7. https://doi.org/10.1057/s41599-019-0224-y

Acerbi, A., & Mesoudi, A. (2015). If we are all cultural Darwinians what's the fuss about? Clarifying recent disagreements in the field of cultural evolution. *Biology & Philosophy*, *30*(4), 481–503.

Akbari, M., Bahadori, M. H., Khanbabaei, S., Milan, B. B., Horvath, Z., Griffiths, M. D., & Demetrovics, Z. (2023). Psychological predictors of the co-occurrence of problematic gaming, gambling, and social media use among adolescents. *Computers in Human Behavior*, *140*, 107589. https://doi.org/10.1016/j.chb.2022.107589

Alfano, M., Fard, A. E., Carter, J. A., Clutton, P., & Klein, C. (2021). Technologically scaffolded atypical cognition: The case of YouTube's recommender system. *Synthese*, *199*(1), 835–858. https://doi.org/10.1007/s11229-020-02724-x

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

Almaatouq, A., Becker, J. A., Bernstein, M., Botto, R., Bradlow, E., Damer, E., Duckworth, A. L., Griffiths, T., Hartshorne, J. K., Law, E., Lazer, D., Liu, M., Matias, J. N., Rand, D. G., Salganik, M., Satlof-Bedrick, E., Schweitzer, M.,

Shirado, H., Suchow, J. W., … Yin, M. (2021). Scaling up experimental social, behavioral, and economic science [Preprint]. *Open Science Framework*. https://doi.org/10.31219/osf.io/wksv8

Alston, P. (2020). The parlous state of poverty eradication: Report of the Special Rapporteur on extreme poverty and human rights. *Human Rights Council, Forty-Fourth Session*.

Altay, S., de Araujo, E., & Mercier, H. (2022). "If This account is True, It is Most Enormously Wonderful": Interestingness-if-true and the sharing of true and false news. *Digital Journalism*, *10*(3), 373–394. https://doi.org/10.1080/21670811.2021.1941163

Aly, G. (2007). *Hitler's Beneficiaries: How the Nazis Bought the German People*. Verso Books.

Amadae, S. M. (2003). *Rationalizing Capitalist Democracy*. University of Chicago Press.

Amadae, S. M. (2016). *Prisoners of Reason: Game Theory and Neoliberal Political Economy*. Cambridge University Press.

Anderson, B. (1983). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso Books.

Andre, J., & Morin, O. (2011). Questioning the cultural evolution of altruism. *Journal of Evolutionary Biology*, *24*(12), 2531–2542.

Andrews, K. (2015). The folk psychological spiral: Explanation, regulation, and language. *The Southern Journal of Philosophy*, *53*, 50–67. https://doi.org/10.1111/sjp.12121

Asch, S. E. (1955). Opinions and social pressure. *Scientific American*, *193*(5), 31–35.

Atari, M., Davani, A. M., Kogon, D., Kennedy, B., Ani Saxena, N., Anderson, I., & Dehghani, M. (2021). Morally homogeneous networks and radicalism. *Social*

Psychological and Personality Science, *13*(6), 999–1009. https://doi.org/10.1177/19485506211059329

Atran, S., & Henrich, J. (2010). The evolution of religion: How cognitive by-products, adaptive learning heuristics, ritual displays, and group competition generate deep commitments to prosocial religions. *Biological Theory*, *5*(1), 18–30. https://doi.org/10.1162/BIOT_a_00018

Aumann, R. J. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, *1*, 67–96.

Aumann, R. J. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometricia*, *55*, 1–18.

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology, 149*(8), 1608–1613. https://doi.org/10.1037/xge0000729

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 3–8.

Bailey, R., & Misra, P. (2022). Interoperability of social media: An appraisal of the regulatory and technical ecosystem. *Available at SSRN*. https://doi.org/10.2139/ssrn.4095312

Bala, V., & Goyal, S. (1998). Learning from neighbours. *Review of Economic Studies*, *65*(3), 595–621. https://doi.org/10.1111/1467-937X.00059

Bar-Yosef, O. (2007). The archaeological framework of the Upper Paleolithic Revolution. *Diogenes*, *54*(2), 3–18. https://doi.org/10.1177/0392192107076869

Baranowski-Pinto, G., Profeta, V. L. S., Newson, M., Whitehouse, H., & Xygalatas, D. (2022). Being in a crowd bonds people via physiological synchrony. *Scientific Reports*, *12*(613). https://doi.org/10.1038/s41598-021-04548-2

Bargh, J. M., Chen, M., & Burrows, L. (1996). The automaticity of social behavior: Direct effects of trait concept and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244.

Barkow, J. H., Cosmides, L., & Tooby, J. (Eds.). (1992). *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* Oxford University Press.

Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press.

Bebbington, K., MacLeod, C., Ellison, T. M., & Fay, N. (2017). The sky is falling: Evidence of a negativity bias in the social transmission of information. *Evolution and Human Behavior*, *38*(1), 92–101. https://doi.org/10.1016/j.evolhumbehav.2016.07.004

Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, *30*(3), 141–164. https://doi.org/10.1257/jep.30.3.141

Benkler, Y. (2020). A Political Economy of the Origins of Asymmetric Propaganda in American Media. In W. Bennett & S. Livingston (Eds.), *The Disinformation Age* (pp. 43–66). Cambridge University Press.

Bennett, J. (1978). Some remarks about concepts. *Behavioral and Brain Sciences*, *1*(4), 557–560.

Bennett, W., & Livingston, S. (2020a). A Brief History of the Disinformation Age: Information Wars and the Decline of Institutional Authority. In W. Bennett & S. Livingston (Eds.), *The Disinformation Age* (pp. 3–40). Cambridge University Press.

Bennett, W., & Livingston, S. (Eds.). (2020b). *The Disinformation Age*. Cambridge University Press.

Benoit, S. L., & Mauldin, R. F. (2021). The "anti-vax" movement: A quantitative report on vaccine beliefs and knowledge across social media. *BMC Public Health*, *21*(2106). https://doi.org/10.1186/s12889-021-12114-8

Bentley, J. W. (2020). Improving the statistical power and reliability of research using Amazon Mechanical Turk. *SSRN Scholarly Paper No. 2924876*. https://doi.org/10.2139/ssrn.2924876

Berlin, I. (1969). *Four Essays on Liberty*. Oxford University Press.

Bermúdez, J. L. (2003). The domain of folk psychology. *Royal Institute of Philosophy Supplements*, *53*, 25–48.

Bhargava, V. R., & Velasquez, M. (2021). Ethics of the attention economy: The problem of social media addiction. *Business Ethics Quarterly*, *31*(3), 321–359. https://doi.org/10.1017/beq.2020.32

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Biebricher, T. (2018). *The Political Theory of Neoliberalism*. Stanford University Press.

Biebricher, T. (2020). Neoliberalism and authoritarianism. *Global Perspectives*, *1*(1), 11872. https://doi.org/10.1525/001c.11872

Binmore, K. (2005). *Natural Justice*. Cambridge University Press.

Birch, K., & Bronson, K. (2022). Big tech. *Science as Culture*, *31*(1), 1–14. https://doi.org/10.1080/09505431.2022.2036118

Blaine, T., & Boyer, P. (2018). Origins of sinister rumors: A preference for threat-related material in the supply and demand of information. *Evolution and Human Behavior*, *39*(1), 67–75. https://doi.org/10.1016/j.evolhumbehav.2017.10.001

Bolhuis, J. J., & Wynne, C. D. L. (2009). Can evolution explain how minds work? *Nature*, *458*(7240). https://doi.org/10.1038/458832a

Borbáth, E., Hutter, S., & Leininger, A. (2023). Cleavage politics, polarisation and participation in Western Europe. *West European Politics*, *46*(4), 631–651. https://doi.org/10.1080/01402382.2022.2161786

Bork, R. (1978). *The Antitrust Paradox.* Free Press.

Bourne, R. (2019). Is this time different? Schumpeter, the tech giants, and monopoly fatalism. *Cato Institute Policy Analysis*, *872*, 1–34.

Boyd, R., & Richerson, P. J. (1985). *Culture and the Evolutionary Process*. University of Chicago Press.

Boyd, R., & Richerson, P. J. (1988). An Evolutionary Model of Social Learning: The Effects of Spatial and Temporal Variation. In T. R. Zentall & B. G. Galef, Jr. (Eds.), *Social Learning: Psychological and Biological Perspectives* (pp. 29–48). Lawrence Erlbaum Associates.

Boyer, P., & Parren, N. (2015). Threat-related information suggests competence: a possible factor in the spread of rumors. *PLOS ONE*, *10*(6), e0128421. https://doi.org/10.1371/journal.pone.0128421

Brady, W. J., & Crockett, M. J. (2019). How effective is online outrage?. *Trends in Cognitive Sciences*, *23*(2).

Brooks, A. S., Yellen, J. E., Potts, R., Behrensmeyer, A. K., Deino, A. L., Leslie, D. E., Ambrose, S. H., Ferguson, J. R., d'Errico, F., Zipkin, A. M., Whittaker, S., Post, J., Veatch, E. G., Foecke, K., & Clark, J. B. (2018). Long-distance stone transport and pigment use in the earliest Middle Stone Age. *Science*, *360*(6384), 90–94. https://doi.org/10.1126/science.aao2646

Brown, W. (2015). *Undoing the Demos: Neoliberalism's Stealth Revolution*. MIT Press.

Brown, W. (2019). *In the Ruins of Neoliberalism*. Columbia University Press.

Buller, D. (2005). *Adapting Minds: Evolutionary Psychology and the Persistent Quest for Human Nature*. MIT Press.

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, *28*(5), 606–637. https://doi.org/10.1111/mila.12036

Byars, S. G., Ewbank, D., Govindaraju, D. R., & Stearns, S. C. (2010). Natural selection in a contemporary human population. *Proceedings of the National Academy of Sciences*, *107*, 1787–1792. https://doi.org/10.1073/pnas.0906199106

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen- through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development*, *21*(2), 315–330.

Case, A., & Deaton, A. (2020). *Deaths of Despair and the Future of Capitalism*. Princeton University Press.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton University Press.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology*, *113*(3), 702–734. https://doi.org/10.1086/521848

Chalmers, D. J. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. Allen Lane.

Chartrand, T., & Bargh, J. (1999). The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893–910.

Chater, N. (2018). *The Mind is Flat.* Allen Lane.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A., & Cutler, D. (2016). The association between income and life expectancy in the United

States, 2001–2014. *JAMA*, *315*(16), 1750–1766. https://doi.org/10.1001/jama.2016.4226

Chomsky, N. (1959). A review of B. F. Skinner's verbal behavior. *Language*, *35*, 26–58.

Christensen, D. (2007). Epistemology of disagreement: The good news. *The Philosophical Review*, *116*(2), 187–217.

Chwe, M. S. Y. (2001). *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton University Press.

Clare, L. (2020). Göbekli Tepe, Turkey: a brief summary of research at a new World Heritage Site (2015–2019). *e-Forschungsberichte*, § 1–13. https://doi.org/10.34780/efb.v0i2.1012

Clark, A. (1994). Beliefs and desires incorporated. *Journal of Philosophy*, *91*(8), 404–425.

Clark, A. (1997). *Being There.* MIT Press

Clark, A. (2003). *Natural-born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford University Press.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

Clarke, E., & Heyes, C. (2017). The swashbuckling anthropologist: Henrich on The Secret of Our Success. *Biology & Philosophy*, *32*(2), 289–305. https://doi.org/10.1007/s10539-016-9554-y

Coady, C. A. J. (1992). *Testimony: A Philosophical Study*. Clarendon Press.

Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–198. https://doi.org/10.1145/2959100.2959190

Csibra, G., & Gergely, G. (2006). Social Learning and Social Cognition: The Case for Pedagogy. In Y. Munakata & M. H. Johnson (Eds.), *Processes of Change in Brain and Cognitive Development* (pp. 249–274). Oxford University Press.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153.

Cusumano, M. A. (2021). Section 230 and a tragedy of the commons. *Communications of the ACM*, *64*(10), 16–18. https://doi.org/10.1145/3481354

d'Errico, F. (2007). The origin of humanity and modern cultures: Archaeology's view. *Diogenes*, *54*(2), 122–133. https://doi.org/10.1177/0392192107077652

Davies, M., & Stone, T. (1995). *Folk Psychology: The Theory of Mind Debate.* Blackwell.

Davies, W. (2016). *The Limits of Neoliberalism: Authority, Sovereignty and the Logic of Competition*. Sage.

De Jaegher, H. (2009). Social understanding through direct perception? Yes, by interacting. *Consciousness and Cognition*, 18(2), 535–542.

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, *1*(4), 568–570.

Dennett, D. C. (1987). *The Intentional Stance*. MIT press.

Dennett, D. C. (1991a). *Consciousness Explained.* Little, Brown and Co.

Dennett, D. C. (1991b). Real patterns. *The Journal of Philosophy*, *88*(1), 27–51.

Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. Allen Lane.

Dewey, J. (1963). *Liberalism and Social Action*. Capricorn Books.

DiRusso, C., & Stansberry, K. (2022). Unvaxxed: a cultural study of the online anti-vaccination movement. *Qualitative Health Research*, *32*(2), 317–329. https://doi.org/10.1177/10497323211056050

Dunbar, R. (1993). Co-evolution of neocortex size, group size and language in humans. *Behavioral and Brain Sciences*, *16*, 681–735.

Dunn, B. (2016). Against neoliberalism as a concept. *Capital and Class 41*(3), 435–54.

Dupre, J. (2012). Against Maladaptationism: Or, What's Wrong with Evolutionary Psychology?. In J. Dupre (Ed.), *Processes of Life: Essays in Philosophy of Biology* (pp. 245–260). Oxford University Press.

Eisenstein, E. L. (1980). *The Printing Press as an Agent of Change*. Cambridge University Press. https://doi.org/10.1017/CBO9781107049963.011

Enders, A. M., & Armaly, M. T. (2019). The differential effects of actual and perceived polarization. *Political Behavior*, *41*(3), 815–839. https://doi.org/10.1007/s11109-018-9476-2

Farina, M. (2013). The evolved apprentice. How evolution made humans unique. *Phenomenology and the Cognitive Sciences*, *12*(4), 915–923.

Feher, M. (2009). Self-appreciation; or, the aspirations of human capital. *Public Culture*, *21*(1), 21–41.

Fenici, M. (2015). A simple explanation of apparent early mindreading: Infants' sensitivity to goals and gaze direction. *Phenomenology and the Cognitive Sciences*, *14*(3), 497–515.

Fenici, M. (2017). Rebuilding the landscape of psychological understanding after the mindreading war. *Phenomenology and Mind*, *12*, 142–150.

Fenici, M., & Zawidzki, T. W. (2021). The origins of mindreading: How interpretive socio-cognitive practices get off the ground. *Synthese*, *198*(9), 8365–8387. https://doi.org/10.1007/s11229-020-02577-4

Fisher, M. (2009). *Capitalist Realism: Is There No Alternative?* John Hunt Publishing.

Floridi, L. (2014). *The 4th Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press.

Floridi, L. (2015). The new grey power. *Philosophy & Technology*, *28*(3), 329–332. https://doi.org/10.1007/s13347-015-0206-y

Floridi, L. (2021). The end of an era: From self-regulation to hard law for the digital industry. *Philosophy & Technology*, *34*(4), 619–622. https://doi.org/10.1007/s13347-021-00493-0

Fodor, J. (1975). *The Language of Thought*. Harvard University Press.

Fodor, J. (1980). Methodological solipsism considered as a research strategy in cognitive science. *Behavioral and Brain Sciences*, *3*, 63–73.

Fodor, J. (1983). *The Modularity of Mind*. MIT Press.

Foucault, M. (2008). *The Birth of Biopolitics: Lectures at the Collège de France, 1978–1979*. Palgrave Macmillan.

Franks, B. (2011). *Culture and Cognition: Evolutionary Perspectives*. Palgrave Macmillan.

Franssen, M., & Koller, S. (2016). Philosophy of Technology as a Serious Branch of Philosophy: The Empirical Turn as a Starting Point. In Franssen, M., Vermaas, P. E., Kroes, P., & Meijers, A. W. (Eds), *Philosophy of Technology After the Empirical Turn* (pp. 31–60). Springer.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Frischmann, B., & Selinger, E. (2018). *Re-engineering Humanity*. Cambridge University Press.

Funkhouser, E. (2022). A tribal mind: Beliefs that signal group identity or commitment. *Mind & Language, 37*(3), 444–464. https://doi.org/10.1111/mila.12326

Gallagher, S. (2004). Understanding interpersonal problems in autism: Interaction theory as an alternative to theory of mind. *Philosophy, Psychiatry, & Psychology*, *11*(3), 199–217.

Gallagher, S. (2008). Direct perception in the intersubjective context. *Consciousness and Cognition*, *17*(2), 535–543.

Gelfand, M. J. (2019). *Rule Makers, Rule Breakers: How Tight and Loose Cultures Wire Our World*. Scribner.

Gelfand, M. J., & Lun, J. (2013). Ecological priming: Convergent evidence for the link between ecology and psychological processes. *Behavioral and Brain Sciences*, *36*(5), 489–490.

Gelfand, M. J., Harrington, J. R., & Jackson, J. C. (2017). The strength of social norms across human groups. *Perspectives on Psychological Science*, *12*(5), 800–809.

Gelfand, M. J., Jackson, J. C., Pan, X., Nau, D., Pieper, D., Denison, E., Dagher, M., Van Lange, P. A. M., Chiu, C.-Y., & Wang, M. (2021). The relationship between cultural tightness–looseness and COVID-19 cases and deaths: A global analysis. *The Lancet Planetary Health*, *5*(3), e135–e144. https://doi.org/10.1016/S2542-5196(20)30301-6

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., & Arnadottir, J. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, *332*(6033), 1100–1104.

Gergely, G. (2011). Kinds of Agents: The Origins of Understanding Instrumental and Communicative Agency. In U. Goshwami (Ed.), *The Wiley-Blackwell Handbook of Childhood Cognitive Development* (pp. 76–105). Wiley-Blackwell.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences, 7*, 278–292.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*, 755–756.

Germani, F., & Biller-Andorno, N. (2021). The anti-vaccination infodemic on social media: A behavioral analysis. *PLOS ONE*, *16*(3), 1–14. https://doi.org/10.1371/journal.pone.0247642

Germar, M., & Mojzisch, A. (2019). Learning of social norms can lead to a persistent perceptual bias: A diffusion model approach. *Journal of Experimental Social Psychology, 84*, 103801.

Gershon, R. A. (2013). *The Transnational Media Corporation: Global Messages and Free Market Competition*. Routledge.

Gerstle, G. (2022). *The Rise and Fall of the Neoliberal Order: America and the World in the Free Market Era*. Oxford University Press.

Gethin, A., & Morgan, M. (2018). Brazil divided: Hindsights on the growing politicisation of inequality. *WID.World Issue Brief*, *2018*(3), 1–8.

Geva, D. (2021). Orbán's ordonationalism as post-neoliberal hegemony. *Theory, Culture & Society*, *38*(6), 71–93. https://doi.org/10.1177/0263276421999435

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gnanadesikan, A. E. (2008). *The Writing Revolution: Cuneiform to the Internet*. Wiley-Blackwell.

Goldman, A. (2006). *Simulating Minds*. Oxford University Press.

Goldman, A., & O'Connor, C. (2021). Social Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University.

Golman, R., Loewenstein, G., Moene, K. O., & Zarri, L. (2016). The preference for belief consonance. *Journal of Economic Perspectives, 30*(3), 165–188. https://doi.org/10.1257/jep.30.3.165

Gordon, R. (1986). Folk psychology as simulation. *Mind and Language 1*, 158–171.

Graeber, D., & Wengrow, D. (2021). *The Dawn of Everything: A New History of Humanity*. Penguin.

Graham, G. (1993). The Origins of Folk Psychology. In S. Christensen and D. Turner (Eds.), *Folk Psychology and the Philosophy of Mind* (200–220). Lawrence Erlbaum.

Griffiths, J. (2019). *The Great Firewall of China: How to Build and Control an Alternative Version of the Internet*. Zed Books.

Grzanka, P., Mann, E., & Sinikka, E. (2016). The neoliberalism wars, or notes on the persistence of neoliberalism. *Sexuality Research and Social Policy 13*(4), 297–307.

Guala, F. (2016). *Understanding Institutions: The Science and Philosophy of Living Together*. Princeton University Press.

Guala, F. (2018). Coordination, team reasoning, and solution thinking. *Revue d'Economie Politique*, *128*(3), 355–372.

Guala, F. (2020). Solving the Hi-lo Paradox: Equilibria, Beliefs, and Coordination. In A. Fiebich (Ed.), *Minimal Cooperation and Shared Agency* (pp. 149–168). Springer.

Guala, F., & Mittone, L. (2010). How history and convention create norms: An experimental study. *Journal of Economic Psychology*, *31*(4), 749–756.

Guess, A. M., Lockett, D., Lyons, B., Montgomery, J. M., Nyhan, B., & Reifler, J. (2020). "Fake news" may have limited effects beyond increasing beliefs in false claims. *Harvard Kennedy School Misinformation Review*, *1*(1). https://doi.org/10.37016/mr-2020-004

Gupta, A. K. (2004). Origin of agriculture and domestication of plants and animals linked to early Holocene climate amelioration. *Current Science*, *87*(1), 54–59.

Hameleers, M., & Van der Meer, T. G. (2020). Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers?. *Communication Research*, *47*(2), 227–250.

Hardin, G. (1995). *One For All: The Logic of Group Conflict.* Princeton University Press.

Harrington, J. R., Boski, P., & Gelfand, M. J. (2015). Culture and national well-being: Should societies emphasize freedom or constraint? *PLOS ONE*, *10*(6), 1–14. https://doi.org/10.1371/journal.pone.0127173

Harrison, G. W., Monroe, B. & Ulm, E. R. (2022). Recovering subjective probability distributions: A Bayesian approach. *CEAR Working Paper* 2022-03.

Harrison, G. W., & Ross, D. (2016) The psychology of human risk preferences and vulnerability to scare-mongers: Experimental economic tools for hypothesis formulation and testing. *Journal of Cognition and Culture* 16, 383–414.

Harsanyi, J. (1967). Games with incomplete information played by 'Bayesian' players, parts I-III. *Management Science*, 14, 159–182.

Harvey, D. (2005). *A Brief History of Neoliberalism*. Oxford University Press.

Haslanger, S. (2019). Cognition as a social skill. *Australasian Philosophical Review*, *3*(1), 5–25. https://doi.org/10.1080/24740500.2019.1705229

Hauser, D., Paolacci, G., & Chandler, J. (2019). Common concerns with MTurk as a participant pool: Evidence and solutions. In F. Kardes, P. M. Herr, & N. Schwarz

(Eds.), *Handbook of Research Methods in Consumer Psychology* (pp. 319–337). Routledge.

Hayek, F. A. (1944). *The Road to Serfdom.* Chicago University Press.

Heintz, C. (2018). Cultural Attraction Theory. In H. Callan (Ed.), *The International Encyclopedia of Anthropology*. Wiley. https://doi.org/10.1002/9781118924396.wbiea2311

Hendricks, V. F., & Vestergaard, M. (2019). *Reality Lost: Markets of Attention, Misinformation and Manipulation*. Springer Nature.

Henrich, J. (2016). *The Secret of Our Success: How Culture is Driving Human Evolution, Domesticating Our Species and Making Us Smarter.* Princeton University Press.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., … Tracer, D. (2005). "Economic Man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*, 795–855.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*, 1767–1769.

Henshilwood, C. S., d'Errico, F., van Niekerk, K. L., Dayet, L., Queffelec, A., & Pollarolo, L. (2018). An abstract drawing from the 73,000-year-old levels at Blombos Cave, South Africa. *Nature*, *562*(7725), 115–118. https://doi.org/10.1038/s41586-018-0514-3

Herman, E. S., & Chomsky, N. (1988). *Manufacturing Consent: The Political Economy of the Mass Media*. Pantheon Books.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, *317*(5843), 1360–1366.

Herrmann, E., Hernández-Lloreda, M. V., Call, J., Hare, B., & Tomasello, M. (2010). The structure of individual differences in the cognitive abilities of children and chimpanzees. *Psychological Science*, *21*(1), 102–110.

Hewlett, B. S., Fouts, H. N., Boyette, A. H., & Hewlett, B. L. (2011). Social learning among Congo Basin hunter–gatherers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1168–1178. https://doi.org/10.1098/rstb.2010.0373

Heyes, C. (2011). Automatic imitation. *Psychological Bulletin*, *137*(3), 463.

Heyes, C. (2012). New thinking: The evolution of human cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2091–2096. https://doi.org/10.1098/rstb.2012.0111

Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, *9*(2), 131–143.

Heyes, C. (2016). Blackboxing: Social learning strategies and cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1693). https://doi.org/10.1098/rstb.2015.0369

Heyes, C. (2017). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Harvard University Press.

Heyes, C. (2018). Enquire within: Cultural evolution and cognitive science. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1743).

Hickman, L.A. (1990). *John Dewey's Pragmatic Technology*. Indiana University Press.

Hickman, L.A. (2017). Dewey, Pragmatism, Technology. In S. Fesmire (Ed.) *The Oxford Handbook of Dewey* (pp. 491–506). Oxford University Press.

Hirschman, A. O. (1970). *Exit, Voice, and Loyalty*. Harvard University Press.

Hochschild, J. L., & Einstein, K. L. (2015). *Do Facts Matter? Information and Misinformation in American Politics*. University of Oklahoma Press.

Hodder, I. (2007). Çatalhöyük in the context of the middle eastern Neolithic. *Annual Review of Anthropology*, *36*, 105–120. https://doi.org/10.1146/annurev.anthro.36.081406.094308

Huang, H. (2015). Propaganda as signaling. *Comparative Politics*, *47*(4), 419–437.

Hughes, T. (1994). Technological Momentum. In M. R. Smith and L. Marx (Eds.), *Does Technology Drive History?: The Dilemma of Technological Determinism* (pp. 101–113). Massachusetts Institute of Technology.

Humphrey, N. K. (1976). The Social Function of Intellect. In P. P. G. Bateson & R. A. Hinde (Eds.), *Growing Points in Ethology* (pp. 303–317). Cambridge University Press.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

Hutto, D. D. (2008). *Folk Psychological Narratives*. MIT Press.

Hutto, D. D. (2009). ToM Rules, but it is Not OK! In I. Leudar & A. Costall (Eds.), *Against Theory of Mind* (pp. 221-238). Palgrave Macmillan.

Hutto, D. D., Herschbach, M., & Southgate, V. (2011). Social cognition: mindreading and alternatives. *Review of Philosophy and Psychology*, *2*(3), 375–395.

Iamamoto, S. A. S., Mano, M. K., & Summa, R. (2021). Brazilian far-right neoliberal nationalism: Family, anti-communism and the myth of racial democracy. *Globalizations*, *20*(5), 1–17. https://doi.org/10.1080/14747731.2021.1991745

Ibáñez, J. J., Ortega, D., Campos, D., Khalidi, L., & Méndez, V. (2015). Testing complex networks of interaction at the onset of the near eastern Neolithic using modelling of obsidian exchange. *Journal of The Royal Society Interface*, *12*(107), 20150210.

Jagiello, R., Heyes, C., & Whitehouse, H. (2022). Tradition and invention: The bifocal stance theory of cultural evolution. *Behavioral and Brain Sciences*, *45*, 1–18. https://doi.org/10.1017/S0140525X22000383

Jarvis, J. (2011). *Public Parts: How Sharing in the Digital Age Improves the Way We Work and Live*. Simon & Schuster.

Johnson, S. (2013). *Future Perfect: The Case For Progress in A Networked Age*. Penguin.

Joshi, H. (2022). Debunking creedal beliefs. *Synthese*, *200*(6), 514. https://doi.org/10.1007/s11229-022-03991-6

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising*, *46*(1), 141–155. https://doi.org/10.1080/00913367.2016.1269304

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social learning strategies: Bridge-building between fields. *Trends in Cognitive Sciences*, *22*(7), 651–665. https://doi.org/10.1016/j.tics.2018.04.003

Keulartz, J., Schermer, M., Korthals, M., & Swierstra, T. (2004). Ethics in technological culture: A programmatic proposal for a pragmatist approach. *Science, Technology, & Human Values*, *29*(1), 3–29. https://doi.org/10.1177/0162243903259188

Kiely, R. (2018). *The Neoliberal Paradox*. Edward Elgar Publishing.

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

King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, *111*(3), 484–501. https://doi.org/10.1017/S0003055417000144

Kiper, J. (2022). Remembering the causes of collective violence and the role of propaganda in the Yugoslav wars. *Nationalities Papers*, 1–24. https://doi.org/10.1017/nps.2022.53

Kirschner, S., & Tomasello, M. (2010). Joint music making promotes prosocial behavior in 4-year-old children. *Evolution and Human Behavior*, *31*(5), 354–364.

Klaehn, J. (Ed.). (2010). *The Political Economy of Media and Power*. Peter Lang.

Knobe, J. (1995). A talent for bricolage: An interview with Richard Rorty. *The Dualist, 2,* 56–71.

Kuran, T. (1997). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press.

Larrouy, L., & Lecouteux, G. (2018). Choosing in a large world: The role of focal points as a mindshaping device. *GREDEG Working Paper, No. 2018-29.*

Lavelle, J. S. (2021). The impact of culture on mindreading. *Synthese*, *198*(7), 6351–6374. https://doi.org/10.1007/s11229-019-02466-5

Leudar, I., & Costall, A. (2009). *Against Theory of Mind*. Palgrave Macmillan.

Leudar, I., Costall, A., & Francis, D. (2004). Theory of mind: A critical assessment. *Theory & Psychology*, *14*(5), 571–578.

Levy, R. & Mattsson, M. (2023). *The effects of social movements: Evidence from #MeToo.* SSRN. https://doi.org/10.2139/ssrn.3496903

Lewens, T. (2014). The Evolved Apprentice: How Evolution Made Humans Unique. *The British Journal for the Philosophy of Science*, *65*(1), 185–189. https://doi.org/10.1093/bjps/axt003

Lewens, T. (2015). *Cultural Evolution: Conceptual Challenges*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199674183.001.0001

Lewis, D. (1969). *Convention: A Philosophical Study.* Harvard University Press.

Lippmann, W. (1922). *Public Opinion.* Brace.

Lordkipanidze, D. (2017). The History of Early Homo. In M. Tibayrenc & F. J. Ayala (Eds.), *On Human Nature* (pp. 45–54). Academic Press.

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, *7*(1), 74–101. https://doi.org/10.1038/s41562-022-01460-1

Maiese, M., & Hanna, R. (2019). *The Mind-Body Politic.* Springer.

Mameli, M. (2001). Mindreading, mindshaping, and evolution. *Biology and Philosophy*, *16*(5), 595–626.

Mann, M. (1986). *The Sources of Social Power: Volume 1: A History of Power from the Beginning to AD 1760*. Cambridge University Press.

Marlowe, F. W. (2005). Hunter-gatherers and human evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *14*(2), 54–67.

Marotzke, J., Milinski, S., & Jones, C. D. (2022). How close are we to 1.5 degC or 2 degC of global warming? *Weather*, *77*(4), 147–148. https://doi.org/10.1002/wea.4174

Masson-Delmotte, V., Zhai, P., Pörtner, H. O., Roberts, D., Skea, J., Shukla, P. R., … & Waterfield, T. (2018). Global warming of 1.5 C. *An IPCC Special Report on the impacts of global warming of 1*.5C.

McCoy, J., Rahman, T., & Somer, M. (2018). Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist*, *62*(1), 16–42. https://doi.org/10.1177/0002764218759576

McGeer, V. (2001). Psycho-practice, psycho-theory and the contrastive case of autism. How practices of mind become second-nature. *Journal of Consciousness Studies*, *8*(5-6), 109–132.

McGeer, V. (2007). The Regulative Dimension of Folk Psychology. In D. D. Hutto & M. Ratcliffe (Eds.), *Folk Psychology Re-assessed* (pp. 137–156). Springer.

McGeer, V. (2015). Mind-making practices: The social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, *18*(2), 259–281.

McGeer, V. (2020). Enculturating folk psychologists. *Synthese*, *199*(2), 1039–1063. https://doi.org/10.1007/s11229-020-02760-7

McMahon, A. (2020). Early urbanism in northern Mesopotamia. *Journal of Archaeological Research*, *28*(3), 289–337. https://doi.org/10.1007/s10814-019-09136-7

Mehta, J., Starmer, C., & Sugden, R. (1994). The nature of salience: An experimental investigation of pure coordination games. *American Economic Review*, *84*, 658–673.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, *31*, 838–850.

Mercier, H. (2020). *Not Born Yesterday*. Princeton University Press.

Mercier, H., & Morin, O. (2019). Blind imitation or a matter of taste? *International Cognition and Culture Institute*. http://cognitionandculture.net/blogs/hugo-mercier/a-matter-of-taste/

Mesoudi, A. (2021). Cultural selection and biased transformation: Two dynamics of cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1828). https://doi.org/10.1098/rstb.2020.0053

Mesoudi, A., Chang, L., Murray, K., & Lu, H. J. (2015). Higher frequency of social learning in China than in the West shows cultural variation in the dynamics of

cultural evolution. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1798).

Meta Platforms. (2022). Number of monthly active Facebook users worldwide as of 1st quarter 2022 (in millions). *Statista*. Retrieved June 21, 2022, from https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Milanovic, B. (2016). *Global Inequality*. Harvard University Press.

Miller, D. J., Duka, T., Stimpson, C. D., Schapiro, S. J., Baze, W. B., McArthur, M. J., … & Sherwood, C. C. (2012). Prolonged myelination in human neocortical evolution. *Proceedings of the National Academy of Sciences*, *109*(41), 16480–16485.

Mirowski, P. (2013). *Never Let a Serious Crisis go to Waste: How Neoliberalism Survived the Financial Meltdown*. Verso Books.

Mirowski, P. (2014). *The political movement that dared not speak its own name: The neoliberal thought collective under erasure* (Working Papers Series No. 23). Institute for New Economic Thinking.

Mirowski, P. (2019). Hell is truth seen too late. *Boundary 2*, *46*(1), 1–53. https://doi.org/10.1215/01903659-7271327

Mirowski, P. & Plehwe, D. (Eds.). (2009). *The Road from Mont Pèlerin: The Making of the Neoliberal Thought Collective*. Harvard University Press.

Mirowski, P., & Nik-Khah, E. (2017). *The Knowledge We Have Lost in Information*. Oxford University Press.

Mogan, R., Fischer, R., & Bulbulia, J. A. (2017). To be in synchrony or not? A meta-analysis of synchrony's effects on behavior, perception, cognition and affect. *Journal of Experimental Social Psychology*, *72*, 13–20. https://doi.org/10.1016/j.jesp.2017.03.009

Monaco, D. (2022). The rise of anti-establishment and far-right forces in Italy: Neoliberalisation in a new guise? *Competition & Change 27*(1), 224–243. https://doi.org/10.1177/10245294211060123

Monbiot, G. (2016). Neoliberalism – The Ideology at the Root of All Our Problems. *The Guardian.* Retrieved 2 May, 2022, from https://www.theguardian.com/books/2016/apr/15/neoliberalism-ideology-problem-george-monbiot

Moore, R. (2016). Gricean communication and cognitive development. *The Philosophical Quarterly*, *67*(267), 303–326.

Morgan, T. J. H., Uomini, N. T., Rendell, L. E., Chouinard-Thuly, L., Street, S. E., Lewis, H. M., … & Laland, K. N. (2015). Experimental evidence for the co-evolution of hominin tool-making teaching and language. *Nature Communications*, *6*(1), 1–8. https://doi.org/10.1038/ncomms7029

Morin, O. (2016a). *How Traditions Live and Die.* Oxford University Press.

Morin, O. (2016b). Reasons to be fussy about cultural evolution. *Biology & Philosophy*, *31*(3), 447–458.

Morin, O. (2019). Did social cognition evolve by cultural group selection? *Mind & Language*, *34*(4), 530–539. https://doi.org/10.1111/mila.12252

Morin, O., Jacquet, P. O., Vaesen, K., & Acerbi, A. (2021). Social information use and social information waste. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1828). https://doi.org/10.1098/rstb.2020.0052

Morozov, E. (2012). *The Net Delusion: The Dark Side of Internet Freedom*. Public Affairs.

Morozov, E. (2019). Digital Socialism? *New Left Review*, *116/117*, 33–67.

Morton, A. (1996). Folk psychology is not a predictive device. *Mind*, *105*(417), 119–137.

Morton, A. (2003). *The Importance of Being Understood: Folk Psychology as Ethics.* Routledge.

Muhammed, S. T., & Mathew, S. K. (2022). The disaster of misinformation: A review of research in social media. *International Journal of Data Science and Analytics*, *13*(4), 271–285. https://doi.org/10.1007/s41060-022-00311-6

Mullaney, T. S., Peters, B., Hicks, M., & Philip, K. (2021). *Your Computer is on Fire*. MIT Press.

Nakahashi, W., Wakano, J. Y., & Henrich, J. (2012). Adaptive social learning strategies in temporally and spatially varying environments. *Human Nature*, *23*(4), 386–418.

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, *17*(2), 141–161. https://doi.org/10.1017/epi.2018.32

Niaki, A. A., Cho, S., Weinberg, Z., Hoang, N. P., Razaghpanah, A., Christin, N., & Gill, P. (2020). ICLab: A global, longitudinal internet censorship measurement platform. *2020 IEEE Symposium on Security and Privacy (SP)*, 135–151. https://doi.org/10.1109/SP40000.2020.00014

Nichols, S., & Stich, S. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford University Press.

Nielsen, M., & Tomaselli, K. (2010). Overimitation in Kalahari Bushman children and the origins of human cultural cognition. *Psychological Science*, *21*(5), 729–736.

O'Callaghan, P. (2020). Reflections on the Root Causes of Outrage Discourse on Social Media. In M. C. Navin & R. Nunan (Eds.), *Democracy, Populism, and Truth* (pp. 115–126). Springer.

O'Connor, C. (2019). *The Origins of Unfairness: Social Categories and Cultural Evolution*. Oxford University Press.

OECD. (2020). Abuse of dominance in digital markets. *Global Forum on Competition.* JT03476721.

Ofek, H. (2001). *Second Nature: Economic Origins of Human Evolution*. Cambridge University Press.

Ofer, G. (1987). Soviet economic growth: 1928-1985. *Journal of Economic Literature*, *25*(4), 1767–1833.

Ostrom, E., Burger, J., Field, C. B., Norgaard, R. B., & Policansky, D. (1999). Revisiting the commons: Local lessons, global challenges. *Science*, *284*(5412), 278–282.

Otten, S. (2016). The Minimal Group Paradigm and its maximal impact in research on social categorization. *Current Opinion in Psychology*, *11*, 85–89. https://doi.org/10.1016/j.copsyc.2016.06.010

Page, S. (2007). *The Difference.* Princeton University Press.

Palumbo, R. V., Marraccini, M. E., Weyandt, L. L., Wilder-Smith, O., McGee, H. A., … & Goodwin, M. S. (2017). Interpersonal autonomic physiology: A systematic review of the literature. *Personality and Social Psychology Review*, *21*(2), 99–141.

Parsell, M. (2008). Pernicious virtual communities: Identity, polarisation and the Web 2.0. *Ethics and Information Technology*, *10*(1), 41–56.

Pasquale, F. (2016). Platform neutrality: Enhancing freedom of expression in spheres of private power. *Theoretical Inquiries in Law*, *17*(2), 487–513.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, *66*(11), 4944–4957.

Peters, U. (2019). The complementarity of mindshaping and mindreading. *Phenomenology and the Cognitive Sciences*, *18*(3), 533–549.

Petersen, M. B. (2020). The evolutionary psychology of mass mobilization: How disinformation and demagogues coordinate rather than manipulate. *Current Opinion in Psychology*, *35*, 71–75. https://doi.org/10.1016/j.copsyc.2020.02.003

Pickard, V. (2020). The Public Media Option: Confronting Policy Failure in an Age of Misinformation. In W. Bennett & S. Livingston (Eds.), *The Disinformation Age* (pp. 238–258). Cambridge University Press.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(02). https://doi.org/10.1017/S0140525X04000056

Pika, S., Sima, M. J., Blum, C. R., Herrmann, E., & Mundry, R. (2020). Ravens parallel great apes in physical and social cognitive skills. *Scientific Reports*, *10*(1), 1-19.

Piketty, T. (2014). *Capital in the Twenty-first Century*. Harvard University Press.

Pinker, S. (2002). *The Blank Slate: The Modern Denial of Human Nature*. Penguin.

Pinter, B., & Greenwald, A. G. (2011). A comparison of minimal group induction procedures. *Group Processes & Intergroup Relations*, *14*(1), 81–98. https://doi.org/10.1177/1368430210375251

Pitt, J. C. (2011) *Doing Philosophy of Technology: Essays in a Pragmatist Spirit.* Springer.

Planer, R., & Sterelny, K. (2021). *From Signal to Symbol: The Evolution of Language*. MIT Press.

Popiel, P. (2018). The tech lobby: Tracing the contours of new media elite lobbying power. *Communication, Culture and Critique*, *11*(4), 566–585. https://doi.org/10.1093/ccc/tcy027

Porter, E., & Wood, T. J. (2022). Political misinformation and factual corrections on the Facebook news feed: Experimental evidence. *The Journal of Politics*, *84*(3), 1812–1817. https://doi.org/10.1086/719271

Powell, A., Shennan, S., & Thomas, M. G. (2009). Late Pleistocene demography and the appearance of modern human behavior. *Science*, *324*(5932), 1298–1301. https://doi.org/10.1126/science.1170165

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a Theory of Mind?. *Behavioral and Brain Sciences*, *1*(4), 515–526.

Pylyshyn, Z. (1984). *Computation and Cognition*. MIT Press.

Pyo, J., & Maxfield, M. G. (2021). Cognitive effects of inattentive responding in an MTurk sample. *Social Science Quarterly*, *102*(4), 2020–2039. https://doi.org/10.1111/ssqu.12954

Rabaglia, C. D., Marcus, G. F., & Lane, S. P. (2011). What can individual differences tell us about the specialization of function? *Cognitive Neuropsychology*, *28*(3–4), 288–303. https://doi.org/10.1080/02643294.2011.609813

Rahman, K. S. (2018). Regulating informational infrastructure: Internet platforms as the new public utilities. *Georgetown Law and Technology Review*, *2*(2), 234–251.

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology*, *44*(3), 875–881.

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26). https://doi.org/10.1073/pnas.2024292118

Richerson, P. J. (2017). Recent critiques of dual inheritance theory. *Evolutionary Studies in Imaginative Culture*, *1*(1), 9. 875–881.

Robinson, G. (2017). “Down to the very roots”: The Indonesian army’s role in the mass killings of 1965–66. *Journal of Genocide Research, 19*(4), 465–486. https://doi.org/10.1080/14623528.2017.1393935

Rorty, R. (1989). *Contingency, Irony and Solidarity*. Cambridge University Press.

Rorty, R. (1991a). Moral identity and private autonomy: The case of Foucault. In R. Rorty (Ed.) *Essays on Heidegger and Others: Philosophical Papers* (pp. 193–198). Cambridge University Press.

Rorty, R. (1991b). *Objectivity, Relativism, and Truth: Philosophical Papers (Vol. 1).* Cambridge University Press.

Rorty, R. (1998). *Achieving Our Country: Leftist Thought in Twentieth Century America*. Harvard University Press.

Rorty, R. (2021). *Pragmatism as Anti-Authoritarianism*. Belknap Press.

Ross, D. (2005). *Economic Theory and Cognitive Science: Microexplanation.* MIT Press.

Ross, D. (2007). H. sapiens as ecologically special: What does language contribute?. *Language Sciences, 29*(5), 710–731.

Ross, D. (2008). Classical game theory, socialization and the rationalization of conventions. *Topoi, 27*(1-2), 57–72.

Ross, D. (2014). *Philosophy of Economics*. Palgrave Macmillan.

Ross, D. (2015). A Most Rare Achievement: Dennett’s Scientific Discovery in *Content and Consciousness*. In C. Muñoz-Suárez & F. De Brigard (Eds.), *Content and Consciousness Revisited: With Replies by Daniel Dennett* (pp. 29–48). Springer.

Ross, D. (2019). Consciousness, language, and the possibility of non-human personhood: Reflections on elephants. *Journal of Consciousness Studies*, *26*(3–4), 227–251.

Ross, D. (2022). Economics is converging with sociology but not with psychology. *Journal of Economic Methodology, 30*(2), 135–156. https://doi.org/10.1080/1350178X.2022.2049854

Ross, D., & Stirling, W. (2021). Economics, Social Neuroscience, and Mindshaping. In J. Harbecke & C. Herrmann-Pillath (Eds.), *Social Neuroeconomics* (pp. 174–201). Routledge.

Ross, D., & Stirling, W. (2023). Mindshaping, conditional games, and the Harsanyi Doctrine. *CEAR Working Paper 2023-03*.

Ross, M. H. (1988). Political organization and political participation: Exit, voice and loyalty in preindustrial societies. *Comparative Politics, 21*(1), 73–89.

Salganik, M. J. (2018). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Savage, L. (1954). *The Foundations of Statistics*. Wiley.

Schelling, T. (1960). *The Strategy of Conflict.* Harvard University Press.

Schmidt, M. F., Butler, L. P., Heinz, J., & Tomasello, M. (2016). Young children see a single action and infer a social norm: Promiscuous normativity in 3-year-olds. *Psychological Science, 27*(10), 1360–1370.

Schüll, N. D. (2014). *Addiction by Design: Machine Gambling in Las Vegas*. Princeton University Press.

Schweppe, J., & Walters, M. A. (Eds.). (2016). *The Globalization of Hate: Internationalizing Hate Crime?*. Oxford University Press.

Seymour, M. J. (2011). Mesopotamia. In T. Insoll (Ed.), *The Oxford Handbook of the Archaeology of Ritual and Religion* (pp. 775–94). Oxford University Press.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193.

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge University Press.

Skyrms, B. (2003). *The Stag Hunt and The Evolution of Social Structure*. Cambridge University Press.

Slobodian, Q. (2018). *Globalists*. Harvard University Press.

Sperber, D. (1996). *Explaining Culture: A Naturalistic Approach.* Oxford University Press.

Sperber, D. (2017, August 5). Cecilia Heyes on the social tuning of reason. *International Cognition and Culture Institute*. http://cognitionandculture.net/blogs/dan-sperber/cecilia-heyes-on-the-social-tuning-of-reason/

Starr, P. (2019). How neoliberal policy shaped the internet—and what to do about it now. *The American Prospect*, *2*.

StatCounter. (2022). Worldwide desktop market share of leading search engines from January 2010 to January 2022. *Statista*. Retrieved June 21, 2022, from https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

Sterelny, K. (2003). *Thought in a Hostile World*. Blackwell.

Sterelny, K. (2010). Minds: Extended or scaffolded? *Phenomenology and the Cognitive Sciences*, *9*(4), 465–481. https://doi.org/10.1007/s11097-010-9174-y

Sterelny, K. (2012). *The Evolved Apprentice: How Evolution Made Humans Unique*. MIT Press.

Sterelny, K. (2017a). Adaptation Without Insight. In R. Boyd (Ed.), *A Different Kind of Animal: How Culture Transformed Our Species* (pp. 135–151). Princeton University Press.

Sterelny, K. (2017b). Cultural evolution in California and Paris. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *62*, 42–50. https://doi.org/10.1016/j.shpsc.2016.12.005

Sterelny, K. (2018). Why reason? *Mind & Language*, *33*(5), 502–512. https://doi.org/10.1111/mila.12182

Sterelny, K. (2020). Afterword: Tough questions; hard problems; incremental progress. *Topics in Cognitive Science*, *12*(2), 766–783. https://doi.org/10.1111/tops.12427

Sterelny, K. (2021). *The Pleistocene Social Contract: Culture and Cooperation in Human Evolution*. Oxford University Press. https://doi.org/10.1093/oso/9780197531389.003.0002

Storey, G. R. (Ed.). (2009). *Urbanism in the Preindustrial World: Cross-Cultural Approaches*. The University of Alabama Press.

Stout, D., & Chaminade, T. (2012). Stone tools, language and the brain in human evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1585), 75–87. https://doi.org/10.1098/rstb.2011.0099

Sugden, R., & Zamarrón, I. E. (2006). Finding the key: The riddle of focal points. *Journal of Economic Psychology*, *27*(5), 609–621.

Sullivan, E., Sondag, M., Rutter, I., Meulemans, W., Cunningham, S., … & Alfano, M. (2020). Can real social epistemic networks deliver the wisdom of crowds? In E. Sullivan, M. Sondag, I. Rutter, W. Meulemans, S. Cunningham, B. Speckmann, & M. Alfano (Eds.), *Oxford Studies in Experimental Philosophy Volume 3* (pp. 29–63). Oxford University Press. https://doi.org/10.1093/oso/9780198852407.003.0003

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202

Taylor, L. (2021). Public actors without public values: Legitimacy, domination and the regulation of the technology sector. *Philosophy & Technology*, *34*(4), 897–922.

TechCrunch. (2021). Share of social media users who regularly get news from selected social media sites in the United States in 2020 and 2021. *Statista*. Retrieved June 21, 2022, from https://www.statista.com/statistics/330638/politics-governement-news-social-media-news-usa/

Thompson, A. (Ed.). (2007). *The Media and the Rwanda Genocide*. Pluto Press.

Tomlin, T. J. (2014). *A Divinity for All Persuasions: Almanacs and Early American Religious Life*. Oxford University Press.

Tooby, J., & Cosmides, L. (1995). Mapping the Evolved Functional Organization of Mind and Brain. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (pp. 1185–1197). MIT Press.

Trouche, E., Johansson, P., Hall, L., & Mercier, H. (2018). Vigilant conservatism in evaluating communicated information. *PLOS ONE*, *13*(1). https://doi.org/10.1371/journal.pone.0188825

Tunçgenç, B., & Cohen, E. (2016). Movement synchrony forges social bonds across group divides. *Frontiers in Psychology*, *7*(782).

Turchin, P., Whitehouse, H., Korotayev, A., Francois, P., Hoyer, D., … Currie, T. E. (2018). Evolutionary pathways to statehood: old theories and new data. *SocArXiv*.

Udehn, L. (2001). *Methodological Individualism: Background, History and Meaning*. Routledge.

United States. (2021). United States Code. Title 47, section 230, Office of the Law Revision Counsel, Congress, House.

Vadde, A. (2021). Platform or publisher. *PMLA*, *136*(3), 455–462.

Vallier, K. (2021). Neoliberalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/neoliberalism/

Van Horn, R. (2009). Reinventing Monopoly and the Role of Corporations. In P. Mirowski & D. Plehwe (Eds.), *The Road from Mont Pèlerin: The Making of the Neoliberal Thought Collective* (pp. 204–37). Harvard University Press.

Venugopal, R. (2015). Neoliberalism as concept. *Economy and Society*, *44*(2), 165–187. https://doi.org/10.1080/03085147.2015.1013356

Verbeek, P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things*. University of Chicago Press.

Vlerick, M. (2020). The cultural evolution of institutional religions. *Religion, Brain & Behavior*, *10*(1), 18–34. https://doi.org/10.1080/2153599X.2018.1515105

Weatherson, B. (2021). David Lewis. In E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Winter 2021). https://plato.stanford.edu/entries/david-lewis/

Whiten, A. (2017). Social learning and culture in child and chimpanzee. *Annual Review of Psychology*, *68*(1), 129–154. https://doi.org/10.1146/annurev-psych-010416-044108

Wijenayake, S., van Berkel, N., Kostakos, V., & Goncalves, J. (2020). Impact of contextual and personal determinants on online social conformity. *Computers in Human Behavior*, *108*. https://doi.org/10.1016/j.chb.2020.106302

Wikström, V., Saarikivi, K., Falcon, M., Makkonen, T., Martikainen, S., … & Tervaniemi, M. (2022). Inter-brain synchronization occurs without physical co-

presence during cooperative online gaming. *Neuropsychologia*, *174*, 108316. https://doi.org/10.1016/j.neuropsychologia.2022.108316

Williams, D. (2022a). Signalling, commitment, and strategic absurdities. *Mind & Language*, *37*(5), 1011–1029. https://doi.org/10.1111/mila.12392

Williams, D. (2022b). The marketplace of rationalizations. *Economics & Philosophy*, *39*(1), 99–123. https://doi.org/10.1017/S0266267121000389

Wilson, C., & Jumbert, M. G. (2018). The new informatics of pandemic response: Humanitarian technology, efficiency, and the subtle retreat of national agency. *Journal of International Humanitarian Action*, *3*(1), 1–13.

Wiltermuth, S. S., & Heath, C. (2009). Synchrony and cooperation. *Psychological Science*, *20*(1), 1–5. https://doi.org/10.1111/j.1467-9280.2008.02253.x

Wise, J. (2022). Life expectancy: Parts of England and Wales see "shocking" fall. *BMJ*, *377*. https://doi.org/10.1136/bmj.o1056

Wolfendale, P. (2019). The reformatting of *Homo sapiens*. *Angelaki*, *24*(1), 55–66.

Wong, K., Garza, L., & Robbins, M. (2021). *"Everything is Theory Versus Practice": Department Policy Changes in the Time of Black Lives Matter*. Poster presented at the International Research Conference for Graduate Students, San Marcos, Texas.

Wraight, T. (2019). From Reagan to Trump: The origins of US neoliberal protectionism. *The Political Quarterly*, *90*, 735–742. https://doi.org/10.1111/1467-923X.12709

Wu, T. (2017). *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*. Vintage.

Youngblood, M., Stubbersfield, J. M., Morin, O., Glassman, R., & Acerbi, A. (2021). *Negativity bias in the spread of voter fraud conspiracy theory tweets during the 2020 US election* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/2jksg

Zakrzewski, C. (2022). Tech companies spent almost $70 million lobbying Washington in 2021 as Congress sought to rein in their power. *The Washington Post*. Retrieved 21 December, 2022, from https://www.washingtonpost.com/technology/2022/01/21/tech-lobbying-in-washington/

Zamorano-Abramson, J., Michon, M., Hernández-Lloreda, M. V., & Aboitiz, F. (2023). Multimodal imitative learning and synchrony in cetaceans: A model for speech and singing evolution. *Frontiers in Psychology*, *14*, 1061381. https://doi.org/10.3389/fpsyg.2023.1061381

Zawidzki, T. W. (2008). The function of folk psychology: Mind reading or mind shaping?. *Philosophical Explorations*, *11*(3), 193–210.

Zawidzki, T. W. (2012). Unlikely allies: Embodied social cognition and the intentional stance. *Phenomenology and the Cognitive Sciences*, *11*(4), 487–506. https://doi.org/10.1007/s11097-011-9218-y

Zawidzki, T. W. (2013). *Mindshaping: A New Framework for Understanding Human Social Cognition.* MIT Press.

Zawidzki, T. W. (2015, July 1). Communication without metapsychology. *International Cognition and Culture Institute*. http://cognitionandculture.net/webinars/speaking-our-minds-book-club/communication-without-metapsychology/

Zawidzki, T. W. (2018). The Many Roles of the Intentional Stance. In B. Huebner (Ed.), *The Philosophy of Daniel Dennett* (pp. 36–61). Oxford University Press. https://doi.org/10.1093/oso/9780199367511.003.0003

Zawidzki, T. W. (2019). A new perspective on the relationship between metacognition and social cognition: Metacognitive concepts as socio-cognitive tools. *Synthese, (198)*7, 6573–6596.

Zentall, T. R. (2004). Action imitation in birds. *Animal Learning & Behavior*, *32*(1), 15–23. https://doi.org/10.3758/BF03196003

Zollman, K. J. (2007). The communication structure of epistemic communities. *Philosophy of Science*, *74*(5), 574–587.

Zollman, K. J. (2013). Network epistemology: Communication in epistemic communities. *Philosophy Compass*, *8*(1), 15–27. https://doi.org/10.1111/j.1747-9991.2012.00534.x

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books.