The Ethics of Cognitive Security

Andrew Ward Buzzell

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PHILOSOPHY
YORK UNIVERSITY
TORONTO, CANADA

November, 2023

Abstract

This dissertation concerns ethical and epistemic assessment of the use of state power to defend against information threats and hostile activities, especially in digital information environments, an activity which has been described as the pursuit of cognitive security. I have three main aims. Firstly, to motivate scholarly interest in what I call the ethics of cognitive security - an interdisciplinary effort to provide coordinated empirical, theoretical, and ethical input into this exercise of power, specifically by democratic states. To the extent that these are decide to develop offensive and defensive strategies for the conduct of information warfare, we can ask if there are empirical and ethical considerations that should inform this. I consider at length the literature on epistemic paternalism and the extent to which such efforts might be paternalist and thus objectionable or in need of special justification. Secondly, to articulate an ecological conception of our epistemic interdependence on the information environment that can help to describe the public interest in its responsible stewardship. This helps to generate less-securitized formulations of the aims and constraints of defensive operations that interface better with our ethical interests in the practice of cognitive security. I draw on the social epistemology literature, especially where it concerns testimony and epistemic dependence, and develop a conception of epistemic environmental dependence and trust. I connect this to environmental policy and philosophy scholarship that concerns the ethical relations that arise as a result of interdependencies, and the sorts of policy instruments that can be appropriate for protecting this kind of shared interest. Finally, I apply this conceptual framework to the specific cognitive security problem of hostile disinformation. Drawing on a conceptual analysis of epistemic pollution I argue that we should favour root cause analysis and remediation, even where it might seem to involve more substantial interference, rather than reactive efforts to filter and remove these, for reasons that include both empirical adequacy and respect for relevant ethical interests.

Acknowledgements

It has been an enormous privilege and pleasure to work with Regina Rini as my supervisor. Her guidance, advice, critical acumen and encouragement have helped me immeasurably. I've learned so much from our conversations, reading drafts and comments and assisting with research. I'm especially grateful for the opportunity to collaborate on several projects, which have been valuable to me in many ways. I wish I could say that I had absorbed more of her wonderful writing style.

I would like to thank Neil Levy for serving as my external examiner and providing extremely helpful comments, questions, and suggestions, and travelling to Toronto to participate in person in my defence. I'd like to thank Julianne Chung and Henry Jackman for agreeing to be on my committee and providing valuable comments and interesting conversations about this work as it has progressed. Thanks to Robert Gehl for agreeing to be an examiner and offering challenging and useful feedback and commentary, and to Parisa Moosavi for helping with the defence.

Alice MacLachlan has provided invaluable support throughout my time at York for which I'm very grateful, including helping me with a successful SSHRC application in my first year and helping to organize a mock interview in my last. After having spent almost ten years away from academia, I had the good fortune to have Claudine Verheggen and Robert Myers as instructors during the intensive first-year seminar which was a challenging but fantastic experience that I remember fondly. Thanks to Jacob Beck who offered me the opportunity to work with him as in a directed reading assignment and as a research assistant early in my PhD which helped me to improve my research and analytical skills and to catch up on the philosophical work on representation and perception.

I'd like to thank Jay Shaw at the University of Toronto Joint Centre for Bioethics for offering me a place in the Scholars Program in AI Ethics and Health, the opportunity to collaborate on research, and for advice, guidance and support over the last few years.

Finally, I owe a debt to everyone at York and in my department with whom I've had the pleasure of working with, especially my fellow cohort members Jef Delvaux and Rand Hirmiz.

Funding: This research was supported with external funding provided by a Doctoral Fellowship from the Social Sciences and Humanities Research Council of Canada (SSHRC), as well as the University of Toronto Joint Centre for Bioethics Scholars Program in AI Ethics and Health. It has also been supported by internal awards provided by Faculty of Graduate Studies at York University, and the Susan Mann Dissertation Scholarship.

Table of Contents

Introduction

Countries around the world are enacting policies and developing strategies that, whether they use the term or not, pursue cognitive security. This is the extension of national security interest to the information environment, especially to digital media systems. It follows the articulation of cybersecurity interest in previous decades, which likewise expanded the national security frame to new, previously private, areas of concern. In both cases, new technologies which digitized and networked our public and private activities expanded the scope and scale of access, analysis, and action, both friendly and hostile.

The immediate aim of cognitive security policy is to defend against information warfare, especially where it targets political and institutional capacities that are essential for governance. Information operations have already affected elections, public health responses to crises, and sparked and amplified serious conflicts around the world. Tactics such as targeted disinformation are supplanted with techniques designed to generate mistrust, to demoralize, to polarize, and to frustrate and impede consensus building and information-seeking. Even where outright manipulation is unfeasible, effective information warfare can paralyze, or at least frustrate, the means by which states react to information and events, a stated goal of several nation's offensive cognitive warfare doctrine.

As I write this, in mid-October 2023, reports of an Israeli attack on a hospital in Gaza reverberated around the world, generating anger, sparking protests and violence, and leading to the cancellation of a summit meeting between several involved states. In mere hours the story spread across social media platforms to mainstream media around the world. At this time it is still unclear what happened, but there is strong evidence that disinformation played a critical role in spreading and distorting the dominant narratives, at least one of which is plainly false. Even days later as evidence began to accumulate, in some respects it didn't matter, because many people disbelieved the substance and sources of that evidence. From a cognitive security perspective, what is particularly important are the preconditions that make this kind of event possible.

The tempo at which it unfolded outpaced reactive strategies such as filtering, fact-checking, labelling, de-ranking, or otherwise modulating misinforming content in digital media systems. It exploited properties of the information environment that are both intentional and accidental and was scaffolded on the already threadbare and contested terrain of the ordinary mechanisms that generate shared consensus on matters of public interest. Broader-spectrum information operations prepare the ground for this sort of inflection point by encouraging distrust, fragmenting audiences, and developing influence capabilities, which often is difficult to distinguish from legitimate skepticism and ordinary disagreement. For democratic countries attempting to formulate countermeasures, a significant problem is protecting one while inhibiting the other.

We might agree in principle that it's dangerous for our information systems to be so easily used to subvert and undermine our considered articulations of public and national interest. But it's difficult to develop a defensive strategy that can prevent significant adverse events, and respond effectively, that can also sit comfortably alongside democratic values, such as we and our institutions understand and endorse these. For some, any form of domestic cognitive security intervention is prima facie illiberal and illegitimate. From an epistemological point of view, thinking about media systems as the infrastructure of knowledge production and social coordination, many find the idea of imposition of top-down state or corporate management of content within these as fundamentally misaligned with what we know about how these processes function best. Some philosophers have argued that many kinds of management efforts are objectionably paternalist.

We do not expect that our information systems are relentlessly factual. These systems are increasingly built and managed as centralized monoliths, but this fits poorly with the ways we use them. We don't imagine they yield only expert opinions and acceptable views. We don't suppose that some qualification be required to engage with them, or that they should be designed with only serious matters in mind. One animating value of liberal democratic thought is that we don't know in advance what information is valuable, what views are correct, which framings are most useful, what best resonates with the public interest and mood. Instead, it embraces uncertainty and epistemological anarchy,

supposing that free inquiry will sort things out most efficiently. This is where cognitive warfare aims to exploit openness and uncertainty to tip the scales for some political gain.

My project aims firstly to motivate scholarly interest in what I call the ethics of cognitive security - an interdisciplinary effort to provide coordinated empirical, theoretical, and ethical input into the exercise of institutional power to defend against hostile information operations. To the extent that states are forced to develop offensive and defensive strategies for the conduct of information warfare, we can ask if there are empirical and ethical constraints that demand consideration.

Secondly, I will argue that we should articulate the public interest we have in the information environment to inform an account of information ecosystem health. This offers a less reactive and less-securitized theoretical approach to the problem, by formulating a conception of cognitive security based on our shared epistemic dependencies. Just as with the ecological environment, we have complex epistemic interdependencies, and this offers an alternate perspective on cognitive security - instead of deciding post-hoc what is a pollutant, what is a dangerous process, we might identify the conditions of ecosystem health against which such judgments are made. These might be better goals for democratic cognitive security policy to pursue. We might hope that even when adverse events occur, we can trust the resilience of a healthy information environment. Sunstein, writing about this problem over 30 years ago, notes that " we do not know what a well- functioning marketplace of ideas would look like"(Sunstein,

1992), and in a more recent response to Sunstein's observation, Hwang (Hwang, 2020) advocates a more modest goal, just to identify what "unfair competition in the marketplace of ideas" might look like. I don't think we can do the latter without the former, and in either case, we should expect our findings to be tentative and uncertain and to require continuous public and political discussion. It should not be surprising if some of the findings are quite radical, that we have centralization in the wrong places, or that some public systems that should be private, or some private ones should be public. Many of the sociotechnical systems that intermediate so much of our social and political activity aren't themselves reflective of the purposes for which we rely on them and the dependencies we engage with them.

Here's how my argument develops in the coming chapters. In the first, I provide a historical sketch of information warfare and the current state of play and articulate a conception of cognitive security. In the second, I consider, and largely dismiss, concerns that most of what we might do in the pursuit of defensive cognitive security goals will be objectionably paternalist. The third chapter draws on social epistemology, especially the epistemology of testimony, to paint a picture of our collective dependence on the adequacy of information environments to furnish the contextual information we need to acquire knowledge within them. I describe this as an epistemic environmental dependence, and one that makes us vulnerable to misplaced trust. In the fourth chapter I connect this to the role that interdependence plays in environmental ethics and policy, and I argue that ecology and environmental philosophy offers a rich and productive

metaphor for thinking about information environment health and our public interest in maintaining it. The final chapter critiques and re-imagines cognitive security from this ecological perspective. I apply the ecological view to the problem of disinformation, with the first-order goal of conducting a productive analysis, but also a second-order goal, to demonstrate the type of engagement that I envision the ethics of cognitive security as a research project would undertake.

Chapter 1 - Cognitive Security

1.1 Cognitive security and the epistemic environment

This dissertation has three overarching goals. The first is programmatic - to argue for the establishment of the ethics of cognitive security as field of inquiry. The second is methodological - that philosophical work on social epistemology is a rich and productive resource for such a project to draw on. The third is conceptual - that the pursuit of cognitive security demands an epistemically, empirically, and ethically grounded conception of a healthy information environment that can support the normative claims entailed by much of the current cognitive security discourse. For example, can we make use of concepts like fake news and disinformation in ways that diagnose real problems in the infosphere, and not just to express dissatisfaction with content we find there? Are cognitive security concerns necessarily partisan and securitized, or can they be articulated in ways that express genuine stewardship of a public good? This first chapter will focus on context and concepts central to the first goal and lay some groundwork for the other two.

Cognitive security is an emerging concern in national security discourses, with historical and conceptual connections to discussions of information warfare, psychological operations, and propaganda. It describes the vulnerability of political entities to interference in the epistemic substrate of political and social systems they depend on, and which risks a range of harms, including existential ones. Characteristic

threats include disinformation, propaganda, distrust, election interference, and various kinds of sabotage to information and media systems. I argue that the concept of cognitive security, the empirical picture we can paint today of the range of action in this domain, and that which we can forecast in the near future, warrants coordinated interdisciplinary study. When I speak of the ethics of cognitive security, I mean to ask not just how its aims can be pursued ethically, as though those aims are given, and can be found independently of ethical reflection, but more broadly, how we should determine what those aims should be.

We are currently in the reactive early phases of the transition to a pervasive, borderless, always-on conception of the challenges of cognitive security, responding ad hoc to whatever emergencies reach our attention, guided by whatever institutions happen to be involved or mobilizable. There is an urgent need to gain a unified perspective on this field of endeavour so that we can respond deliberately, within frameworks of democratic governance, and in ways that can be genuinely responsive to the public good. Or, if it should turn out that there really is no such conception of cognitive security, then theory and practice should reflect the practical aims that animate policy and should eschew the insinuation that a higher purpose is pursued. Interventions in the service of cognitive security are all around us, but we lack a coordinated civilian effort to understand the conditions of justification and the risks for harm. My hope is that this project can contribute to the development of an organized interdisciplinary effort to produce and advocate sound policy in this arena.

The domain of cognitive security, and the actions taken in the service of it, are distinctively epistemic, in that it targets epistemic states, what a community knows, what it does not know, by way of epistemic modalities - how some community comes to regard information as knowledge. There is an implied rupture between the normal and the hostile within the messy practices we use to collectively make sense of our experiences and acquire knowledge - that malicious actors deliberately intervene in ways to put a finger on the scale, in ways that will undermine collective interests. These do not only involve what we might think of as straightforwardly false claims. For example, Boichak (Boichak, 2023) observes that Russian social media influence operations are increasingly indirect, generating a broad spectrums of content that helps to crystalize narratives that construct a context in which some particular narrative, such as Russia's role in the genesis of WWII appears credible. Chomsky describes the deliberate "bounding of the thinkable" (Chomsky, 2003) as both a goal and a mechanism of modern propaganda.

I will argue at length in Chapter 5 that mere falsehood does not capture the diagnostics of undesirable content entailed by most conceptions of cognitive security. Therefore, I consider an approach to this problem by asking more broadly what kinds of normative and epistemic claims are implicit in conceptions of cognitive security, and what kinds of epistemic engineering might be effectively and justifiably undertaken to promote them. I argue that social epistemology and allied social sciences provide a rich resource to help understand some of the pragmatic and ethical problems that arise. I will argue that

philosophical work on epistemic dependence is particularly relevant to understanding the challenges this raises, but also, that what we know about epistemic values and practices creates valuable countervailing pressure on approaches to cognitive security that expand the securitization frame too far, or that narrowly focus on veritistic properties of the information environment. It is because of our essential epistemic interdependence, and the specific ways our epistemic infrastructure reflects and structures this, that cognitive security vulnerabilities arise, and where the epistemic, moral, and practical risks of actions undertaken in the service of cognitive security emerge.

Much of this chapter will veer away from traditional philosophical topics, as it provides a historical and technical sketch of information operations and capabilities. Before I begin with that, I want to orient the reader to the underlying philosophical interest I bring to this.

Writing about Confederate soldiers killed in Gettysburg during the American Civil War, Lear (Lear, 2022) responds to criticism that he's unduly sympathetic. He argues this his sympathy is just for those who tried to live good and honorable lives, but, "... for historical and cultural reasons, along with character flaws of their own, get caught in a vision that is wildly wrong and profoundly unjust due to misunderstandings and misperceptions and social pressures - and then waste their lives, sometimes doing terrible harm, in a cloud of misapprehension and falsity". (Lear 2022 p. 91). We might extend this same concern for those who are manipulated by information operations, those caught

in webs of propaganda, and we might consider that there are cases where they themselves are epistemically blameless - that they have followed reasonable and virtuous procedures, have arrived at beliefs that are locally true and justified by evidence, but by some external measure strike outsiders as quite mistaken, even terribly so. Several defendants charged with involvement in the January 6 2021 attack on the US capital made just this claim in court, that they behaved reasonably on the basis of information from credible sources (including the President!) that the election had been stolen (Kosseff, 2023). Whatever we might say about the proper apportioning of individual responsibility, I think we find it natural in these sorts of cases to direct some of the blame at the epistemic environment itself, and I'm interested here in what we can say more precisely about this

Part of what is stake when we ask this question is the extent to which what justifies our beliefs lies within us, and epistemologists have disputed the extent to these are entirely internal to our cognition, or whether some of it lies beyond, out of our control. Much of this turns on technical issues, but it's common to think that one problem with the external account is that it fails to deliver substantive norms of epistemic conduct, that blame and accountability fade almost entirely from the picture. The worry is that if what we believe is substantially impacted by conditions outside us, and we conduct ourselves in epistemically responsible ways with the evidence we have, then we can have false-yet-justified beliefs, and true-yet-unjustified beliefs. However, in "Radical Externalism", Srinivasan (Srinivasan, 2020) argues that externalism about justification can in fact deliver this normative content, by diagnosing cases where structural conditions external

to the knower prevent the formation of true beliefs when we might otherwise think they ought to. Here the account of how we are justified detaches from that of how we should behave as epistemic agents. As Levy (Levy, 2022) playfully describes this kind of case, bad beliefs can happen to good people, and a causal explanation of what has gone wrong must include conditions of the external epistemic environment, which Srinivasan calls 'bad ideology', and Levy describes as 'epistemic pollution'. The normative problem for externalism is supposed to be that it can't tell us how we should conduct ourselves as epistemic agents, but it might yet make normative claims about how conditions external to the agent should be.

For instance, we should not carelessly form beliefs about the safety of childhood vaccinations that will influence choices we might make on behalf of children in our care. But our ability to make good on this expectation has substantial environmental dependencies, and somewhere along the line we will run into questions about whether this is an "ought" that often lacks an accompanying "can". In my view, this points to a new kind of normative content that an externalist epistemology might generate, that we can evaluate the adequacy of one's epistemic environment to the demands we make of it. These questions multiply in an age of networked propaganda and ubiquitous information operations, and lead us directly to a substantive philosophical question, whether we can say something about this putative environmental dysfunction over and above criticism we might make of the specific content that is encountered within it.

Hinting at an Aristotelian conception of epistemic agency, Srinivasan argues that "[v]irtue requires being embedded in a cooperative world.", and that "... the epistemic goods really worth having are those that cannot be had by mere individual effort."(Srinivasan, 2020, p. 37). In my account of this, we have environmental epistemic dependencies - our capacity to have mostly true and useful beliefs, and ones that by and large won't lead us to harm ourselves or others, depends on properties of the epistemic environment we inhabit. As a result, we have duties and responsibilities to ensure that these external conditions are as conducive to the pursuit of those epistemic goods as they can be.

This suggests a guiding principle for cognitive security - the health of the information environment. This in turn suggest a central question for the ethics of cognitive security - what is ecological health in the information environment, and in what ways can we effectively and ethically foster this, so that we are best positioned to pursue those epistemic goods worth having? This is a question about acceptable forms of epistemic engineering. It's discomforting to entertain a view that there is information that should be withheld from us because it might be bad for us, or for our epistemic community, or that our encounter with this information should be mediated to foster specific doxastic outcomes. Likewise, it's distasteful to consider subjecting others to persuasive efforts that are not just liminal and discrete, a pamphlet in the marketplace of ideas, but that are baked into the information environment in ways we can't discern. Yet not only are these types of interventions taking place, they are also increasingly common, and undertaken

by a range of individual, institutional, commercial, and state actors. We are in the early stages of a huge and significant change to our epistemic environment, where sophisticated technologies afford powerful methods of epistemic engineering to a wide variety of actors, some hostile, some commercially motivated, but also some that pursue defensive and altruistic ends.

1.2 Information operations

In 1942 the United States Government established the Office of War Information (OWI), centralizing under federal authority efforts to inform the public about the war America had just entered. Before the OWI, a variety of institutions were tasked with providing facts and figures, monitoring public sentiment, disseminating government communications, and conducting public diplomacy. Concerned that such efforts, especially the latter, amounted to objectionable domestic propaganda, Roosevelt at first resisted the creation of the OWI, and hoped that once the "facts of war"(Weinberg, 1968, p. 74) were made plain, public support would follow. As this optimism proved unfounded and public morale flagged, he supported a dedicated information operations effort, under the banner of civil defense, similar to the Ministry of Information (MOI) already established in the United Kingdom in 1935 (itself a vestige of similar efforts in the latter years of WW1).

Offensive information warfare is as old as war itself, and propaganda has long been used to bolster public support for state actions. But there was something new about this approach, taking the epistemic functioning of the public as an information system as a matter for proactive state interest. In the UK the MOI deputized the public in a variety of efforts that required understanding and alignment with strategic goals, and in the US the OWI produced media designed to avoid the loss of public support for a war which many did not believe benefitted America directly, firstly by aiming to educate the public with truthful accounts of the conflict and the reasons for engaging in it, later, and more controversially, censoring and manipulating media to alter public perception of the conflict. This later activity included direct interference with domestic journalism, such as the reporting of the Soviet massacre of Polish officers in Katyn (Szymczak, 2010), where the OWI pressured US media to suppress evidence that America's erstwhile ally had committed this crime, to avoid domestic political pressure to dissolve the alliance.

I take the kinds of campaigns undertaken by the MOE and the OWI to be clear examples of national security initiatives in the domestic information environment. The very idea of projecting state power into the domestic information environment in ways that might target specific content or ideas, and actively intervene to shape the prevailing epistemic climate, is uncomfortable for western democracies and not easily accommodated by the principles that underlay them. Even the public discussion of such activities by hostile states can become uncomfortable and reveal contradictions. The Institute for Propaganda Analysis (IPA) was formed in the United States in 1937, whose

founders included influential public relations pioneers, and which engaged in a highly visible campaign to teach propaganda analysis to the public and introduce curricula to the public school system. In the run-up to WWII, the IPA came under pressure to soften its views. Critics complained that it "... was analyzing any propaganda, regardless of its intent, and that this approach was counterproductive, hindering the US war effort against Nazi Germany" (Fondren, 2021, p. 280) When America joined the war, recognizing the likelihood that institutions such as the MOI and OWI would be needed, the IPA dissolved, worrying that teaching propaganda analyses "...could be misused for undesirable purposes by persons opposing the government's effort" (Sproule, 2005, p. 176)

The OWI was shut down after the Second World War, but soon after the U.S. Information and Educational Exchange Act of 1948 was enacted, which authorized the federal government to undertake information operations, but specifically prohibited their use domestically, in large part due to concerns that they might become an organ of state propaganda. The United States Information Agency (USIA) was established in 1953 with the mission of engaging in public diplomacy - overt efforts ".. to understand, inform and influence foreign publics in promotion of the national interest" (Chodkowski, 2012, p. 2) and was restricted from targeting domestic audiences. This limitation was also placed on the U.S. Agency for Global Media (USAGM) a Cold War instrument of information operations policy which was prohibited from operating in democracies with free presses,

and instead was to target foreign nations which "... lack adequate sources of free information". (22 U.S.C. § 1431).

An increasingly transnational infosphere created challenges for this approach even in the early days of widespread internet access. A partially declassified US Defense Department report admitted the "...likelihood that PSYOP messages will be replayed to a much broader audience, including the American public" (Wall, 2010, p. 292). A former commander of the Combined Joint Information Campaign Task Force commented that the State department adopted a policy not "...to devote resources by policy to ‘psyoping’ the American citizens" (ibid p. 291), while tacitly acknowledging that they would nonetheless be exposed due the porousness of digital information systems.

Roosevelt and the IPA both recognized that protecting domestic epistemic infrastructure and information environments is critical to the security of a nation at war, and both found themselves adopting uncomfortable approaches, with Roosevelt reluctantly acquiescing to domestic IO, and the IPA conceding that at times the state requires epistemic and psychological leverage over the public. Attempts to limit direct interference in domestic affairs, and with those of democracies with free presses, can be seen as rudimentary way to create principled constraints that reflects this hesitation to articulate a full-throated defense of domestic information operations, but ones that we will see are increasingly difficult to operationalize and justify in a fragmented and digitized world where the structure of the information environment threatens distinctions

between domestic and foreign actors, media systems, and threats. Even distinctions between human actors and automated bots are muddied by systems that enable coordination and semi-automation of human activity (Chu et al., 2012; Schreckinger, 2016)

Information operations have played an increasingly central role in national security since the start of the Cold War, from pre-digital efforts such as the Soviet Union's planting of false accounts of US involvement in the AIDS epidemic in 1983 (Boghardt, 2009), which eventually circulated widely and has generated misperceptions that persist to this day. Cold war disinformation campaigns conducted by the Soviets included forged evidence that the KKK was going to target white athletes at the LA Olympics, that the Korean Airlines passenger plane it shot down in 1983 was in fact a spy plane, and that United States was importing Latin American children for organ harvesting (Brantly, 2020; Palca, 1988) which received wide airplay at the time and is a recurring theme in disinformation campaigns even today.

The United States was initially dismissive of Soviet disinformation campaigns, but eventually conceded their real impact on national interests. In US Senate debate in 1987, the Deputy CIA Director Robert Gates, who was largely skeptical of the seriousness of information operations (so-called "active measures" by the Soviets), under questioning from then senator Joe Biden, conceded that in at least two instances, one of which was Spain's NATO referendum in 1986, Soviet information operations had a significant

impact on events. This exchange was a bellwether that eventually lead to the creation of the Active Measures Working Group (AMWG), a US agency with a mandate to combat disinformation harmful to American security and national interests. A USIA report in 1992 noted that "...Soviet active measures apparatus dwarfed, by a factor of perhaps 20 or 30 to one, the US governmental apparatus set up to analyze and counter its activities" (Abrams, 2016, p. 7). The AMWG produced targeted and effective counterpropaganda throughout the decade in which it was active but was shut down without equivalent replacement in 1992.

In many ways, this corresponds not just with the end of the Soviet Union, but with a shift in the information space. Targeted provocations and propaganda from state actors, which could be countered with effective counterpropaganda, were supplanted by a broader spectrum of epistemic threats, often from smaller actors, with more indirect and non-specific aims. In an internal US State department document released in 2015, then undersecretary of public diplomacy Richard Stengel assessed that US and allied efforts to combat Islamic State in the information environment had failed and urged the creation of a coordinated multichannel counter-messaging effort (Mazzetti & Gordon, 2015) a call for engagement in was called "memetic warfare" (Giesea, 2015). The reason why one would reach for such a term as 'memetic' is because the methods are deliberately designed to elide overt rational consideration and slip past cognition to influence deliberation and belief.

After the closure of the AMWG most of the United States anti-disinformation efforts were conducted by the Pentagon, with a much smaller effort in the hands of the State department. Some critics worried this resulted in an excessively militarized approach, and when the Pentagon's office was shut down by the Obama administration in 2009 this was greeted with approval by a range of critics across the political spectrum (Brown, 2009). This left a much smaller office that was itself shifting away from public diplomacy as a tool of influence in favour of an approach that preferred a more nebulous goal of "engagement", this was explicitly aimed at foreign audiences. When the State Department attempted to modernize its approach, it began to target social media platforms, an effort where "... the government may be trying to do the impossible, ie to plant carefully worded and controlled messages on platforms that sprang up precisely to avoid the kind of influence that the State Department seeks to exert via them." (Morozov, 2009). These digital media platforms were disruptive and subversive to efforts to control political narratives, and many observers interpreted the Arab Spring as a paradigmatic example of the anti-authoritarian effect of this new media environment. However, it soon became clear that, if anything, the new tools and technology were equally, if not more effective, in the hands of authoritarians, and as instruments to erode democracies and stifle thought, rather than nourish them. (Glasius & Michaelsen, 2018; Tufekci, 2017; Walker, 2018)

A 2020 NATO report warns of "... hybrid or memetic warfare employing social media deception, diplomatic warfare and influence operations... ", which can "... undermine, delay or frustrate [friendly] forces, nations and populations" (Reding & Eaton, 2020, p.

47). Elections have been a particularly salient site of concern, as the increasingly porous nature of the digital infosphere provided vectors for foreign interferences, decisive at times, in elections such as Kenya in 2008 (Mäkinen & Wangu Kuira, 2008), Trinidad and Tobago in 2010, (Pham et al., 2022), the United States (Benkler et al., 2018), and the Brexit referendum (Moore & Tambini, 2018). Militants involved in conflicts in Pakistan engaged in highly effective campaigns against Polio vaccination initiatives (Ahmad, 2007), and similar campaigns occurred in Nigeria (Jegede, 2007), and even before the COVID-19 pandemic, states engaged in hostile anti-vaccination disinformation campaigns (Broniatowski et al., 2018).

Information operations committed to specific narratives or ideas are being displaced by ones with more diffuse goals such as the creation and exacerbation of discord, polarization, and distrust (Martin et al., 2019; Weissmann et al., 2021). This in turn calls for a shift in defensive posture. Estonia is often considered exemplary in this regard. In response to Russian hybrid aggression that reached a crisis point in 2007 Estonia's domestic defense organization expanded the role of its cyber-defense activities to include anti-disinformation efforts that included active shaping of the information environment in addition to reactive responses to disinformation campaigns (Robbins, 2020). This whole-of-society effort to inoculate Estonia from information attacks includes policies such as the integration of media and propaganda literacy in education, not just as topics in their own right, but as lenses through which other subjects such as statistics and history are studied. Particular attention was placed on the active inclusion of Russian-speaking

Estonian audiences in the generation of programming to counteract information operations that would target this demographic and cut it off from Estonian information sources and the erode trust in these. Similar efforts have been undertaken in other Baltic and Scandinavia countries which have faced cyber and information aggression, and included the seeding of favourable narratives, the cultivation of media ecologies resistant to radicalization, deliberate debunking and pre-bunking of disinformation, and the activation of voluntary citizen efforts to combat disinformation and propaganda. I take these as canonical examples of initiatives in the service of cognitive security.

This sketch of a history of IW/IO provides some historical context as I turn to the current state of information operations. The current legal and operational norms emerged in this historical context where discourses of national defense and national security were the natural home for planning and analysis. But as we'll see in the next section, it is outgrowing these conceptions, not only operationally, but also theoretically and as a matter of public interest, and is no longer the domain of a few state actors, anchored to specific inflection points in international affairs such as declared wars. Where the OWI and the MOI were narrowly focused on specific conflict, actors, and content, and the AMWG and USIA tended to respond to specific narrative events, we'll see in the next section that modern information operations are deployed in a vast range of domains, across many channels, and often without clearly defined objectives.

1.3 From information operations to cognitive security

In the 1980's the United States developed a computational "Early Warning and Monitoring System (EWAMS)" (Hopple, 1980), to monitor communications and measure sentiment, with the hope of forecasting political conditions by mining a dataset of largely public information. The idea was that significant developments in international affairs could be predicted by analyzing content from official sources of information. At the time, there wasn't much in the way of non-official sources, certainly not that were easily tractable to mass analysis, and, at any rate, the number of people whose opinion could have immediate impact on international affairs was relatively small and largely knowable. The output of the system was considered by officials in policymaking processes, though the extent to which it was used remains classified.

As information and communications technology (ICT) has evolved, so has its reach into institutional and individual behaviour, providing even more fine-grained data and increased opportunity to test the retroactive ability for prediction methods to forecast known events - generating models that might then be applied looking forward to the future. Where EWAMS relied on a few official sources of data, later iterations of this concept, such as the Integrated Crisis Early Warning System (ICEWS) could ingest vast troves of unstructured data ranging from social media posts, economic, media, and instrumented physical systems to predict conflict and instability (O'brien, 2010). Private companies such as Google began to mine their data for predictive opportunities, an early

example of which is the "Google Flu" project, which was able to outperform the CDS's prediction of influenza by analyzing search data (Househ et al., 2017). Public data sets such as the GDELT Project ("Global Database of Events, Language, and Tone") allow researchers to experiment with their algorithms and models to generate predictions, and private industry, especially finance, collects and analyzes enormous datasets to predict market conditions (Leetaru & Schrodt, 2013). The rapid pace of digitization and ubiquitous computing continues to increase the scale and scope of the data available to build predictive systems, and capacities that were once only in reach of state actors are increasingly available to private institutions, even individuals.

There is a rapid escalation of the ability to gather data that can forecast sentiment and behaviour. But what if the goal is not just to monitor and predict but to influence? Consider the seeding and propagation of the narrative that AIDS was caused by US bioweapons programs, an example of offensive information warfare designed to influence public opinion and generate local resistance in countries hosting US military bases that reverberates today. The story originated in a Russian-controlled news outlet, the New Delhi Patriot, set up almost 20 years beforehand as a propaganda organ. (US Department of State 1987 p. 44) It developed a physical circulation of over 30,000 by the time the AIDS story was placed, and the testimony of one alleged expert, Jacob Segal, was subsequently manipulated and disseminated to bolster the claims (ibid. p. 35).

The resources required to orchestrate this campaign were enormous, and in the infosphere at the time it was difficult to reverse-engineer the story fast enough to counter it. Exploiting asymmetries in the capacity to respond appropriately and at an effective tempo is still central to disinformation efforts. Research on defensive strategy has aimed to identify the optimal inflection points for countering disinformation (Wardle & Derakhshan, 2017b) - strike too early and you amplify the story, causing some people to form false beliefs on the strength of unintentionally misleading higher order evidence. Strike too late and it can't be dislodged (Brashier et al., 2021). The contemporary information environment offers far more sophisticated avenues for the dissemination of disinformation and requires far less resources to do so. Chapter 5 will discuss in detail some specific epistemic effects of anti-disinformation efforts.

Early examples of active measures frequently involved the careful construction of forgeries, often with painstakingly assembled contextual materials. As the volume of information operations increased, so did the range technical affordances for the gathering of data and the production and dissemination of narratives and evidence. Forgeries, and other kinds of disinformation, don't need to fool experts if they can be distributed directly to the public, and elaborate covert networks of promulgation are unnecessary if the public can be enlisted to amplify and reproduce them. As opportunities for spreading disinformation have expanded and increasingly disintermediated gatekeepers that could be focal points for countermeasures, the field of concern for defensive efforts widens to include a great deal of behaviour that would not ordinarily be considered part of the

security frame, especially in states with strong protections of rights to free speech, association, and publication.

The extent to which many information operations techniques rely on the unwitting or at least partially unwitting cooperative epistemic activities of the public (Rini, 2017; Starbird et al., 2019) brings into scope the structure and governance of the communications platforms and networks that enable this. Social media platforms have a mix of public and self-interested incentives to prevent manipulation, and platform design can generate both resilience and vulnerability to hostile efforts. Subtle changes to communications platforms can have far-reaching effects, and where once it was possible for the MOI to arrange a quiet word with the editor of a major newspaper, meaningful interventions with communications networks are difficult to orchestrate and have effects that are hard to forecast and monitor. A major motivation for cognitive security interventions is the worry that past some point of degradation, it might become impossible for authoritative voices to cut through the noise in an emergency.

Alongside the advances in digital ICT have made it easier to deploy information operations, there are improvements in technologies that make it easier to predict the likelihood of decisive impact. The development of behavioural economics, and the subsequent theoretical and political efforts to apply these theories to business and public policy objectives, is based in part on the claim that we can take into account the impact of familiar sorts of cognitive bias (Thaler & Sunstein, 2009; Tversky & Kahneman, 1973)

when developing public policy and instruments for its implementation. If we want people to do something, there are subtle interventions that can be crafted, so-called "nudges", that will cause aggregate behaviour to conform, even without strong coercion, incentives, or disincentives. For example, if you want to reduce the amount of fake news being shared on a social media platform, you can modify the interface in ways that demonstrably reduce the likelihood users will share a dubious story (Pennycook, McPhetres, et al., 2020; Roozenbeek et al., 2021) without actually rationally engaging with the user. Rather than give reasons why the story is or is not accurate, you alter the interface to prompt the user to consider accuracy before they are able to re-publish it.

Sunstein and Thaler argue at length that the insights of behavioural economics have substantial implications for both the articulation of policy objectives and the ethical justifications and constraints on various policy tools. They argue that policymakers can and should exploit cognitive bias by way of the manipulation of choice architecture to improve public decision-making. Dedicated government offices such as Australia's Behavioural Economics Team, the UK's Behavioural Insights Team, Singapore's Human Experience Lab, and the US Social and Behavioral Sciences Team, bring expertise from behavioural economics directly to bear on government policymaking. Businesses now use increasingly fine-grained and comprehensive data to analyze users and expose them to persuasive efforts to influence their behaviour, and increasingly offer this service to actors in the political sphere. I'll discuss nudges (and their broader theoretical grounding in libertarian paternalism) in the much greater detail in the next chapter, for now, I just

want to highlight the extent to which the prospect of changing people's minds, and behaviour, without them necessarily knowing and understanding that they have been influenced, has been widely viewed as practical, effective, and justifiable, even where there is disagreement about the extent of the influence and the conditions of justification.

Behavioural economics offers a collection of methods, based the study of individual and collective cognition and its failure modes (in Chapter 2 I call these 'cognitive strategies'), and a related body of literature that studies how to use leverage these to specific ends, to assess the results, and which attempts to delineate the kinds of purposes it can justifiably be used for. As this field has developed, so has the scale and breadth of datafication and the sophistication of analytics technologies. Many digital platforms afford the aggregation of data from multiple sources, which can include payment processors, credit reporting agencies, location data from telecommunications providers, and data collected ambiently from Internet of Things (IOT) devices and smartphones as well as that collected actively in the use of digital services. When combined and linked this can enable highly granular analysis and predictions of behaviour and can help to guide and assesses influence efforts. We might permit an app to track our music listening habits to recommend playlists but be surprised to learn this data has been used by a third party to make inferences about our mental health(Allen Anderson, 2015; Anderson et al., 2021), which can be labelled and enriched with other data, and made available for sale. Data from a wide range of sources can be re-taxonomized as health data (Martinez-Martin et al., 2018) and used to make psychographic inferences that can in turn inform

persuasive messaging - illustrating the interconnections between technologies of surveillance and influence. Biometric data such as facial images and voice samples are not only used to track individuals across systems, they are also used to infer physical, psychological, and political traits, capacities which are not just used for marketing, but also used in campaigns designed to influence public sentiment (Bakir, 2020; González, 2017). Distinctions between the commercial and political use of these technologies, and domestic versus foreign actors and employments, are increasingly murky and analytically unhelpful as so much of this data is generated and processed trans-nationally, and political activity conducted by sub-state and non-state actors.

We can see then that three connected disruptive factors - opportunity, capability, and technique, increase the apparent vulnerability of public beliefs to hostile influence. Digital ICT creates opportunities to intervene, generates data to inform and measure interventions, and both factors produce empirical feedback that helps to improve the cognitive techniques that inform influence strategies. Whether these techniques are effective, and to what extent and in what conditions, perhaps surprisingly, is somewhat orthogonal to questions about their relevance to cognitive security concerns. Despite huge investments in active measures over decades, the Soviet's viewed them as "... instruments in a larger game, as mere chess pawns, but capable of damaging opponents at the margins and perhaps opening the way for larger gains later" (Kinahan, 1990, p. 305). Even when marginal, the ability to unsettle public views on some topic, or to shift perceptions of authority and expertise, can have effects that are desirable to adversaries. Hostile

disinformation campaigns that do not decisively shift majority sentiment often have the result of increased partisan polarization (Corstange & Marinov, 2012; Tomz & Weeks, 2020), which can itself be a motivation for interference.

Similarly, whether or not microtargeting is effective for its immediate aim, the mere use of it has strong effects. The microtargeting of advertisements, especially during election campaigns, is widely touted by practitioners and critics alike as powerfully effective. One of the few currently available meta-analyses suggests that "...microtargeting exceeds the persuasive impact of alternative messaging strategies by an average of 70% or more."(Tappin et al., 2022, p. 6). Although other empirical results are mixed, it serves the interests those who provide and wield these techniques to exaggerate their strength (Baldwin-Philippi, 2019), but, even if they are not as effective as advertised, merely knowing that these technologies are being used, and believing that they are influencing some, is itself a vector for manipulation and the deliberate creation of epistemic effects. As Rini (Rini, 2021) argues, Russia's use of the Internet Research Agency to interfere in the 2016 US presidential election undermines US interests and furthers Russia's interests, even more by being discoverable. More powerful than subtly changing one's mind is letting it be known that one might do so, that one might actively be doing so, and as a result one should not believe one's eyes. The use of sophisticated microtargeting, even when it is not effective, can be shown to undermine trust in democratic institutions (Matthes et al., 2022) in exactly the way Rini predicts, by undermining democratic testimonial networks. That there is any fake news at all gives a

much sharper edge to cries of "fake news" when an unflattering story about one's co-partisans should surface - as though Descartes's evil demon was not just a bare possibility but tucks a calling card into our morning newspaper now and then. Even if one is a diligently critical reader, the scale and complexity of digital information environments makes it impractical to wield one's skepticism effectively, to bring into scope all of the features of the system that are potentially relevant to the evaluation of a piece of evidence.

Contested attribution further complicates predicted epistemic effect - foreign disinformation campaigns that would ordinarily be expected to cause discord can in fact have a unifying effect when the attribution is widely perceived to be hostile to shared values (Bauer & Wilson, 2022) Even where we might doubt that some particular effort is likely to have a significant effect, there are good reasons to worry about distal harms. Low-circulation attempts at memetic warfare, such as the "Posing as Patriots" network which attempted to influence the 2022 US midterm elections (Ronzaud et al., 2022) can not only seed narratives that can in turn jump between platforms (Krafft & Donovan, 2020), but their discovery itself provides an evidential basis for allegations of inauthentic behaviour, including more serious ones that are more commonly the domain of hostile interference, such as the manipulation of evidence involving public figures such as politicians and journalists. Evidence for some attempted manipulation justifies more general suspicion of evidence. The increased prevalence and visibility of this activity, and defensive responses, fosters an 'everyone is doing it' atmosphere that can rationally

ground distrust in evidence and media systems. This "normalization of deviance" (M. Innes et al., 2021), as instigated by malefactors, percolates into domestic politics as the tactics of digital influence engineering become both materially available and seemingly less extraordinary.

Influence over the operators, moderators, and regulators of information system is an even higher value target for hostile actors. In April of 2022 the US government announced the formation of the Disinformation Governance Board, with the stated aim of providing education and consultation to assist government agencies with information security concerns, and immediately attracted furious criticism accompanied by the online propaganda and influence efforts that now typically coalesce around polarized, and polarizable, issues, including violent and misogynist attacks on the board's Executive Director Nina Jankowicz (Lorenz, 2022). Critical attention focussed on the concept of disinformation itself, framing it as inherently subjective and partisan, and, consequently, government efforts to control it as illegitimate meddling in protected speech. The Disinformation Governance Board was paused within weeks and then shuttered, leaving the prospects for an accepted operational definition of disinformation, let alone defensive counter-strategy, looking bleak. In response, (Murphy, 2023) argues that we should redouble efforts to develop a more acceptable conception of disinformation but the problem appears to run deeper. It is difficult to see how conceptual engineering of the concept of disinformation can mitigate disinformation attacks on the social coordination of this very act of engineering when the information environment in which such efforts

must succeed is fractured, polarized, and where the functionally operational norms are non-epistemic and even adversarial to epistemic norms. This is one motivation for my interest in reframing the question, from "what is disinformation" to "what is a healthy information environment" - the hope that this offers a deeper and more ecumenical conceptualization of the legitimate aims of cognitive security.

The use of cognitive strategies developed and deployed using big data and digital intermediations offer expansive opportunities to influence beliefs. A variety of terms have been used to describe such capabilities, "persuasive design" (Fogg, 2002), "organized persuasive communication" (Bakir, 2020), "digital influence engineering" (M. Innes et al., 2021). Explicitly hostile applications of these techniques can be described with a connected family of terminology and concepts that include psychological warfare, information warfare, memetic warfare, information operations, strategic communication, psychological operations, civilian public diplomacy, propaganda, counterpropaganda, hybrid warfare, and cognitive warfare. Some of these are more properly considered as more narrow concerns in security domains, or have constrained meanings, and others are hard to distinguish at all. In some cases, distinctions that might apply to their offensive use are immaterial to activities that aim to defend against them. None of these terms has gained a dominant foothold in the policy literature, and in a review of exactly this problem of terminological proliferation and overlap, the authors argue that "...the growing number of overlapping terms is symptomatic of a weak understanding" (Wanless & Pamment, 2019, p. 8), an assessment I'm quite sympathetic to.

As early as 2013 the Canadian Department of National Defence observed that "... sociotechnical networks will continue to offer opportunities for foreign states to perform influence activities against the interests of Canada and its allies" (Department of National Defence, 2014) The Potomac Institute for Policy Studies, the successor to the US Office of Technology Assessment, charged with studying science, technology and national security issues, warns in a recent report that "... in today's world, it is necessary to combat adversarial use of perception management, disinformation, and strategic deception." (Pearson & Moxham, 2022, p. 5). In the UK, writing about the application of big data and artificial intelligence to amplify hostile influence campaigns, an advisor to several UK security thinktanks and the prime minister's office, argues that "... the West must respond – and it will likely mean turning the tools used to manipulate into the weapons of defence." (Dear, 2021). Similar calls and warnings are increasingly present in memos, articles, and conference proceedings from defense and security professionals, especially in NATO countries.

A name and some conceptual scaffolding for this kind of defensive response can be found in a 2017 Rand Corporation Report to the US Senate Armed Forces Committee, where a distinct field of national security study is proposed - that of "Cognitive Security" (Waltzman, 2017). Cognitive security is focussed on distinct security vulnerabilities that, while they often accidentally coincide with economic, biological and cyber security concerns, are of their own unique kind. Goals include protecting individuals,

communities, and states in a future where "researchers, governments, social platforms, and private actors will be engaged in a continual arms race to influence" (Waltzman, 2017, p. 7). The report defines the field of cognitive security as one that will ".... bring together expertise in cognitive science, computer science, engineering, social science, security, marketing, political campaigning, public policy, and psychology to develop a theoretical as well as an applied engineering methodology for managing the full spectrum of information environment security issues" (ibid.). Since 2017, the term has slowly been adopted and can be found in academic and professional literature on warfare (Claverie & Du Cluzel, 2022; Le Guyader, 2022; Nesic, 2022), philosophy and information systems (Hassan et al., 2018), cybersecurity (Smith, 2023), and disinformation (Hung & Hung, 2022; Terp & Breuer, 2022).

I want to linger briefly on a terminological distinction that helps illustrate the unique challenges and epistemological dimensions of cognitive security. This is the distinction between cognitive warfare and information warfare as developed by (Rogers, 2021). On this view, information warfare is a contest for information anchored to specific interests and events - for some definable bundle of knowledge, perhaps about a political figure, an economic initiative, the cause of a pandemic, or the military ambitions of a rival. One view of the strategic position of the United States and its allies at the beginning of the digital media era was that it held a "... comparative advantage in its ability to collect, process, act upon, and disseminate information, an edge that will almost certainly grow over the next decade." (Rogers, 2021, p. 83). Of course, the production of knowledge

from information is not a mere technical undertaking, it is also a collective epistemological project in which sense is made of information, trust in expertise relied upon, contexts and frames of reference established, and supporting media ecosystems created and maintained. The shift Rogers observes is one of emphasis from the former to the latter aspects of the production and accumulation of knowledge, especially by adversaries who sought to undermine this sociotechnical information advantage by attacking the softer targets of its epistemic dependencies.

For instance, analysis of so-called "political micro-influencers" deployed by Russian state-affiliated actors early in the 2022 invasion of Ukraine (Boichak, 2023) shows that much of the activity involved the production of relatable identities and the positive framing of ideas that indirectly support the contextual setting of more overt Russian messaging efforts to justify and explain the war. This fits well with the way Rogers describes cognitive warfare as a conflict of sensemaking. "...IW involve actors contending over information within specified and assigned contexts in which the orientation of the context to the contending is settled. CW, conversely, involves actors contending within unspecified and unassigned contexts, in which the orientation of the context literally is the contest. (Z. Rogers, 2021, p. 87) This is one reason why fact-checking as an anti-disinformation strategy founders in conditions of cognitive warfare. Facts do not wear their content and their veracity on their sleeves, but require interpretation, and are only verifiable through social processes that require agreement about context, and are only factual in relation to some frame of reference and shared

activity. A study of Birdwatch, the crowd-sourced fact-checking system used on Twitter, notes that "...context features, and specifically partisanship are highly predictive of both misleadingness and helpfulness ratings..." (Allen et al., 2022) substantially more so than content features of the tweets such as sentiments and sourcing.

Scholars of digital information systems have observed the phenomena Rogers has identified in the form of context collapse (Frost-Arnold, 2021; Marwick & boyd, 2011). Rogers argues that, quite apart from the operations of hostile states, the manipulative behaviour of commercial actors is substantially responsible for a crisis of trust in the digital information environment. "The net effect of the fragmentation and disutility of the information environment is not merely one of many more contested narratives. It is of no narratives." (Rogers, 2020). Whole-of-society defensive strategies such as democratic deterrence (Wigell et al., 2021), or Finland's "Security Strategy for Society", are subverted if the informational and cultural threads that bring people together are broken.

In this project I'll use the broader term "epistemic engineering" to capture both friendly nudges and hostile influence as well as the defensive and remedial efforts that arise in response and employ more specific terminology to distinguish particular acts of epistemic engineering along dimensions such as the intentions and reasons that govern their exercise. In the next chapter I'll use the concept of paternalism to focus on epistemic engineering that's done without our consent, for our own benefit, and explore a

conception of stewardship that might be appropriate for acts of epistemic engineering that aim to secure collective goods such as the health of the information environment.

This section has been concerned with describing the conditions that lead to the development of cognitive warfare and cognitive security as concepts and describing these. I've argued that while IO was once anchored in the specific concerns of nations in overt conflict, it now describes an atmosphere of contestation and vulnerability in the information environment. Even when the threats emerge by accident, from the operations of commercial actors pursuing their own ends in digital information spaces, social and political risks emerge. The scope of activity that IO threat models must consider has expanded considerably and continues to do so. The tactics, tools, and procedures used by explicitly hostile, state-backed entities increasingly overlaps with those used by friendly entities, domestic actors, and corporations and other commercial bodies. The extent to which any of these actors can be clearly designated along the foreign/domestic axis is increasingly unclear, particularly for the purposes that regulations and norms around IO are concerned with (Marangione, 2021). Collectively, we are forced to contend with capacities and opportunities for epistemic engineering that have a reasonable likelihood of creating significant effects on the beliefs and sentiments of individuals at varying levels of demographic aggregation - and even when efficacy is unclear, the underlying goals can be distal and indirect in ways that allow them to be advanced by campaigns that considered in isolation appear to be unsuccessful. Worse - it is difficult to devise defensive strategies that don't exacerbate the problem.

1.4 Ethics and cognitive security

A sketch of the territory of philosophical interest here can begin with the observation that we are faced with challenges of both identification and implementation. Firstly, consider the problems of implementation. Roosevelt's OWI was expected to address three complaints, "...first, that there was too much information; second, that there wasn't enough of it; and third, that in any event it was confusing and inconsistent" (Weinberg, 1968, p. 77). These problems will be familiar to those who study the epistemology of digital information systems, and the challenges they pose to practices of mass communications, often which involve the problem of ensuring signals of trust and reliability can be made salient in noisy communication channels. This framing falls on the IW side of Roger's distinction - there is some factual message that we broadly agree is factual and the problem is aiding its communication and recovery. Within this paradigm, epistemologists might cash out the notion of signal and noise along the veritistic lines suggested by Goldman (Goldman, 1999) - true propositions are the signal, false propositions are the noise, and the processes of individual and group belief formation that reliably fix on the most true propositions and the fewest false ones are best. Good epistemic engineering might seek to facilitate maximal production and uptake of these true propositions.

However, at some threshold of scale and kind, the noise, and the strategies we must use to function in spite of it (for example, hastily delegating epistemic authority to

gatekeepers - a strategy that can be undermined by what I call trust-herding in Chapter 5), we find that we face the more difficult problems of cognitive warfare and security. There is a fragmentation of contexts, narratives, and epistemic dependencies that diminishes the auxiliary consensus that we would mobilize to solve signal problems. The challenge of cognitive warfare is that how we give meaning to information, the trust we have when encountering the product of interpretative labour from specialists, and how we assess significance, becomes destabilized, can be manipulated at a tempo that outpaces remedial efforts, and can thus subsequently interfere with the coordination of collective action. There is not just a diminishment of political and collective autonomy, as national security discussions have tended to understand the problem, but also our individual autonomy. I argue in Chapter 3 that our epistemic autonomy is relational and can be limited and undermined by conditions in the epistemic environment, such as bad ideology (Srinivasan, 2020), oppressive socialization (Benson, 1991), and epistemic pollution (Levy, 2018).

The interventions that we might reach for to improve the situation will all have complex epistemic effects that require specialized expertise to forecast. Some of these effects may have obvious ethical dimensions, perhaps depriving some people of information they have a right to encounter or providing information that can be expected to mislead and create harms for others. Others will have knock-on effects that undermine the epistemic interests they were designed to serve. The 2022 Russian invasion of Ukraine spurred information operations targeted at allied nations to influence public

approval for support of Ukraine, but some critics observed mission creep from the prevention of the circulation of falsehoods to the deliberate suppression of unhelpful truths, even among mainstream information source (Rimbert & Halimi, 2022). This risks erosion of trust to the very communication channels whose efficacy such policies depend on and creates opportunities to destabilize and undermine the operational value of concepts such as disinformation. Similarly, clumsy handling of the management of narratives about the origin of COVID-19 on social media platforms, where for a time there was direct filtration and censure of all content probing this on at least two major platforms (Shir-Raz et al., 2023), provided evidence that support conspiratorial claims about censorship and government control of speech. This undermined not only the immediate goal of protecting the information space for public health interventions but revealed the invisible hand of what Douek (Douek, 2020) calls "content cartels" that control discourse on social platforms, often with formal and informal liaison with political entities, with policies that are largely secret or unarticulated, and enforced with procedures that lack transparency.

A pointed way to put the implementation question is to ask - what is the epistemic equivalent of the concept of collateral damage in kinetic and diplomatic contexts? Suppose on some theory of just war and the acceptable exercise of information operations capabilities that a state targets the cognitive domain of an adversary for a putatively justifiable purpose. Distrust and disinformation is not a precision weapon, and we currently lack organized efforts to forecast the likely consequence of interventions. For

example, clumsy efforts to promote vaccines and combat disinformation have resulted in falling rates of all vaccinations (Eisenstein, 2022). Identifying and managing collateral effects is critical to any defensive strategy against IW and CW. In Chapter 2 I'll consider questions about what kind of interventions we might make to advance collective epistemic goals in light of the literature on epistemic paternalism. When can we interfere with the epistemic efforts of others for their own good, and what justificatory conditions obtain?

Secondly, consider the problem of identifying which kinds of content and channels are legitimate targets for defensive action. How can we identify some subset of epistemic infrastructure and behaviour as appropriate objects of security interests? The COVID-19 pandemic revealed a willingness across many nations to engage in coordinated constructive, defensive and hostile, information operations to bolster domestic public health or destabilize that of adversaries. This suggests that many states adopt an expansive view of the kinds of information and channels that might be of sufficient concern to the public interest to justify coordinated intervention. In Chapter 4 I'll consider collective interdependence as a condition on justification for such efforts, and I will argue that the kinds of epistemic interdependence identified by recent work in social epistemology is supportive of collective public interest in the information environment. In Chapter 5 I adopt Wittgenstein's conception of hinge commitments as a way of describing the structure of this interdependence and locating critical aspects of it.

It is difficult to give cognitive security the sort of airtight definition philosophers often devise and expect. One substantial problem is the extent to which the concept of security itself demands interrogation - whose security, and from what? Discourse on cognitive security that exists today is heavily influenced by its origins in security think-tanks and adjacent literatures, and often framed as an attack on "open societies", which, even if we agree are worth fostering, often distribute openness and its benefits unevenly and unfairly. Cognitive security is often viewed through the lens of societal conflict and great power competition, which can give the appearance that a baseline partisanship and value-ladenness is unavoidable. Former CIA Director General Michael Hayden voiced concerns that cognitive warfare vulnerabilities threaten "...many of the premises on which we have based our governance, policy, and security" (Hayden, 2019). One might worry that a full-throated defence of these will founder on the problem that, as one Reagan speechwriter put it, "... American democracy is less a form of government than a romantic preference for a particular value structure" (Wirthlin 1985). Critics of the securitization of public health (DeLaet, 2014) and securitization more broadly observe that the expansion of the security frame tends to reproduce and project power and economic relations in ways that can be harmful and which import domain-inappropriate values.

The problem of securitization is important but sustained engagement with it falls outside the scope of this project. Firstly, because I am interested in the ethics of cognitive security as it exists today - and the securitized framing is operational in most of the domains in which cognitive security ends are pursued as such and is therefore a condition

of current information environment. A security orientation also helps distinguish it from the individual-harm based justification for intervention and regulation of information environments, a much more developed area of theoretical and applied scholarship, but one with very different interests. Content including representations of abuse, threats of violence, violations of privacy, the reproduction of patented or copywritten material, have all been regulated or deemed regulable by governments, companies, and institutions, but not for the kind of directly epistemic reasons that motivate worries about cognitive security.

I think that security is actually a perspicuous way to demarcate cognitive warfare concerns from the larger domain of the ethics of epistemic engineering. It should be clear from the previous sections, especially the way Rogers distinguishes information warfare from cognitive warfare, that what is distinctive about cognitive vulnerabilities is the extent to which they aren't part of traditional security discourse and concerns. At some inflection point they nonetheless become relevant to security. Not only does this property define cognitive warfare, but it is also the major challenge cognitive security confronts - what kinds of epistemic vulnerabilities demand institutional response, and by which institutions? A narrow view, wary of creeping securitization and militarization, faces the challenge of the ubiquity, novelty, speed and scale of threats, emerging from outside traditional security domains.

A broad view, on the other hand, risks authoritarianism and an oddly limited way of looking at information spaces that doesn't square with conceptions of normal scientific and cultural practices and individual liberties. Information systems aren't just there for official business, authorized users, and serious inquiry, and their proper functioning can't always be described in terms of the maximal production of true propositions. Cognitive security is concerned with properties of, and phenomena within, information environments that are dangerous not just because of the particular content that is associated with them. Writing about the difficulty in crafting ethical and legal principles that could govern social media platform content moderation, Grimmelmann notes that the very same meme, occurring in different contexts and at different scales, can be treated as "... a practice, a parody of the practice, and a commentary on the practice" (Grimmelmann, 2017, p. 222), each of which demands different handling. Throughout this project I develop an argument that we need a different way of describing the nature of these risks that isn't anchored to known incidents, or poorly defined pollutants such as "fake news", and which don't presuppose a specific ideology.

In the coming chapters I explore some conceptual engineering that can yield less politically valanced content for cognitive security. This offers a less securitized perspective, and one less prone to nationalism. Rogers warns that in response to cognitive warfare we "... need to develop an understanding of a heterogenous type of cognitive violence which can be at once public and deeply private, non-lethal and highly destructive to human intellectual, emotional, and psychological states [...] the type of

cognitive violence in mind can easily cause major disruption in the normal functioning of societies as well as significant changes in behavior without being assigned a specified meaning." (Rogers, 2021, p. 87). This concern suggests philosophical questions we can ask about the kinds of epistemic conduct, institutions, and infrastructure that are responsive to human needs and conducive to a good life that are substantially less loaded with overtly political content. We have an epistemic environmental dependency on the mediums through which we encounter each other, and this affords a perspective from which we can ask "What is a healthy information environment?". The answers we can give to this question provide an orientation for cognitive security that is grounded in our collective epistemic needs and interests, rather than the desire to eliminate some particular class of content.

In the next chapter I'll consider a challenge that can be raised against efforts to deliberately engineer our epistemic environment to pursue cognitive security goals. On some conceptions of epistemic paternalism, these efforts are paternalistic in that they interfere with the free exercise of inquiry, speech, and epistemic agency more broadly, for the purposes of improving our own wellbeing, where in fact we should be left to make these determinations for ourselves.

Chapter 2 - Epistemic Paternalism

This dissertation is concerned with cases where, in the pursuit of cognitive security and resilience to information warfare, state power is used to promote and conduct deliberate engineering of the epistemic environment to improve it, to make it safer, to decrease tendencies to disinform on significant matters. But the very idea of this sort of engineering has struck many philosophers as potentially paternalist, and possibly objectionable and unjustifiable as a result. The goal of this chapter is to consider, and largely dismiss, concerns that many forms of epistemic engineering will be objectionably paternalist.

I situate worries about epistemic paternalism within the broader context of philosophical thinking about paternalism generally, with the aim of finding precise conditions for the application of the concept, and clarity about what is at stake when it should be applied. The literature on epistemic paternalism often employs the notion of paternalism in ways that come into tension with the conceptual structure of the term. Many forms of epistemic management have some, but not all, of the attributes of paternalism, and objections motivated by anti-paternalism should not be applied to them. Anti-paternalism aims to protect a distinctively individual form of autonomy that does not map well to our interests in the shared information space, and I will argue that worries about epistemic paternalism are often expressed in ways that cast an implausibly broad net.

Some kinds of epistemic engineering may be justified in spite of the appearance of paternalism, for one of two reasons. In many cases, they should not be considered paternalist at all, and are instead forms of self-binding or stewardship. A goal of later chapters will be to develop a conception of stewardship of the information environment that describes these cases where deliberate efforts are made to alter epistemic systems in pursuit of some public good. This category would encompass many of the kinds of interventions often thought to be potentially paternalist, but which this chapter argues should not be considered such.

Most interventions we might undertake in the service of cognitive security goals are better described as forms of self-binding, management and stewardship, which lack much of the ethical and justificatory baggage of paternalism. This is important if the category of epistemic paternalism should mark out some class of activity as objectionable, or as requiring special oversight or justification. I find that there is a smaller class of cases that can exhibit the hallmarks of paternalism. Institutions with substantial epistemic power can implement policies that affect the exercise of epistemic agency by individuals not only to avoid harms to others, but also to improve their epistemic wellbeing. In later chapters I'll show why I think these tend to become unjustifiable on other grounds.

2.1 Definitions of paternalism

Scholarly interest in paternalism has largely been grounded in the work of Mill, and subsequent political and legal theory that invokes the harm principle as a limit on state power, where prevention of harm is taken to be a potentially legitimate justification for state interference with liberty. Prohibition on paternalism places a condition on this – that the harm must be to someone other than the person whose liberty is violated. A first cut at a definition of paternalism is that an act or policy is paternalist when it interferes with a person's autonomy for their own benefit. At the extremes, interference with the commission of murder is thus easily justified, and interference with suicide is not. There's a prominent tradition in political philosophy that views paternalist interference as inherently objectionable. In this view, the boundaries of the paternalistic are the boundaries of the acceptable exercise of coercive power.

This simple definition is elaborated and expanded in the literature on paternalism in response to two pressures. One to broaden the scope and handle cases where it seems there is paternalism, but some more narrow definition would not capture it. The other to constrain the scope, blocking attributions of paternalism in cases where it is arguably unwarranted. These reflect two motivations for using the term, one to describe the limits of law, and the other to express a moral concern for autonomy, and in following sections I try to disentangle these in order to get a clear view of what might be objectionable, or regulable, about the kinds of epistemic engineering this dissertation is concerned with.

Gerald Dworkin has developed a definition of paternalism over the course of several articles and books that has been widely influential. His first formulation emphasized interference and self-benefit, "... the interference with another's liberty of action justified by reasons referring exclusively to the welfare, good, happiness, needs, interests or values of the person being coerced..." (Dworkin, 1972, p. 67). However, this definition has trouble handling cases where the subject has not expressed a specific preference at the time of interference. Gert and Culver (Gert & Culver, 1976) provide a counterexample, where the subject objects to blood transfusions on religious grounds, becomes unconscious, and as their condition deteriorates a physician decides that a transfusion should be administered. This does not interfere with liberty of action, as there is "... no attempt to control behaviour, indeed there was no behaviour to control" (ibid. p 46). The problem here is that there are cases where our moral intuitions are such that the charge of paternalism seems apt, but where the interference cannot be said to violate autonomy in a straightforward way. The subject may not have preferences about the exact case, or not be in a position to express them, may not have had the opportunity to reflect on them, or the preferences may be faulty in ways the subject would or should appreciate with prompting. Hershey (Hershey, 1985) argues that a definition of paternalism must therefore include non-consent as a condition, which helps capture interventions that the subject is not aware of (and therefore, where the intervention can't be said to be against their will). Instead, Hershey offers this account:

An action, x, initiated by A (an individual or group) with regard to B (another individual or group) is paternalistic if and only if:

(I) x is primarily intended by A to benefit B,

(II) B's consent or dissent is not a relevant consideration for A. (ibid. p. 179)

Clarke (Clarke, 2002) likewise attempts to capture cases where preferences aren't directly and coercively ignored, where the paternalist act "... aims to close an option that would otherwise be open to Y, or X chooses for Y in the event that Y is unable to choose for himself; and 2) to the extent that X does so in order to promote Y's good". (ibid p. 82). Dworkin published several subsequent articles refining and defending this conception, which resulted in the addition of a non-consultation condition. To the extent that there is a canonical definition of paternalism today, it is probably Dworkin's recent version:

X acts paternalistically towards Y by doing (omitting) Z if and only if:

1) Z (or its omission) interferes with the liberty or autonomy of Y.

2) X does so without the consent of Y

3) X does so just because doing Z will improve the welfare of Y (where this includes preventing his welfare from diminishing), or in some way promote the interests, values, or good of Y. (Dworkin, 2015, p. 21)

There are three core components - interference, non-consent, and self-benefit. Debate about whether some proposed intervention should count as paternalist tends to focus on the extent to which it truly violates autonomy, and whether consent is really given. Concerns about autonomy, agency, and consent become tangled when we are called to discern one's rational preferences or ideal preferences from instances of non-ideal expression, especially when their expression would result in actions that appear to be clearly detrimental to the agent. For example, some argue for a distinction between means-paternalism, which interferes with the way people pursue their goals and ends-paternalism, which attempts to change those goals or prohibit their pursuit. This distinction is also invoked by (Kleinig, 1983) as that between positive and negative paternalism, and by (Dworkin, 1981) as that between weak and strong. We can also find it in (Raz, 1986), who argues that means-paternalism is acceptable since this only prohibits actions that conflict with more important ends the agent values. Confronted with a subject planning some risky activity, where means-paternalism might require that they sign a waiver, use a safety device, acquire a license, or demonstrate training, the ends-paternalist seeks to prevent the activity.

Le Grand and New observe that this distinction can be wielded in ways that obfuscate significant disagreement about what counts as means versus ends, as they put it "... people are almost always striving to achieve an appropriate balance between largely uncontroversial ends" (Le Grand & New, 2015, p. 36). One might agree with one's ski companion about the value of reaching the bottom of a slope safely but disagree at the

extent to which speed or style also matter. Dworkin argues that there is a risk that ends-paternalism is self-undermining - that if the subject of interference cannot agree with the value of the end that has been substituted by the paternalist for their own, then the intervention cannot be said to target their self-benefit. I'll return to this issue in Section 4 where I will consider the extent to which endorsement of ends is required for epistemic interference to count as paternalist. This is important because if the subject cannot agree that some epistemic improvement, they receive is beneficial, a claim that the intervention that created the improvement is paternalist is undermined.

This third condition, that of self-benefit, primarily signals the special protection from the application of the harm-principle that anti-paternalism aims to offer. Laws preventing driving while under the influence of alcohol are justified by the prevention of harm to others. Laws mandating the wearing of seatbelts are widely thought to exemplify paternalism, because the harm they prevent falls exclusively to the subject of the mandate. It has been frequently noted in the literature on paternalism, especially in legal contexts, that even in the seemingly most clear-cut examples of paternalism, the subject of interference is rarely the sole beneficiary, and the extent to which a benefit should count as a benefit to self can become contested. Laws against smoking don't benefit the individual if it happens to the be case that they were not destined to be harmed, and these laws do benefit others, by lowering health care costs and preventing harms to others via second-hand smoke. Seatbelt laws arguably help lower insurance premiums for all and reduce burdens on the healthcare system. Charges of paternalism almost universally

exhibit this tension between concern for a kind of autonomy taken to be fundamentally individual, and conceptions of harms and benefits which don't neatly fit into this individualist framing.

Concerns like these both motivate and challenge a common distinction between pure and mixed paternalism. Pure paternalism occurs when the subject of interference and the intended beneficiary of the interference are significantly aligned (if not identical). Mixed paternalism aims to provide benefits to both the subject of interference and to others. The rarity of instances of pure paternalism, and the near ubiquity of mixed examples in the literature on epistemic paternalism, raises conceptual problems for anti-paternalism. When benefits are mixed, in many cases the harm principle will suffice to justify the intervention in ways that aren't challenged by anti-paternalist arguments. This can be particularly common in policy contexts, where regulations do not always wear their intentions on their sleeves, and the extent to which the aim of the intervention can correctly be described as paternalist becomes contested. Childress (Childress, 2020) observes that such disputes can be well-grounded, because policymakers wary of charges of paternalism are incentivised to produce regulation in ways that emphasize the protection of harm to others. As public health values have gained increasing acceptance as grounds for justification in recent years, Childress observes that there is a temptation to "... highjack the language and norms of public health to cover private harms and thus to invoke public health rather than paternalism to justify proposed interventions." (Childress, 2020, p. 46)

To what extent does intention matter? We might act with the intent to circumvent a person's agency for their own good, but we might also act in a way that did not intend to do so but has that effect. Disputes about facts can manifest as disputes about paternalist outcomes. A mandate to wear a face covering to prevent disease transmission can be argued to have a paternalist outcome if the facts are such that masks only protect the wearers, and do not protect others from transmission. If this was found to be the case, the appreciation of this fact by policymakers would be the litmus test for paternalist intention. But no matter the intention, these facts change the kinds of reasons that can be said to legitimately justify the policy.

A policy might be paternalist even if not intended to be so. De Marneffe (de Marneffe, 2006) argues that a policy is paternalist when "... it cannot be justified by non-paternalist reasons alone, and, because the government adopts it only because someone in the relevant political process takes some paternalistic reason as sufficient to justify it" (ibid. p. 70). But it is not clear why claimed justification should matter. If the legitimate reason for some policy is paternalist, then it doesn't matter what other non-legitimate justifications, motives, reasons, and such might explain why the policy was adopted. A policy can thus be paternalist even if we didn't intend to be, and even if at the time the facts were such that it could be justified on non-paternalist grounds. De Marneff thinks this consequence is unwelcome, because it would not capture policies that aim to be paternalist, but that are ineffective, but I think this conflates concerns about the history of some policy with concerns about its substance and justification. If the policy is

ineffective, it not justifiable as a policy, quite apart from the extent to which it might have been paternalist had it been effective. The extent to which motive, justification, and outcome matter in determining if some act is paternalist is particularly important in Section 5, where I consider systemic interventions where results are hard to predict.

A paternalist intervention can be instrumental in the sense that the provision of self-benefit might be accidental to the goal of the policy to prevent harm to others. This is converse to the mixed/pure distinction, where the mixed act aims to benefit the subject of interference, but where the intervention can't be feasibly implemented in a way that only affects the beneficiary. In the instrumental form, a paternalist policy might be the best way to prevent some harm to others. Shiffrin (Shiffrin, 2000) gives this example, "Suppose a park ranger has the power to refuse permission to climb a steep, dangerous mountain path. ...Suppose the ranger says, `Of course, you make take whatever risks you want to with your life, but I refuse permission because you might die and leave your spouse grief-stricken" (ibid. p. 217)

Here, the paternalist quality of this intervention is instrumental to the real goal of preventing harm to others. As I'll discuss in Section 5, in some cases self-benefit is a non-separable constituent of a collective good and therefore instrumental paternalism is necessary for its pursuit. Dworkin disagrees with Shiffrin that this exemplifies paternalism at all, because the motive and justifying reason is to prevent harm to another person. However, Shiffrin argues for a definition of paternalism that deliberately relaxes

both the self-benefit condition and the autonomy condition, in part because it helps to capture cases where omissions to act, and non-coercive actions, can be paternalist. For Shiffrin, what is crucial for identifying paternalism is the substitution of judgement within the sphere of the subject's sphere of legitimate agency on grounds the subject would not agree with, whether or not this specifically targets some conception of their wellbeing.

Here I've described the considerations that figure into the diagnosis of policies and interventions as paternalist. Dworkin's view is most responsive to the legal motivations, where Shiffrin provides a broader scope that addresses the moral concern. As I will argue in Section 4, worries about substitution of judgement loom large in the literature on epistemic paternalism, often with the consequence of casting a wide net as to what should count as paternalism at all. The information environment is rich with content and affordances that aim to change our minds and manipulate the ways we access and assess information in ways that seem to disintermediate our own judgement, but we should be cautious of conceptions of epistemic paternalism that would bring all these into scope. I take Shiffrin's argument to show that some account of a boundary between influence and substitution of judgement is required to make sense of epistemic paternalism. However, there are many cases where we might meet all the elements of Shiffrin's definition of paternalism on route to producing regulation that straightforwardly responds to harm to others. The self-benefit condition is still important to help identify why concerns about paternalism arise, and when they are analytically useful.

2.2 Libertarian paternalism - influence and persuasion as paternalism

Libertarian paternalism has shaped discussion of epistemic paternalism in three important ways. Firstly, it identifies cognitive and epistemic factors that can both enable and justify paternalist intervention. Secondly, it broadens the scope of the kinds of actions that can implement paternalist goals, adding subtle cognitive and psychological interventions to the more direct restrictions and prohibitions of traditional paternalism. Thirdly, it broadens the scope of what kinds of policies should count as paternalist, including both pro-self and pro-social goals in its conception of paternalism.

If we are worried about the potential for paternalist policies to violate our autonomy, those that change how we act by changing what we believe without consulting us about it might seem to be the most insidious. Even worse if the goal is also epistemic - to change what we think in some way we don't assent to or even perceive. Mill's argument against paternalism hedges on both the conditions under which we can be said to be exercising rational autonomy, and the definition of self-regarding actions and harms. Some harms to self are also harms to others, and sometimes we act with incomplete information that, were we to possess it, would change our minds. For non-absolutists, the determination of the bounds of acceptable paternalism involves an epistemic judgment, that the agent understands what they are doing, and what the consequences are likely to be, as we see in Mill's BRIDGE passage in Chapter 5 of On Liberty:

"If either a public officer or anyone else saw a person attempting to cross a bridge which had been ascertained to be unsafe, and there were no time to warn him of his danger, they might seize him and turn him back, without any real infringement of his liberty; for liberty consists in doing what one desires, and he does not desire to fall into the river. Nevertheless, when there is not a certainty, but only a danger of mischief, no one but the person himself can judge of the sufficiency of the motive which may prompt him to incur the risk: in this case, therefore, (unless he is a child, or delirious, or in some state of excitement or absorption incompatible with the full use of the reflecting faculty) he ought, I conceive, to be only warned of the danger; not forcibly prevented from exposing himself to it." (Mill, 1998, p. 96)

If we have reason to think that the person was unaware of the danger, then, on Mill's account, interference is justified. Perhaps we know they were looking in the wrong direction, or don't understand the language. These sorts of considerations complicate determinations of autonomy violation. Maybe if one possessed all the relevant information different choices would be made, whatever one's preferences and reasons for wanting to cross the river. Or perhaps one has a strong preference to cross the river, even in a dangerous manner - an example of a preference that is stable in the face of new information.

But what about interference that changes one's preferences? Of particular interest to this project is the "nudge", where paternalist goals are pursued by crafting subtle interventions that don't firmly foreclose options, but instead alter the way we perceive them. Over the course of several influential books and papers, Sunstein and Thaler have developed and defended the idea that these kinds of interventions can be effective and autonomy-preserving (Thaler & Sunstein, 2009). Paradigmatic examples of nudges make it easier to make particular choices, often by exploiting cognitive biases or manipulating defaults. We are still free to choose otherwise, but it becomes markedly less likely that we will bother to. These cases are interesting because they often rely on epistemic means, and sometimes target epistemic ends, and can thus often be considered forms of epistemic paternalism.

It is important to observe the way in which both the functional efficacy of the nudge and the justification for its deployment depend on the same underlying basis. I will refer to this as the 'cognitive argument'. Arguments for libertarian paternalism claim that we must consider cognitive bias (Thaler & Sunstein, 2009; Tversky & Kahneman, 1973) when crafting public policy. Thaler and Sunstein argue that policymakers can and should exploit these biases by manipulating choice architecture to improve public decision-making. Nudges can take the form of changes to the interfaces through which we make choices, like opt-in to savings plans, modifications to the layout of forms, or structured procedures required to decline a vaccination, that have strong and systematic effects on outcomes. Sometimes these are manipulated wittingly, sometimes they have evolved over

time to optimize for whatever selective pressures exist in context. The cognitive argument is based on empirical accounts of the psychological properties that explains the efficacy of such nudges, as they rely on predictable cognitive habits, heuristics, and failure modes for their reproducible effects.

A common theme in the critical response to this proposal (Mills, 2018; Mitchell, 2004; Sunstein, 2018; Thaler & Sunstein, 2003) is the 'oxymoron argument', that to whatever extent choice architecture manipulation can be shown to be an effective intervention, by the same token it must be regarded as a threat to autonomy. Sunstein and Thaler respond to this and related concerns by arguing that both choice architecture and cognitive bias is ubiquitous, and that consequently, paternalist design decisions are inevitable (Sunstein, 2005). Whether by deliberate design or selection pressures in situ, many of our choices will be made in contexts that are strongly biased and coercive. Arguments against libertarian paternalism based on strong objections to epistemic interference are therefore allegedly undermined, as there is no non-paternalist position available. Of course, there remains significant scope to debate the limits to acceptable interference, and to try to articulate a notion of autonomy that is responsive to the cognitive argument, and indeed a substantial literature has developed exploring this problem. To my ears, this conflates influence with paternalist coercion.

Some nudges involve preference elicitation, where choice architecture is deliberately engineered to force one to engage reflective cognitive processes in specific decision

contexts. Where one might mindlessly scroll through a social media news feed, tapping a button to share an article based on the headline alone without much thought, such an intervention would insert a step that asks if we read the article. Such prompts have empirically demonstrated effectiveness at reducing the rate at which content is shared – in many cases people will close the prompt and abandon the preference to share the article (Epstein et al., 2023). Other forms of preference elicitation involve the manipulation of defaults and the structuring of opt-in and opt-out decisions.

Consider an example I'll label BRIDGE-NUDGE. In Mill's bridge case, suppose one is interfered with - the police prevent the person from using the damaged bridge (that they didn't know was damaged), but the person argues that they still need to cross, and accept the dangers of swimming. Mill's response is clear – they should be permitted. The libertarian paternalist might want to be certain that the person really is expressing a rational preference – maybe they are angry or distracted, maybe they haven't fully considered the consequences. They are required to view a short public safety advisory video describing the dangers of the river, and it concludes with reminders about one's friends and family.

After this, they are permitted to swim – except they've changed their mind. Recalling the discussion of Shiffrin's argument in Section 1, to what extent did mandatory exposure to persuasion amount to a paternalist substitution of judgement? At stake in the examples I'll consider in Section 3 and 4 is whether manipulation of the kinds of information we

encounter, and the ways we encounter it, can amount to objectionable paternalism. In BRIDGE-NUDGE, one remains free to exercise one's autonomously chosen goal, but one is compelled to interact with a process that has been expertly designed to change one's mind. There is an asymmetry in this relation that is particularly vivid in many computer-mediated choice architectures where teams of experts armed with huge datasets and AI-driven predictive tools are on one side of the lever, and the epistemic agent on the other. The odds are stacked that preferences will change.

Empirical results suggest that all methods of elicitation will have significant effects on the preferences expressed, such as (Bettman et al., 1998), and that all choice architectures have biasing properties, often reflecting the interests of those that construct them. The libertarian-paternalist argues – if this counts as paternalism, an objectionable impingement on our agential autonomy – then this condition is universal and inevitable. This inevitability should therefore cause us to question our evaluation that it is objectionable – it is not feasible to navigate the world without interfacing with choice architectures that will change our preferences and behaviour. Our ethical evaluation should shift from the futile concern to preserve an illusory conception of autonomy, and instead focus on evaluating the architectures themselves – sure they change our minds, but do they do these in ways that are harmful, or that could be helpful, and should we collectively take an interest in consciously representing beneficial goals in the engineering of them?

The distinctive character of the nudge in this literature is the generation of hidden influence that can appear to involve substitution of judgement. Sunstein offers a brief but unsatisfying rebuttal to the idea that nudges work covertly by observing that their mechanisms are by-definition visible, for instance in a GPS mapping app. But a system that raises the salience of some information in turn lowers that of others, and hides even more, a kind of technological mediation of perception (Verbeek, 2015) that is unavoidable. Even when we can see the "what" we cannot see the "why". The extent to which nudges are present, their mechanisms of action, and the interests they serve, are not plainly visible.

Suppose the oxymoron objection is correct - there's a serious tension between the libertarian aim to respect autonomy and the paternalist aim to modify behaviour. Unless there is a problem with the inevitability argument, the challenge would seem to fall to the libertarian, not the paternalist. If the cognitive argument is correct, including the claim that all choice takes place within architectures that are manipulative in some way, then the libertarian owes us an explanation of how we should understand the kind of autonomy that must be safeguarded. There may well be serious challenges to meet in justifying nudging as a policy tool in the face of objections that its use is inconsistent with other values. But if we are always making decisions in contexts where our cognitive biases play a decisive role then there is a challenge for the libertarian no matter what position we take on the use of paternalist nudging.

However, ubiquity cuts both ways, and the effectiveness of a nudge, as opposed to more overt forms of coercion, depends on unstable properties of the information environment in which the intervention is enacted. Mills (Mills, 2018) distinguishes dual-system nudges, which leverage known patterns of behaviour and cognition, from acute nudges, which actively attempt to manipulate our reflective reasoning, or prevent us from engaging that system. But it is quite accidental that a particular decision may be manipulated via a dual-system nudge, instead of requiring an acute nudge, and it reflects a contingent relation between policy objectives and the extant type, strength, and valence of friction costs.

This generalizes to an instability problem that all nudges face. Any policy goal that can be accomplished with paradigmatic examples of nudges can have the viability of those nudges destabilized using techniques that take advantage of the same cognitive properties that the nudges rely on. Some persuasive technologies can increase vaccination rates, others can generate opposition, causing dual-system nudges to lose their effectiveness, and thus demanding more coercive interventions if there is still political will to pursue the objective. Where once there were no preferences, strong preferences can take root, and these may have been formed in ways that take advantage of biases and cognitive limitations that we might assess to be irrational, even involuntary. It is accidental that nudging can serve some particular policy goal, and this very contingency is borne out by the nudge's vulnerability to countermeasures that raise awareness of the

nudge itself, that try to inculcate opposing preferences, and that draw attention to the hidden elements of choice architecture.

As a category, what is most distinctive about nudges is that they work by influence, not force, and this invites us to consider the category of epistemic paternalism - whether in terms of the use of epistemic instruments, or the pursuit of epistemic outcomes. The combined effect of the cognitive and inevitability arguments erode a presumptive restriction on epistemic paternalism. The cognitive argument claims we often are not able to rationally assess our interests and make voluntary and rational decisions. The inevitability argument suggests we always face some degree of choice architecture manipulation. The instability problem is that these only work because of enabling epistemic conditions that can themselves become the target of intentional or accidental manipulation which the paternalist might need to pre-emptively guard against.

These then are the first two reasons why libertarian paternalism has had significant impact on discussion of epistemic paternalism. It highlights the importance of cognitive and epistemic factors in both enacting and justifying paternalist interventions, and as a result broadens the scope of potentially paternalist action to include influence and persuasion, in addition to outright restrictions and prohibitions.

2.3 From libertarian paternalism to epistemic paternalism

The third significant impact of libertarian paternalism on discussion of epistemic paternalism is the way it broadens the scope of what should count as paternalist by tending to ignore the self-benefit condition. Nudges usually exemplify mixed paternalism, and often focus more on benefits to others than benefits to self. Congiu & Moscati (Congiu & Moscati, 2021) observe in a systematic review that nudges often involve "pro-social" rather than "pro-self" aims. They concede that nudges can have either aim and that the category is defined by distinctly cognitive forms of manipulation. Indeed, the literature on libertarian and epistemic paternalism tends to treat all interventions that influence behaviour via cognitive interventions as kinds of paternalism.

A common example of nudging involves menus - they can be designed and changed in ways that will affect our choices. Schools might want to encourage healthier choices, a restaurant to maximize profits. The inevitability argument treats these cases as equal, but only one aims to benefit the subject of the interference, which is the third condition for paternalism, as Dworkin puts it, where the justifying reason refers "…exclusively to the welfare, good, happiness, needs, interests or values of the person being coerced" (Dworkin, 1972, p. 65). The line between influence and paternalism blurs, which we also find in the discussions of epistemic paternalism which uses the concept in ways that brings into scope a large swath of activity. But this is a problem, because by substantially inflating the category of interventions that can be paternalist to include all manner of subtle forms of influence, in addition to more direct ones such as the police on the bridge, we risk losing our grip on the category totally, and with it the moral and legal motivation

to mark out some zone of sovereignty where intrusions for our own good require elevated justification.

A lot of epistemic engineering affords some degree of substitution of judgement, which will benefit or harm the individual in a mix of direct and indirect ways. However, if the category of paternalism is to remain responsive to the moral and legal motivations for limiting paternalism, it must be in part with reference to the aim of this substitution - that it be our own good, not the prevention of harm to others. We accept that some kinds of information will be controlled because it's publication would cause harm to others, but much less readily do we concede this if the harm is to ourselves. In Chapters 3 and 4 I'll give reasons we cannot so easily separate these.

We can see an early acknowledgement of the possibility of the kinds of cognitive and epistemic interventions favoured by libertarian paternalism in Buchanan (1978), who argues that, "Granted the complexity of the relations between information and action, it seems plausible to expand the usual characterization of paternalism as follows: Paternalism is interference with a person's freedom of action or freedom of information, where the alleged justification of interfering or misinforming is that it is for the good of the person who is interfered with or misinformed." (Buchanan, 1978)

Buchanan has in mind situations such as when a physician withholds information from a patient with the belief that the patient would otherwise make poor decisions. Similarly,

libertarian paternalist nudges often work by changing the salience of information causing us to be more or less likely to consciously consider it. In this sense, Mill's BRIDGE passage, and my BRIDGE-NUDGE example, where the libertarian paternalist establishes a choice architecture that results in the subject preferring not to attempt to cross the river, are examples of what we can think of as instrumental epistemic paternalism.

However, just because the means of intervention are epistemic does not necessarily mean that we have reason to think that epistemic paternalism forms a special category. This would result in a large range of paternalist interventions counting as epistemic. Consider the definition of epistemic paternalism proposed by (Ahlstrom-Vij, 2013) that an act is epistemically paternalist when it meets all three of these conditions:

1. Interference: practice interferes with a subject's ability to access, collect, and evaluate information in whatever way they see fit.

2. Non-consultation: the interference is undertaken without consulting the subject

3. Improvement: one of the motivations for the interference is that it will improve the subject's epistemic standing.

Here the epistemic character must be found in the nature of the benefit, not only the mode of intervention. Ahlstrom-Vij argues that the benefit must not be merely incidental,

but that it be necessary for the instrument to function successfully. In BRIDGE-NUDGE the subject's appreciation of their risky endeavour is essential to achieving the paternalist aim. If instead the intervention was crafted in some way as to merely take advantage of some quirk of the subject's psychology, then any epistemic improvement would incidental. Ahlstrom-Vij (p. 58) uses an example I'll label SPOOK that also involves what we should think of as incidental epistemic improvement. If a government controls the distribution of secret information pertaining to national security, but where it happens to be the case that the information is complex, difficult to assess, and highly likely to cause false beliefs, then the maintenance of this security delivers an epistemic benefit. It also prevents the formation of false beliefs and production of misleading evidence. However, this benefit is incidental to the aims of the policy and ought not be considered an example of epistemic paternalism.

Consider this example from (Meehan, 2020) I'll label FLATMATE which is epistemic both in mechanism and aim. One has a flatmate who has started to consume a highly polarized and limited information diet. In response:" ... you decide to nudge him away from forming anymore irresponsible beliefs from untrustworthy news sources. Some of the measures you take include offering him a discount for the subscription service for a well-trusted newspaper, warning him about the reliability and trustworthiness of the sources he reads his news from, and leaving neutral, unbiased news programs on the TV" (Meehan 2020, p. 252).

In this example one takes control of the subject's information environment without their consent, and without necessarily targeting any specific doxastic outcome, but rather hoping to generally improve the veritistic quality of their beliefs. A question I will return to in the final section of this chapter is that of whom we should say benefits from this improvement, but for now let's suppose the aim is just to benefit the subject. Grundman (2021) calls this kind of effort "doxastic nudging" and argues that it is in effective and feasible intervention that "… can make people believe certain propositions by rendering those propositions particularly salient or framing them in especially persuasive ways." (Grundmann, 2023) Grundman argues that this is an example of epistemic libertarian paternalism. We can see the connection to libertarian paternalism in the way that the cognitive argument identifies both the methods and justifications for a class of intervention that aims to change how we behave, and what we want, in ways we don't directly perceive and endorse.

Another kind of epistemic nudge that has been considered an instance of epistemic paternalism is the ACCURACY NUDGE, which is of special interest because of the indirect way it provides a self-benefit. Accuracy nudges (Pennycook, McPhetres, et al., 2020; Roozenbeek et al., 2021) attempt to cause users of digital information systems to think about accuracy as a value while engaging with user interfaces that afford the viewing and sharing of content. Where one might use such systems without thinking about accuracy, tapping a button to share an article, these prompts appear and require interaction before the action is completed, exposing the user to a message that causes

them to consider the accuracy of the article, inducing reflective cognition in the user. They are typically deployed as a part of efforts to combat misinformation in digital information environments, but the mechanism of action is aimed at the reduction the distribution of misinformation, not to interfere in the doxastic context in which the subject views the content. The goal is to prevent harm to others, both epistemically, by protecting them from misleading content, but also non-epistemically, given that such policies are typically enacted to mitigate specific threats to our wellbeing, such as misinformation about public health measures.

Dworkin makes an important distinction that has been neglected in much of the literature on epistemic paternalism, one that applies to policies such as prohibitions on misleading advertising, or rules that prevent people from entering into harmful contracts. I'll call this the self-binding argument. “There are restrictions which are in the interests of a class of persons taken collectively but are such that the immediate interest of each individual is furthered by his violating the rule when others adhere to it. In such cases the individuals involved may need the use of compulsion to give effect to their collective judgment of their own interest by guaranteeing each individual compliance by the others.” (Dworkin, 1972, p. 69)

When we agree that false advertising should be prohibited, or that there should be safety standards for consumer goods, it is not paternalist to prevent us from violating these rules when we want to. The self-benefit condition would appear to be satisfied, but

examples with this self-binding structure are distinct from paternalism because the good that is pursued is accepted and agreed upon by those bound. These preferences are projected and protected by this kind of autonomy-limiting self-regulation, and thus the non-consent condition of paternalism is not met.

We consent to self-binding when we participate in the regulated social practice of advertising. ACCURACY-NUDGE can take on this structure, but we can imagine disagreement about what counts as accuracy, when it matters, and its relation to other adjacent aims. The use of white propaganda during times of war, where state media efforts attempt to improve morale and shape public opinion in strategically favourable ways, can be a kind of self-binding where we agree to be manipulated to improve our capacity to defend ourselves. These are cases that might otherwise look like a kind of epistemic paternalism - where we are manipulated into changing our beliefs without our consent and even knowledge at the time.

Paternalist benefit-to-subject doesn't count if the subject does not agree they have benefitted, for example, by receiving an epistemic improvement they don't care about. Ahlstrom-Vij argues that in some cases it's sufficient that we might have an "informed" interest, rather than an active one, in the epistemic improvement we receive. We have an informed interest when, were we to possess the relevant knowledge, we would come to possess an interest in the epistemic improvement, and thus endorse it, satisfying the self-benefit condition. I might not believe I benefit from learning about where my coffee is

made, but if I come to learn that it funds a political party I am opposed to, I now have an interest in this knowledge, and can agree that I benefit from acquiring it.

In many cases, when we accept management and stewardship, we engage just this kind of self-binding, where we accept that our will might be frustrated by the future decisions of the stewards, but, so long as those decisions are in line with the goals that motivated vesting power in them, we should say no violation of our autonomy took place. Newspapers might decline my op-eds, political advertising I might wish to undertake is regulated, new anchors might lag behind what I believe to be the state of the facts, regulators might try to persuade me to eat better, forum moderators might delete posts that are crass or unproductive, and so forth. That discussion of epistemic paternalism often carves out exceptions for education and interactions with children admits in small measure what I think we should admit more broadly - that there is a great deal of gatekeeping and shaping of information in our everyday life that we assent to in principle, even if we dislike it at times when we face its implementation. It would be incorrect, and analytically unhelpful, for these to fall into the category of paternalism, which is supposed to come with special obligations and justificatory requirements. Instead, we should see these as forms of stewardship, re-focusing critical scrutiny on its efficacy, fairness, and alignment with the interests that motivate it.

This section argued that libertarian paternalism operates with a broad conception of paternalism that elides the self-benefit condition. It introduced epistemic paternalism,

situated it within the context of the cognitive argument of libertarian paternalism, and noted the extent to it is vulnerable to the same tendency to overlook the self-benefit condition. Given that there's nothing distinctive in the method of intervention alone that can pick out paternalist from non-paternalist policies, conditions on the kind of policy goal, as identified in Section 1, are required if paternalism is to be the basis of a substantial objection or basis for justification.

2.4. Defining epistemic paternalism

The previous section argues that we should reject definitions of epistemic paternalism that are overly broad along the paternalist dimension, and that the self-benefit condition should be retained. In this section, I advocate constraints along the epistemic dimension, specifically, that the benefit pursued should not just be instrumental to some non-epistemic one. Examples in the epistemic paternalism literature can be divided into two types, the agential and the social. The agential ones tend not to be interestingly epistemic because they are usually targeted at non-epistemic improvements. I'll argue that the social ones aren’t obviously paternalist since they tend to involve motivations that focus on harms to others. The underlying concern is to cut the space of genuine epistemic paternalism down to size - that it should be properly epistemic and properly paternalist to offer a distinctive objection to some policy. What remains is important for this project, and the topic of Section 5.

Agential examples are common in bioethics contexts, where there are frequent asymmetries of power and knowledge. Bullock considers cases where a physician disregards a patient's request to be shielded from information (Bullock, 2018a) Ahlstrom-Vij argues that an AI system that guides clinicians can be epistemically paternalist, circumventing judgement in pursuit of better outcomes. Bandini (Bandini et al., 2020) imagines a cancer patient who resists potentially lifesaving surgery, and a physician who resorts to manipulation and persuasion to change their view and obtain consent for the procedure (ibid. p. 125). That they resort to epistemic means is only to further a non-epistemic end, and only because consent is required by professional and legal norms that foreground autonomy in direct response to concerns about paternalism. Misak (Misak, 2004) observes a disjuncture, that patient wishes must routinely be ignored in the provision of care, and yet this is in serious tension with the avowed anti-paternalist norms of clinical practice, and wonders if there is something implausible about strong anti-paternalism in medical contexts. Physicians must keep a "double set of books" (ibid. p. 421), one tracking the needs of the person, the other that of the medical problem they must solve. The need to resort to epistemic means is some cases is almost salutatory, and concern is focused on shaping the patient's decisions, not their reasons or beliefs.

Many examples of epistemic paternalism of the social type offer justification aimed at the protection of our ability to participate in coordinated epistemic endeavors (Ahlstrom-Vij, 2013; Croce, 2018; Hausman, 2018; McKenna, 2020; Pritchard, 2013). Our individual interests depend on things like juries and clinical trials functioning correctly,

but it's somewhat tenuous to construe these interests as distinctively self-regarding, or epistemic. Ameliorative efforts aimed at directly improving the social function do not necessarily directly improve the epistemic standing of the subject of interference. In Ahlstrom-Vij's examples we find the instrumental adoption of epistemic means, but to protect an epistemic function that is public, and which we benefit from only indirectly. In JURY he argues that the control of evidence to jurors is epistemically paternalist on substitution-of-judgment grounds. Likewise, in CLINICAL TRIAL, he argues that scientific norms such as removing decisions about patient allocation from clinical researchers in order to ensure random clinical trials are truly random, are epistemically paternalist. They protect the scientists from errors they might make, probably unwittingly, that would endanger their work, in which they have an interest, against their will. Many of this kind of example will also turn out to be instrumentally epistemic as well.

Where can we look for examples that are properly epistemic, and properly paternalist? Goldman (Goldman, 1991) argues that "... communication controllers are exercising epistemic paternalism whenever they interpose their own judgment rather than allow the audience to exercise theirs" (ibid. p. 119). Goldman takes on board a standard definition of paternalism like that in Section 1 and extends it to realm of what he calls social epistemics, "... social practices, or institutional rules that directly or indirectly govern communication and doxastic decision." (ibid p. 120). Examples of epistemic paternalist practices such as restrictions on evidence presented to juries, control of curriculum, and

policies banning deceptive advertising have in common what Goldman calls a veritistic justification, rooted in the value of promoting practices that increase the likelihood of producing knowledge. Where is the pro-self motive here?

Alluding to Hardwig's (Hardwig, 1985) work on epistemic dependence, Goldman notes that it is generally unfeasible to make decisions using only the fruits of our own epistemic labour, and thus our epistemic autonomy is more permeable than our decision-making autonomy. Controlling and shaping the flow of information can even enhance autonomy, by correcting for predictable cognitive biases (a motive also endorsed by Ahlstrom-Vij). Skillful control of the ways in which we form beliefs can improve our ability to act on our own preferences and improve our epistemic standing.

However, just because our epistemic standing is improved does not mean we can be said to have received a benefit. Bullock (Bullock, 2018a) argues that epistemic improvements are not valuable in themselves, and that they only benefit the subject alongside non-epistemic benefits. Knowing more about nutrition helps us maintain our health, knowing about traffic conditions help us get to work on time, and so forth. A dilemma arises as a result: either epistemic paternalism is parasitic on non-epistemic paternalism, or epistemic paternalism is unjustified. This is because if epistemic ends are valuable in their own right, then their diminishment, as many forms of epistemic paternalism require, cannot be justifiable. In cases such as evidence control in jury trials, filtering categories of misinformation online, persuasive structuring of the presentation of

treatment options in medical settings or teaching false theories as part of progressive teaching lesson, there is deliberate diminishment of the subject's epistemic status. In a later article, Goldman also argues that epistemic improvements are not valuable in themselves. "If practices generate true beliefs in which no relevant agent has an interest, they do not get veritistic credit." (Goldman, 2000, p. 321) Consider this example, hereafter PHYSICS, from Bullock. "Suppose, for example, that I play a series of physics lectures to you whilst you are sleeping, with the intention that you subconsciously learn quantum mechanics. I have good reason to think this will be effective. You happen to have no interest in quantum mechanics and the facts that you learn have no bearing on your wellbeing" (Bullock, 2018a, p. 442).

Bullock raises a challenge to the epistemic paternalist that I'll return to in the final section, that if epistemic paternalism forms a coherent and distinct moral project, there must be some cases where "... epistemic value can counterbalance considerations of wellbeing when the two come into conflict" (ibid. p. 444). For the moment I'm concerned just with the interest condition. Perhaps we have purely epistemic interests that we are not aware of, or do not recognize. As I describe at length in Chapter 3, our epistemic dependence on others, especially via testimony, makes us vulnerable, as Fricker argues, "... it extends one's knowledge base so enormously, [but] lessens one's ability rationally to police one's belief system for falsity" (Fricker, 2006b, p. 242). Some philosophers have entertained the idea the mitigation of this kind of risk licenses epistemic paternalism.

Consider what McKenna (McKenna, 2020) calls "consequential false beliefs", those with important bearing on collective wellbeing. He gives the example of CLIMATE CHANGE, where widespread public misperception about climate science in a democratic state undermines cooperative efforts to respond to the threat. This is just the sort of case that has motivated cognitive security interest in protecting information environments.

McKenna argues that the correction of these is a justifiable policy objective and considers remediation strategies that might be more effective when rational persuasion fails. First "prebunking" or what has been called "inoculation theory"(Lewandowsky & Van Der Linden, 2021) - the deliberate interjection of carefully crafted content that will cause people to be less receptive to misinformation. Secondly, public relations strategies such as persuasive framing, and thirdly, choosing persuasive spokespersons, especially to signal in-group membership to recalcitrant demographics. He argues that these stand or fall together as epistemically paternalist, because the latter three are not substantively different from rational persuasion, and do not detract from our rationality and autonomy. They even enhance it in a "...complicated socio-epistemic environment..." (McKenna, 2020, p. 101) where it is hard for us to succeed epistemically unaided. I'm not concerned here with whether they are justifiable, and Mckenna dismisses out of hand the idea that they are not paternalist, a matter I'll return to in Section 5. For now, I'm interested in the extent to which they could be distinctively epistemically paternalist at all. Following Bullock's strategy, not only might we agree that our interest in climate change belief is

non-epistemic (maybe because we care about climate change itself, or feedback into democratic and scientific institutions), but we might also agree that amelioration that diminishes epistemic standing is consistent with the pursuit of these ends, in the same way we might agree that some forms of non-epistemic paternalism that trades epistemic goods for wellbeing can be justifiable. CLIMATE CHANGE fails the non-epistemic-diminishment test of pure epistemic paternalism - all the aims of the policy can be achieved without improving the epistemic standing of those interfered with.

Worries about consequential false beliefs also motivate Ahlstrom-Vij's effort to define and defend epistemic paternalism. He argues that the "...available psychological evidence regarding our dual tendency for bias and overconfidence not only gives us reason to worry about the former, but also suggests that we cannot rely on ourselves for epistemic improvement" (Ahlstrom-Vij, 2013, p. 36). This is an internal threat to our epistemic autonomy, whereas Fricker argues there is an external threat, that "... [t]he human would-be epistemic autonome on closer investigation is not an ideal, but either paranoid or severely cognitively lacking, or deeply rationally incoherent" (Fricker, 2006b, p. 244) . Ahlstrom-Vij is worried about the limitations of our individual epistemic efforts and looks to epistemic paternalism to help correct them. Fricker, on the other hand, cautions us that even at our best, epistemic self-reliance is an implausible and incoherent ideal. Epistemic dependence might justify autonomy violations in the same way that our dependence on the sobriety and attentiveness of fellow motorists justifies interference with their autonomy, and the question I'll turn to in the final section of this

chapter is whether we should think that there's something distinctively paternalist about attempts to avoid harms that epistemic dependences makes us vulnerable to.

For the individual juror, the non-epistemic-diminishment test is failed, for the criminal procedure as a whole, it is passed. Many epistemic domains exhibit similar structures - it is as though in BRIDGE, all involved are connected by a rope to the subject - if they attempt to cross, everyone ends up in the river. This kind of consideration can reveal cases where self-benefit and social-benefit are so interdependent that we must acknowledge a pro-social aim still risks paternalism. But then we face the other horn of the dilemma - if the goal is avoidance of non-epistemic harms, epistemic paternalism collapses into general paternalism. Maybe pure cases will always have the structure of PHYSICS, and that when we integrate interest considerations as both Bullock and Goldman recommend, we end up with cases like CLIMATE-CHANGE, where the interests that matter appear to be significantly non-epistemic, and potentially not even paternalist. We will have failed to identify a distinctive category of justifiable epistemic paternalism.

However, if we consider the discussion of the cognitive argument and instability problem from Section 3 in the context of the justification for epistemic interference based on our roles in social epistemic practices, we find that there is a form of epistemic paternalism that might be justified on purely epistemic grounds. Our epistemic interdependence generates interests in collective epistemic endeavors. One way to think

about the protection of a purely epistemic good is that in many cases it will be content-neutral. We don't just care that subject acquires a true belief; we are concerned that they have acquired it in an appropriate way with reliable justification (I discuss the social epistemology of justification in the next chapter). An intervention that inculcates a vulnerability to certain classes, or even just higher volumes, of malformed beliefs is an epistemic diminishment because it undermines the subject's standing as an epistemic agent, for themselves and for others.

The intervener in FLATMATE is concerned with the possibility that the subject becomes disconnected from their community's construction of shared reality in a general way, that cannot be said to only create risks for the epistemic subject. It also creates risks for those that depend on the subject, which relate to a multiple of roles the subject might occupy. They are undermined as part of an information system, quite apart from risks posed to their individual social and economic interests. In a reading of FLATMATE where the intervention is justified by this kind of concern, the subject's information environment is interfered with for the sake of making it more diverse and more reliable. Unlike CLIMATE CHANGE, the interference is not aimed at the generation of some specific outcome where it may be that epistemically detrimental means work just as well. This sort of case can potentially meet Bullock's challenge, but it would still have to be shown that the intervention is strong enough that it generates a substitution of judgement if we should think that it is paternalist.

In section 1 I discussed Shiffrin's argument that the diagnostic criterion for paternalism should be loosened from autonomy-violation to the substitution of judgement. In Section 3 I observe that this accords with the broad way the term paternalism is used in the libertarian paternalist literature, and the way it has come to be used in bioethics, where the principle of autonomy has been interpreted to include the autonomy of judgement. Le Grand and New argue that interference is paternalistic when it corrects a person's false beliefs or failures of reasoning, for their own good, and that this kind of paternalism can be justifiable. But Hausman observes this is an implausible constraint on the identification of paternalism– there are many cases where we might think interference would be paternalist even where the subject of interferences was not acting from flawed reasoning. Hausman argues that a policy is paternalist "...if and only if it aims to take over or control what is properly within the agent's own legitimate domain of judgment or action for the benefit of the agent" (Hausman, 2018, p. 65), which is a nice formulation of Shiffrin's substitution of judgement condition that retains the focus on the agent's benefit.

This project is concerned with a set of problems that face an increasingly complex assemblage of state, state-like, and non-state actors, who are able to act in the information environment in ways that can have strong effects on the epistemic standings of others. FLATMATE is an example of epistemic interference that meets the interest condition, passes the non-epistemic-diminishment test, and could be enacted in a way that meets the substitution of judgement condition, and is not instrumental. I think it's a good candidate

example of epistemic paternalism within the parameters that have arisen in this literature. However, it defangs somewhat the force of paternalism as a reason to demand stronger justification on the grounds of autonomy violation. If the self-benefit condition is met only because self and social benefit are constitutively interdependent in the domain in question, we should expect to find relevant conceptions of autonomy are similar mixed. If the subject in FLATMATE has beliefs and behaviours strongly constructed by the information environment, their actions in this domain are similar constrained and shaped. Rather than engaging in a complex form of mixed pro-social and pro-self-paternalism, the intervener, so long as they aren't actively censoring, is trying to improve a shared state of affairs, a kind of stewardship.

FLATMATE exemplifies a kind of intervention that is increasingly feasible and accessible, given the powerful tools we possess in digital information environments to control information flows and harness cognitive biases to modify behavior without engaging reflective cognition, and which libertarian paternalist interventions can wield for the benefit of individuals and their communities. However, these risk various forms of collateral epistemic diminishment in the pursuit of specific ends. In FLATMATE and ACCURACY-NUDGE, the goal is to improve the epistemic properties of the environment in which the individual plays multiple epistemic roles.

It seems that a hybrid epistemic paternalism is possible - where we benefit both in the performance of social epistemic roles, and also from the performance of those roles in

which we have an individual interest. In some of these cases we have a purely epistemic interest in protecting the conditions necessary for the performance of these roles. If there is a category of interference with our epistemic agency that should properly be considered epistemically paternalist, it is this these cases, but I've noted that stewardship may well be a more perspicuous description, especially because most of this stewardship is exercised in ways that do not impose a strong substitution of judgement, and tend to cohere with goals we endorse.

2.5 Epistemic paternalism and epistemic engineering

In the preceding section I defend a narrow conception of epistemic paternalism. I argue that many putative examples should not be considered such - either epistemic improvement is instrumental and fails Bullock's challenge, or it is not paternalist, either because the pro-self element is missing, and the aim is to avoid harm to others, or because they exhibit Dworkin's self-binding structure. The goal of this section is to explore what remains of this space. I describe a type of epistemic engineering that could meet the self-benefit test, aims for a purely epistemic good, and answer Bullock's challenge.

The manipulation of choice architectures to cause us to become better consumers of information illustrated in ACCURACY NUDGE, might appear to be an example of strictly epistemic paternalism. These kinds of epistemic nudges do not aim to avoid any

specific or immediate non-epistemic harm, and improvement can be expected to reduce the likelihood of harms to both the subject of interference and the broader community. Disentangling pro-self from pro-social improvement poses special challenges when collective benefits cannot be enjoyed without individual benefits, and vice-versa. ACCURACY NUDGE interferes with the subject's behaviour, in a way that generates an indirect epistemic benefit, but also which indirectly protects the subject from manipulation, and even helps protect the subject's reputation. Some putative examples of paternalism, and especially of epistemic paternalism, involve goals that indirectly benefit the subject of interference. These are particularly relevant problems in epistemology because if our epistemic wellbeing is ineluctably bound up with that of others, then accounts of both epistemic autonomy and self-benefit are complicated by this constitutive interdependence.

Likewise, FLATMATE aims to protect a shared interest in the health and functioning of our epistemic environment. At times the roommates will depend on each other as sources of information, as will others. Improving the epistemic agency of the roommate doesn't just lower the chances of non-epistemic harm befalling them, it makes them more reliable participants in our collective information systems. We might even accept that for a while this epistemic improvement will have negative impact on the subject's non-epistemic wellbeing - it might cause them to be unhappy or incur costs. A case like this could meet Bullock's challenge, though one would want to show that the non-epistemic costs can be justified. This appears to be a candidate example of epistemic paternalism.

However, we should add one further condition, in response to what Ahlstrom-Vij calls the "epistemic outlier problem". In these cases well-intentioned epistemic paternalism subverts the epistemic agency of an individual who is actually-positioned to do better in the absence of constraints. No one is epistemically better off as a result of this interference, however infrequent the cases, and this threatens a conception of epistemic paternalism that demands epistemic benefit. In genuine instances of justifiable epistemic paternalism there should be good reason to believe the outlier problem is avoided, for example, because a general, content-neutral improvement is sought. FLATMATE thus avoids the outlier problem because it aims for generalized improvement in the subject's information environment, not to cause some specific epistemic outcome that in some cases the subject might be better positioned to enact. However, it does raise the bar for the intervener - if one reshapes the information environment to de-emphasize unreliable sources, one really ought to be sure they are unreliable! An important epistemic motive for protecting informational diversity and intellectual autonomy is to reflect our genuine uncertainty about the best routes to knowledge. On a stewardship conception of epistemic management, which I explore in detail in Chapter 4, uncertainty on this point is built-in and contestable, and the strength of interventions sensitive to the outcome and confidence of these deliberations. On the paternalist conception, it's more difficult to see how to build in justification for the extra strength and lack of accommodation for disagreement that an inescapable intervention would generate.

BRIDGE-NUDGE, JURY, CLINICAL TRIAL, and CLIMATE CHANGE are not examples of epistemic paternalism because they face the outlier problem, and because the harm they seek to avoid is non-epistemic. There is good reason to think that we value and have an interest in some baseline threshold of health and reliability in our information environment on the whole. We need search engines that don't systematically deceive us, sources of evidence that are not tampered with, methods to identify expert testimony, and access to peers who have not been incentivised or manipulated into misleading us. These have non-instrumental epistemic value when we consider them in aggregate, which would be appropriate if their functioning was threatened generally by some mediating process - which is a common condition in the digital information space. Chapter 5 describes how a search engine, or a social network, can generate these kinds of general, content-neutral epistemic environmental threats.

Consider an example I'll call SCIENCE COMMUNICATOR. Kahan (Kahan, 2017) argues that discourse about public health can become dysfunctional when the narrow and highly specific interests of domain experts are reflected into the public sphere, generating misleading evidence. Professionals examining a tiny subset of the information space fixate on dangerous and interesting phenomena, perhaps some pocket of misinformation or a novel conspiracy theory. They worry this might be a bellwether for a threat to the epistemic underpinnings of public health measures and discuss this threat in their professional circles. But this discussion itself is then reflected as a kind of evidence - that professionals are concerned about this information. This can snowball and have the effect

of amplifying misleading information - creating what Kahan calls a polluted science communication environment. Kahan argues we need norms for science communication that reflect the "...need for self-conscious management of the quality of the science communication environment" (ibid p. 421), and "evidence-informed, centrally managed communication by professional risk communicators" (ibid. p. 422). Like SPOOK, this information management would plausibly generate an epistemic good - the prevention of false beliefs about public health measures. And while in the domain of public health this has obvious non-epistemic instrumental value, the example generalizes to other domains where we have an interest but lack domain expertise. Another example of epistemic paternalism with this structure is the deliberate control and censure of information about environmental toxins with potential public health risks, to prevent panic, irrational risk aversion, and political misjudgment.

Recalling Sunstein's inevitability argument, we might think that given that there is no raw, default state of the science communication environment, there is no prima facie objection to giving more thoughtful shape to its contingent structure. This kind of deliberate control of communication satisfies Hausman's definitional conditions on paternalism in an interesting way, because evaluating information I have no training and competence to assess does not fall with in what Hausman calls my "legitimate domain of judgment or action", in any context where my judgement will have consequences that undermine my interests in the veritistic quality of the relevant information environment, or where my predictable mishandling of the evidence will cause harm. This is particularly

true in the many contexts in which our epistemic agency is not merely as passive receivers, but where we also transmit - generating evidence and providing testimony. It is necessary to intermediate institutions and experts between agents and complex sources of information, but this involves substitution of judgement we would normally welcome, and assent to, and which is rarely so totalizing as to outright prevent access to the raw information. However, we do not usually implement this sort of control in ways that are strong enough to prevent determined inquirers from accessing raw data and considering the matter directly.

Requiring a total absence of epistemic outliers, who would be better off without the intervention, helps Ahlstrom-Vij distinguish genuine epistemic paternalism from "epistemic utilitarianism"(Ahlstrom-Vij, 2013, p. 109), whose aims are directed at the common good, and is not bound to deliver individual benefit in all cases. This matters to a project like Ahlstrom-Vij's that aims to describe the conditions in which epistemic paternalism is justifiable - one wants to deal with examples of genuine paternalism. But my project has an orthogonal interest - the extent to which epistemic interventions and engineering aimed at public and individual benefits might count as paternalist and thus demand some special justification. I'm worried that classes of epistemic engineering that we can undertake in the service of cognitive security goals might be found objectionable on the grounds that they are paternalist. If some of these interventions count as epistemic utilitarianism, rather than epistemic paternalism, in part because they don't aim to benefit everyone, and in fact do generate outliers, this is yet another reason to think that

objections based on worries about paternalism won't stick. Being good enough is often going to be sufficient to justify attempts at epistemic utilitarianism (perhaps implemented in the form of stewardship and self-binding), whereas justifiable epistemic paternalism has to deliver benefit in every instance.

However, one might think that efforts that target the protection or improvement of group epistemic endeavours will benefit the subject who is interfered with in all cases, such as ACCURACY NUDGE and SCIENCE COMMUNICATOR, that given the practical impossibility of operating in an unfiltered and unmanaged information space we always benefit from the effort, even if imperfect. Meeting the no-outlier condition in this way might appear to leaves open space for charges of paternalism, and yet, this leads back to the strength of substitution. If implemented in a way that is strong enough to reach that threshold, we might then become dubious that an actual epistemic improvement has been delivered. As I see things, there's very little space left for interventions that are properly epistemically paternalist. Professionals in many fields have their actions, including communication, constrained by standards and regulations that we take to be justifiable in ways that are based on the harm principle or stewardship interests, and not as forms of justifiable paternalism. Could it be paternalist to deny the public direct access to the raw deliberations of scientists, or to enforce a communications strategy to prevent individual consumers being harmed by predictable misunderstandings? One way to argue that these cases are not paternalist is just on the grounds that the substitution of judgement is too weak to count as a violation of

autonomy. Where there is no monopoly on our access to epistemic resources (unlike SPOOK), the intervention is avoidable, and therefore not autonomy threatening.

To target purely epistemic values in a non-instrumental way one needs access to considerable epistemic leverage. In bioethics contexts physicians, caregivers, and surrogates have significant power that raises the risks that their actions might be objectionably paternalist. But in SCIENCE COMMUNICATOR, insufficient epistemic power is applied to generate autonomy-threatening substitution of judgement. In BRIDGE Mill is concerned that the bystander might not be able to understand the "sufficiency of motive" and argues that only the individual agent can make the final judgements about this, even where he seems to permit the bystander to argue a little and inform. Paternalism is supposed to generate justificatory burdens when there is no space for this exercise of individual judgement, which is why the strength threshold is important.

SPOOK involves state power, and we can imagine versions of FLATMATE involving overwhelming control of the subject's information diet. If we find that the centralization of AI models leads to monopoly, for example, where we can't practically avoid relying on a particular language model, the embedded human judgements made in training it for safety, accuracy, toxicity, and the like, might well meet the strength test, and we might easily imagine instances where the presumptive benefit is aimed at the user, and might think that consequently there is a high bar to justify the constrains this places on users

(and it might be expected to generate outliers and thus not be justifiable). The justificatory requirement might be met by introducing modes of transparency and civic engagement, and other management practices to ensure that we are not in disagreement with the constraints it places, or it might just be met by intervening to remove the conditions of ubiquity.

One might worry that the substitution-of-judgment condition rules out epistemic paternalism totally. If our judgement is disintermediated, we might cease to act as epistemic agents in any meaningful sense. But in Chapter 3 on epistemic dependence, I argue that the evolving field of social epistemology complicates individualist pictures of epistemic agency by identifying our dependence on testimony, expertise, interpretation, as well as instrumentation and technologies. To varying degrees regular deferral and substitution of judgement are the norm, not the exception, and often we have little choice in engaging dependencies. Echoing the inevitability argument, we might ask - if substitutions of judgement are everywhere, how can this mark the boundaries of the paternalistic? Where does augmentation end and substitution begin?

As I suggested in the AI model example, one way we can distinguish these is the extent to which the substitution itself cannot be substituted - when alternate delegations are not possible. The use of computer algorithms to control the publication and distribution of content in digital information systems has the appearance of epistemic paternalism in large part to due to strong non-consensual substitution of judgement.

Discussions of the epistemology of these systems typically emphasize they ways in which they risk epistemic (and non-epistemic) harms to users, for the benefit of platform. An algorithm might select what we see, but it does so from a volume of material we couldn't possibly sift through ourselves. There is no un-delegated alternative, and thus it's hard to specify a sense in which a choice we might want to make has been taken from us, and thus there is no epistemic-outlier who might have done better with the unfiltered data. But we might have done better with other ways of filtering the data, and we might make better decisions about evidence when we understand how the data is filtered, and the provision of such affordances, by weakening the substitution, can shift the standards of justification from paternalism to epistemic utilitarianism, stewardship and self-binding. This blocks intuitions that effective shaping of the information environment itself always at risk of epistemic paternalism, and instead points to other facts about the mediating sociotechnical systems.

The goal of this chapter was to survey the literature on epistemic paternalism and determine the extent to which it generates objections that could apply to the kinds of epistemic engineering and management that next chapters of this dissertation are concerned with. I've considered definitions of paternalism, the development of libertarian paternalism and the cognitive and epistemic conditions on agency, choice, and behaviour that this literature has brought to the fore. Then I examined the literature on epistemic paternalism and the challenge of identifying genuine cases of this. I have not been concerned with assessments of the extent to which particular instances of paternalism are

justified, just to identify the extent to which acts of epistemic management and engineering might be properly categorized as such.

If most of them were, this would have the implausible consequence that many of the actions we take that affect the epistemic efforts of others might require special justification. It would classify broad categories of stewardship, self-binding, and communicative action as paternalist. We would need to believe that our persuasive efforts are so powerful that we require certainty about their goals, for example, in CLIMATE-CHANGE, that we are correct that this is an especially consequential belief, and that we have identified the elements of it that are most assuredly true. This strikes me as implausible on both counts.

If none of them were, anti-paternalism would have little traction in the face of the subset of interventions that really do risk violation of our epistemic autonomy for our own benefit - which would disregard a large body of literature that expresses a consensus that there is something objectionable about this, even where there is disagreement about the exact boundaries of the category and the requirements for justification. The increasing centralization of the private sociotechnical systems that mediate many of our epistemic activities can risk genuine unjustifiable paternalism.

For the kinds of epistemic engineering this project is most interested in, we can expect that pro-self and pro-social aims and impacts will be entangled, because individual

epistemic improvements are often non-separable constituents of a collective good. Even when not the direct aim, self-benefit will be a necessary and foreseeable side-effect. This is why libertarian paternalism embraces mixed motivations, but why I think the concept of paternalism is generally a poor fit when assessing efforts to promote epistemic public goods and to prevent epistemic social harms. It imports an individualist frame into domains of constitutively social concern. Where it is supposed to act as a brake on state (and state-like) power over the individual for one's own good, it fails when one's own good is inexorably connected to harms to others.

While it is possible to imagine forms of epistemic engineering in the service of cognitive security interests that meet the narrow conditions of epistemic paternalism I have identified, I will give reasons in the coming chapters to the effect that these will almost all be objectionable for other reasons. For example, cognitive security defense might be enacted by way of strong censorship of disinformation at network levels that are hard, even impossible, to avoid. However, these sorts of "filtration" models involve faulty conceptions of effective cognitive security defense. On the account I develop, almost all effective countermeasures fall squarely into the category of stewardship and self-binding. This is because I'm concerned specifically with the efforts of democracies to enact cognitive security, and in ways that protect our epistemic dependencies. To do so in a way that imagines overwhelming substitution of judgement on matters of content and viewpoint will tend to turn out to fail to protect these animating interests.

Of what might remain, as examples of genuine epistemic paternalism that can potentially meet Bullock's challenge (where we would prioritize epistemic well-being over other types of well-being), there may be a legitimate need for stronger justification. It might need to be shown there is no less restrictive way to protect the targeted public good that is acceptably effective. Such a defense in turn demands the intervener be able to provide evidence that the chosen intervention really will be effective, not just at removing the targeted content, but at protecting the underlying interests that motivated the identification of this content. The empirical and theoretical considerations that might inform both requirements are central questions in the coming chapters - what are our public interests in the health of the information environment, and what kinds of policies can we protect these in the fact of cognitive security threats? I begin to build a conception of this public interest in the next chapter, where I develop a conception of epistemic environmental dependence to describe the extent to which the various forms of epistemic dependence we engage in themselves depend on shared reliance on properties of the epistemic commons.

Chapter 3 - Environmental Epistemic Dependence

The aim of this chapter is to describe epistemic dependence generally and introduce a new conception of epistemic dependence on the health of the information environment in which we seek and encounter evidence and testimony and interact with other agents. This generates public interests in the information environment that should inform stewardship and cognitive security strategies, suggest constraints on such actions, and provide additional grounds to inform assessment of their efficacy and side-effects. The existence of this environmental dependence also provides a source of normativity when we assess actions and technologies that change the information environment. The nature and extent of our dependence will furnish reasons to assess whether some current or forecasted environmental condition undermines some common epistemic good.

This bulk of this chapter is concerned with situating a conception of environmental epistemic dependence within existing social epistemology literature, especially where it concerns epistemic dependence and reliance on the testimony of others, to help draw out the distinct characteristics of dependence on the epistemic environment from that we have on specific individuals, instruments, and communities. I'll argue that accounts of the role of identity in the epistemology of testimony must include an instrumentalist qualifier - we rely on others as part of heuristics that furnish contextual information we need to meet the standards of justification. On this view, disputes about how much we need to know about the speaker to justify belief in their evidence are better understood as

problems of context, not problems of identity. One important function of the information environment is then to furnish this context reliably.

Adequate context evaluation is often impossible for an individual, which underlies a major concern within the epistemology of testimony, that of reductionism, which I will discuss at length in section 4. One way we solve this problem is the division of epistemic labour, relying on others who are better positioned to assess contextual information. Thus, a second important function of the information environment is to allow us to locate epistemic peers and engage with them. I argue that this feature is mundane in ordinary epistemic environments and easily escapes our notice, but online environments generate an enormous scope for mediation of testimony that complicates this reliance.

This chapter concludes with a definition of environmental epistemic dependence that in turn entails a conception of a kind of environmental epistemic trust. These form the basis of argument in Chapter 4 - that our shared epistemic environmental interdependency generates ethical considerations analogously to the way it does in the context of ecology and environmental ethics. There I will argue that if cognitive security policies should directly promote, or at least not diminish, our epistemic environmental dependencies, it should be informed by a conception of epistemic environmental health.

3.1 Justification, trust, and epistemic dependence

What is special about epistemic dependence, and why do I think that this category should include properties of the information environment? In this section I want to introduce a few of the problems in contemporary epistemology that have given rise to attributions of various forms of epistemic dependence, and conceptualizations of what exactly an epistemic dependence consists of.

Epistemologists have long distinguished between knowledge and mere true belief, and what makes knowledge more valuable is whatever entitles us to treat the belief as true. In the language of traditional analytic epistemology, this is the justification component of justified true belief accounts of knowledge, and there is an expansive literature working out the precise conditions that must obtain to secure justification in the face of an array of problem cases. The adequacy of any account of justification involves a two-way relation, with two directions of fit. It must aim for and secure the truth of beliefs - it must fit the world as it is independently of our agency and interests. But it also must be appropriate to our interests and agency. My beliefs about barns aren't usually about barn-parts, or barn-reflections, or the proximal stimulations on my retina, they are about what it is about barns that matter to us. Here the direction of fit is between the belief's justification and social and human constructs. Justifications based on sensory stimulation won't necessarily align with our interests in tying some class of perceptions to barn-concepts, because somewhere along the line we care about the causes of the barn-perceptions, and

the barn-seeming things that give rise to them. We might care that some content we find in a digital information platform is viewed positively by our peers, perhaps indicated by an aggregate rating. But we don't care about rating simpliciter - we care that the rating is motivated by and subject to the same epistemic attitudes and interests with which we assess the information. We care about the norms involved in generating and representing those ratings.

On internalist accounts of justification, having the right internal states secures the claim that we are not just in the possession of some true belief, perhaps by way of some minimal representation of some external state of affairs, but also providing some additional cognitive achievement that can guarantee the representation's correctness Bonjour makes the case for this internalist component in several examples involving clairvoyance (BonJour, 1985), where a person's true beliefs are justified by factors the person does not have evidence for. In one example, Norman has a reliable clairvoyant faculty but no evidence for its effectiveness. From the point of view of Norman, the belief is unjustified. Even if there exists evidence that the faculty is reliable, Norman has not encountered it, it is not part of his cognitions, and without this internal component there is no justification. It matters that the agent is not, in Bonjour's term "subjectively irrational", where "...the acceptance of a belief is seriously unreasonable or unwarranted from the believer's own standpoint" (ibid. p 61). This sort of consideration supports internalist accounts of justification that supervene, at least in part, on internal states of the

agent - it is not enough to have a true belief, and to be justified, but one must possess within one's own cognitions states that provide the justification with sensitivity to defeat.

And yet, we frequently have evidence that we should not trust these states. Our senses sometimes mislead us, our memory sometimes fails, our reliance on others for knowledge is often misplaced, and cognitive biases often erode the rationality of our inquiries - and yet we depend on these as a matter of course. If these conditions can obtain while escaping our detection, and we know that this is possible, we likewise risk subjective irrationality. We need additional reasons to trust, or we need sensitivity to internal defeaters, to justify beliefs based on the exercise of our own cognitive and perceptual agency. From this point of view, we can hold facts about the agent constant, and modify external facts, and find that the justifications for belief vary - we can have the internal states that would justify a belief, but these aren't caused in the right way. Goldman (Goldman, 1976) imagines a case I'll refer to as BARN, where the subject, without realizing it, enters an area with only one real barn, but countless fake paper-mache barns. They see the fake barn, behind which is a real one, and correctly believes there is a barn, justified by the perceptions of the red barn and their knowledge of red-barn concepts and so forth. But had they known of the many fake barns, they would not have believed themselves to be justified, and an observer, knowing of the fakes, would assess the belief as unjustified - its truth was blind luck. In such an environment mere registration of genuine barn-perceptions is insufficient to warrant knowledge of the existence of a barn.

Ziółkowski (Ziolkowski, 2016) offers a similar case, where one selects a thermometer from a box of 100 and uses it to measure the temperature. The thermometer is accurate, the resulting belief is true, and justified based on the instrument and the customs and practices involved in its use and manufacture, but unbeknownst to the agent, the other 99 thermometers in the box were faulty - the correct reading emerges from an improbably lucky event. Norman's beliefs are formed by a reliable process but lack internal properties that would justify them - Norman possesses no evidence for their reliability, only inductive evidence of their production of true hunches. In BARN and THERMOMETER, one has adequate internal properties to justify belief, but external facts threaten justification - there is a reason to not trust the belief, and it undermines justification whether one knows it or not.

Both cases illustrate classes of practical epistemic problems that we face generally, and which epistemic dependencies often help to mitigate. If BARN cases are a risk, communication with others can help us realize this, and to produce evidence to distinguish fake barns from real ones. Where we depend on instruments, we depend on understandings of their accuracy and failure modes communicated by others, and the behaviour of others involved in the production and maintenance of the instruments to hold these facts stable so that we can generate rational credences for the evidence we produce with them. We simply can't check for all these possibilities on our own, all the time, nor can we be attuned to all the possible signals that we need to perform more rigorous checking than is customary.

There is a double reliance on trust - we extend trust to sources of indirect evidence for beliefs, that they indicate what we take them to indicate, but also trust on others to alert us to the existence of abnormal conditions, where for reasons we don't appreciate ourselves, we ought to modulate this trust. This in turn implicates two distinct notions of trust. As Baier (Baier, 1986) argues, when we trust another agent, we make ourselves vulnerable, and we have a reactive attitude of betrayal if that trust turns out to be ill-founded. This is because trust fills in a gap between what we have independent reason to believe about what the agent might feel pressured to do and what they actually decide for themselves to do. We also trust non-agents in the very same sense that we make ourselves vulnerable to breakdowns in their predicted functioning. With the agent, we can't force them to uphold trust, but we believe they will elect to, with the object, we can't verify that our theory of its trustworthy properties is correct, but we act as though it is. Nguyen calls this an "unquestioning attitude" (Nguyen, 2020b), and it marks a boundary between the evidence we firmly have in hand, and that which also has to be the case if the belief really is true. In an example Nguyen paints in compelling detail, a climber trusts a rope against a background of knowledge about such ropes generally, and particular knowledge about a specific rope, and its context of use. Trust enters the picture in the gap between all the things the climber knows about the rope and the things they do not, where the climber suspends the search for further evidence and uses the rope as though the case has been made that it is sound. In fact, that case is never made, and this a permanent condition of our knowledge.

In a vivid thought experiment designed to show that the mind might extend beyond the individual brain, (Clark & Chalmers, 1998) ask us to imagine Otto, an Alzheimer's patient who relies on a notebook to augment his short-term memory. The fact that a cognitive process has been re-implemented in an external medium, is not, according to Clark and Chalmers, a reason why we should not think of the notebook as part of Otto's cognitive states. Whether or not we should think of processes outside the agent as properly cognitive, the illustration is particularly apt to describe external epistemic dependencies. Otto depends on the cognitive process that includes the notebook in order to achieve certain kinds of epistemic standing. An evaluation of his epistemic states or justification for some belief is incomplete if it excludes the contents of the notebook.

Clark and Chalmers take from this a moral about the metaphysics of cognition, whereas I am interested here in that of belief, in particular, when justification for our beliefs is transmitted from others. Dependency is epistemically significant because it is not merely belief content that is externalized, in the sense that the correctness of my belief that it is snowing trivially depends on whether or not it is snowing, but the grounds for justification. In the case of testimony, justification is directly transmitted from the speaker to the hearer if my belief that it is snowing is caused by hearing this reported on the radio, where I possess no evidence whatsoever myself. I treat the radio as a source of information - that it is snowing, and justification - I believe that it is snowing because hearing such on the radio is good evidence that it is the case. Thus, a network of upstream

facts about radio are implicated, and a critical issue is the extent to which I must appreciate these to receive justification.

If, in some refinement of the thought experiment, Otto relies on some other agent to maintain the notebook, we then must bring that agent into scope when assessing beliefs Otto has. Perhaps the agent is colourblind, and Otto knows to make appropriate allowances, Otto believes that the sweater in the closet is green because the assistant recorded in the notebook that it is red - we would still allow that Otto is appropriately connected to truth-making conditions, and that Otto knows that the sweater is green. We need to know about the notebook, and the agent, and the constitutive role they play in the formation and justification of Otto's beliefs. Both are part of the subvenience base for Otto's justification, in addition to the content of Otto's beliefs.

Some dependencies are optional and adopted instrumentally for convenience or economy, but many others are basic in their particulars or functional roles. Social epistemology draws attention to our reliance on the epistemic efforts of others and has expanded the scope of recognized external dependencies as it has evolved. Much of what we believe and know depends in part on reliance on testimony (Fricker, 2006a; Lackey, 2008) , identification of peer disagreement (Christensen, 2007) and experts (Goldman, 2001), the reliability of our information resources (Goldberg, 2020), the diversity of our epistemic environment (Fricker 2013), instruments and their design (Goldberg, 2012), and freedom from distorting external conditions such as a bad ideology and epistemic

pollution (Levy, 2022; Srinivasan, 2020). I'll focus on testimony in this section, as it illustrates the extent to which we do not just depend on externalities for content, representations of what is the case, but for justification. I really can't say anything about the current temperature in Celsius without depending on a good deal of ineliminable externalities, including agreement about sources of evidence, calibration of instruments, and temperature concepts. In the use of instruments, my belief that the temperature is 78 is justified just because the display on the thermostat tells me such - I have outsourced justification in part to the design and maintenance of the instrument forming what Goldberg (Goldberg, 2020, p. 2783) calls "design dependence".

There is a significant literature on the distinction between trust and reliance, but in the epistemic context, especially where non-agential trust is involved, it's useful to think the line is just that of credence. I might treat the fact that the door is open as evidence that the cat got out, for the purposes of beginning my search urgently, while being quite aware that I don't actually believe that the cat is out. Here I rely on the signal of the open door for a pragmatic but non-doxastic reason. But in the background of this is a network of beliefs that I really do trust and depend on, about cats, and doors, and the dangers of my neighbourhood. Here I adopt Nguyen's unquestioning attitude to cut down the amount of uncertainty I need to entertain. Likewise, if I conclude that some proposition is false because I can't find a reference to it using a search engine, to the extent that I believe this conclusion I have extended trust to the comprehensiveness and reliability of the search engine, and the broader causal situation (for instance, that some network layer isn't

filtering out results on this topic). The literature on epistemic dependence is not always attuned to the distinctions we might make between dependence, reliance, and trust, but throughout this chapter I will use the term trust because it best captures our vulnerability to hidden factors that we don't know about.

Epistemic dependence is usually associated with externalist accounts of justification - we ourselves might lack all of the properties that would justify our belief, but form an epistemic relationship to some externality to acquire it. If there is evidence that Norman's clairvoyance is reliable, perhaps possessed by some community of experts who have studied Norman, justification would depend on their methods and practices, but we should still require internal justificatory conditions, and require that Norman possess knowledge of these experts and their reliability. In Clark and Chalmer's original example, Otto has Alzeheimer's, and becomes dependant on the notebook knowingly and gradually, over time, and thus possesses evidence for its reliability. The situation would have to be parsed more carefully in some future state where Otto cannot recall or access these states (a "forgetful Norman" variation). It might seem unpalatable to downgrade our estimation of Otto's total knowledge just because of the loss of appreciation of the justificatory process, but at some threshold we are forced to confront the problem of subjective irrationality. We can replicate the same problem for an agent who has become dependent on access to digital sources of information. We might believe that P and believe that we can reconstruct the reasons for P by searching the internet, and believe that our belief is thus justified, all without re-executing the search procedure. What is at

stake with dependence is that properties outside of one's own agency play decisive roles in securing possession of knowledge, as a basis for justification, not just that one knows P, but that I know P because Q, where Q is external. This brings into scope questions about how we locate and safely interact with dependencies. What if some opaque property of the search engine defeats the justification?

3.2 Dependence, testimony, and context

The seminal work on epistemic dependence is (Hardwig, 1985) who argues that "... appeals to epistemic authority are an essential ingredient to much of our knowledge" (p. 336). We should construe "authority" broadly here - there is a literature within social epistemology on the significance of expert-disagreement, but in a more mundane way we rely on authority that stems just from proximity to evidence. I might not have seen where the cat is, but you may have, and possess epistemic authority as a result. If it is rational to depend on the testimony of others that we have reason to believe are epistemically better positioned, is it rational even when we can't access or assess their evidence? We can have evidence for the belief in some proposition that is not evidence about that proposition, but evidence about the relationship some other knower bears to it - that they have expertise and access to evidence that lends authority to their testimony that P. This raises the problem of reductionism about testimony - do we have an entitlement to believe based on testimony in the absence of defeaters, or do we require information from some other channel that provides positive reason to trust the testimony?

Reductionists take the latter view - testimony can only justify belief when we possess independent reasons to believe it. But this gives rise to difficulties because not only do we lack the expertise and the evidence of the expert we rely on, we also tend to lack adequate higher-order evidence about the evidence testimony provides us, such as the bona fides of the authority. Coady (Coady, 1992) argues that our reliance on testimony is so systematic that we almost always lack this higher-order evidence. Millgram (Millgram, 2015) argues that we can't even check on the evidence from experts if we did devote the time to acquire it, because the expertise is so rarefied and interdependent that it's not possible to reverse-engineer it for individual comprehension even if we had the time and capacity - many spheres of knowledge are irreducibly distributed. Anti-reductionists argue we need no such reason, and that we have a presumptive entitlement to accept testimony. We receive justification by direct transmission, if the speaker possesses it, then the hearer does.

Some examples of testimonial exchange:

BYSTANDER: A stranger in a city I've never been to before tells me how to get to the airport.

PLUTO: An astronomy teacher tells us that Pluto is not a planet

JOHNSON ROD: My mechanic tells me that the sound I complain of indicates costly replacement of both of my vehicle's Johnson Rods.

Croce (Croce, 2022) observes that the current state of this literature oscillates around differing accounts of the appropriate permissiveness with which we can take testimony as a warrant for belief, and results in a dilemma. If we are reductionists, and testimony requires non-testimonial evidence, then we have to know too much to accept testimony, and most of it doesn't count as knowledge. If we are anti-reductionists, the bar for testimonial knowledge is set too low, stamping beliefs with the seal of justification with insufficient connection to truth-making properties; risking gullibility. In either case, testimony itself is an inadequate route to knowledge because it's either too had to get, or too easy. And yet, to Coady's point, the underlying pragmatics of inquiry are such that we rely on it as a matter of course. It might seem we could defend the epistemic value of testimony by way of a particularist/generalist account, where the general presumption is grounded in spot-checking verification effort, but because the function of testimony is to obviate the need for the epistemic labour that in practice we cannot perform, a particularist theory would amount to either reductionism or a regress - we'd need to know when we needed to perform such labour, and to know when it was performed adequately.

Another move one might make is to question the standard of justification required for knowledge, by examining intuitions about when we really know something and the role these play in evaluating accounts of justification, and cases where we must judge the probability of possible defeaters obtaining. Maybe the problem comes from seeking ironclad accounts of justificatory adequacy. Some of the judgments about such intuitions that philosophers have relied on appear to be contingent on culture, language, socio-

economic conditions (Weinberg et al., 2001) - there is substantive disagreement across demographics as to when we can say we know something. Ganeri (Ganeri, 2018) calls this the universality thesis, that "... properties of the English word know [...] that have been studied by epistemologists are shared by the translations of these expressions in most or all languages" (ibid. p. 21), and provides evidence that we should reject it. For example, Ganeri argues that within the Indian philosophical tradition, there are prominent approaches to epistemology that do not give rise to the problems of justification found in the anglophone analytic tradition because they conceptualize the possession of knowledge just in terms of epistemic success. If we can act on the belief that P successfully, we know P. However, such an approach still faces the class of problem that I have in mind. How should we behave as epistemic agents in order to maximize epistemic success and minimize failure - especially where we depend on the behaviours of others? We don't know that the mechanic is honest, that the teacher possesses appropriate expertise, that the bystander knows we didn't mean the private airport on the other side of the city. On any account of the dependency involved in testimony, we have a vulnerability problem, even if we modulate our account of justification. The reasons that might secure justification in PLUTO are of the same sort that might make us wary in BYSTANDER and JOHNSON ROD - they relate to contextual knowledge about the roles and responsibilities of agents we depend on. BYSTANDER may be ignorant but bashful, and guess, or perhaps is playing a prank on us, but we suppose we could tell, or that this is unlikely. The mechanic in JOHNSON ROD has an incentive to mislead me about this fictitious part, but also is exposed to mechanisms of accountability that I may

understand and appreciate. Like Goldman's BARN example, I need to know something about the way the world is in the domain of my inquiry to reliably connect my belief with justifying reasons.

Two competencies are required to assess testimony - one is the determination of relevant context, another is the assessment of the relevant context. These are problems for the reductionist and anti-reductionist alike, because even the lower bar of anti-reductionist requires sensitivity and monitoring for defeaters. One such defeater is deceit, as we might find in the cases of BYSTANDER and JOHNSON ROD. Not only must we have the ability to identify epistemic authorities, we must be aware of their incentives, interests, and possible idiosyncrasies of the testimonial exchange.

Lackey offers this account of the acquisition of knowledge from testimony: "For every speaker A and hearer B, B knows that P on the basis of A's testimony that p if and only if: (1) B believes that p on the basis of the content of A's testimony that p, (2) A's testimony that p is appropriately connected with the fact that p, (3) B has no defeaters for A's testimony that p, and (4) B is a reliable functioning recipient of testimony. 5) the environment in which B receives A's testimony that p is suitable for the reception of reliable testimony" (Lackey, 2003, p. 716)

These conditions arise in the face of a variety of scenarios Lackey describes where B appears to acquire knowledge on the basis of A, but in fact the claim to knowledge is

undermined because the knowledge is acquired only by way of luck, or where there are defeaters A and/or B are unaware of. The fifth condition, that of the environment, is the one that I am most interested in here. Lackey imagines a case similar to BARN that I'll call VILLAGER, where one happens to ask the one honest person in a village where the norm is to not speak the truth to outsiders. One has a true belief, based on accurate testimony, but there are elements of luck and risk that disqualify it as knowledge. Either it is accidental that it is true, or it is accidental that it is justified. Luck precludes knowledge, as in Gettier cases where it is only by happenstance that inferences lead to truths when the causal story does not map to the inferential one - for instance glancing at a broken clock at the moment where it indicates the actual time. Lackey argues that one must thus be in an "epistemically suitable environment" if testimony can transmit knowledge, one that does not contain such things as paper-mache barns and locals who mislead others.

One might wonder if there's something unnecessarily austere about a concept of knowledge that would deny knowledge-attribution in VILLAGER, especially if one focusses narrowly on the particulars of the interaction. It involves a standardly justification-conferring procedure where the subject acquires a true belief. It's not lucky in a stopped-clock sense, because the testifier is perfectly normal and accurate. It is an underlying presumption that is false - that other agents are generally reliable sources of evidence and can transmit justification. But this agent is, and one might even possess some positive reason to trust (satisfying some variants of reductionism). One lacks

appreciation of the elevated risk of undercutting defeaters, including those to positive reasons to trust, and thus one cannot be said to have a method of justification that is connected adequately to truth-making properties. If it was, there would have to be added rigour to account for the fact of probably misleading testimony.

It is these facts about the environment that disqualify the process as reliable. One's appreciation of the environment in which one encounters evidence is impoverished in ways that generate a defeater that must be monitored. This diagnoses the problem in BARN - ordinary barn perceptions generate weaker justification in an environment with many fakes, and thus requires increased corroborative labour. Deep fake videos generate identical difficulties - their presence in an information system reduces the amount of knowledge available with the same about of epistemic work (Fallis, 2021), because presumptive trust is weakened, and elevated proof is required. Because we have to do more work to acquire this proof, the environment is quantitatively poorer in terms of belief-apt information. Knowing that some additional proof is required erodes what Rini calls the "epistemic backstop" and will "...gradually eliminate the epistemic credentials of all recordings" (Rini, 2020, p. 8). An entire type of justification-conferring dependency is undermined, thus depriving tokens of their previous warrant-generating properties.

Croce, after introducing the dilemma about testimony described above, argues that it can be resolved by both the reductionist and the anti-reductionist. The anti-reductionist argues we have a presumed entitlement to belief based on testimony that does not require

epistemic work to acquire, and the worry is that this would lead us to accept all manner of misleading and untrue propositions. Croce argues that because anti-reductionism is consistent with monitoring for defeaters, the worry can be mitigated. Undercutting defeaters bring into question a reason for belief, for instance, the discovery that all the other villagers would have lied (vs rebutting ones, which contain reasons to the contrary of the proposition expressed). Consider the example of interviewing a job applicant, where one has reason to be wary of accepting claims at face value. Croce argues that it is enough just to know this, that recognition of the context of testimony presents a relevant undercutting defeater that would appropriately modulate credence. Because one monitors for them, and because this is consistent with anti-reductionism, the anti-reductionist can thereby reject the charge that the account is implausibly permissive and prone to gullibility. Note that the risk reduction strategy is outsourced here to appreciation of the context.

The reductionist can also escape the dilemma, because while it does require epistemic work to acquire a warrant to accept testimony, the burden does not fall to the individual. The problem for the reductionist was that it makes knowledge acquisition too hard because one must do too much epistemic labour to acquire knowledge, such as identifying and verifying epistemic authority in order to generate positive reasons to believe the source. Consider JOHNSON ROD - while the mechanic has incentive to deceive me, and while I lack the requisite knowledge of auto mechanics to validate their claim, I can rely on my knowledge of the social practice of getting cars fixed, which

entails understanding that the shop is subject to formal and informal mechanisms of accountability. We rely on what Croce calls a "division of epistemic labour" (doubtless inspired by Goldberg's (Goldberg, 2011) use of this term), where we do a little work to connect our appreciation of the situation with what is known by others, such that we know how a putative warrant to accept testimony is subject to existing and ongoing social scrutiny. I can generate various counterfactual justifications - 'if it was not acceptable to believe this mechanic, then I would have learned from reading about them online, or they would have been shut down, or the positive article I found would not have been published', and so forth. Note that here the risk reduction strategy is outsourced to the "epistemic community", which "... takes on the burden of gatekeeping"(Croce, 2022, p. 17).

I think both moves are incomplete as they stand, because they relocate the epistemic challenges that the agent must meet rather than resolve them. It is where they are relocated to that I am interested in - both strategies depend on contextual knowledge, which I will argue is ultimately furnished by the information environment in a general way, not provided in known set of epistemic interactions. And thus we are back to Lackey's environmental condition, but with more nuanced appreciation of what is needed to meet it. Consider the reductionist strategy - we don't depend on community in the sense of some class of individuals doing directly relevant epistemic labour, but instead the various systems, interfaces, and signs that make representations about such labour, which may or may not be reliably connected to the exercise of such. Our dependency

largely falls to these epistemic systems, not individual testimony conveyed by them, and we face the same dilemmas as with testimony in terms of assessing our warrant for trust and sensitivity to defeaters. I think Croce is right that we resolve these by working with others, but this is complicated when this coordination itself takes place within digital information platforms.

I want to flag this dependency as a kind of trust in Nguyen's sense of an unquestioning attitude discussed in section 1. We engage in some minimal verification or monitoring of the signals we have about the dependencies involved in our justification, but ultimately adopt an unquestioning attitude of trust that we have an adequate appreciation of the context to inform our selection of appropriate verification and credence. We have to suppose that we know what kinds of things go wrong, and how to monitor for them, even just to engage productive skepticism. On the anti-reductionist picture, the agent must be capable of identifying and responding to contexts and must additionally be cognizant of their own shortcomings and likely failure modes in this regard. I must know that I'm in a situation where some class of defeater is probable (perhaps including defeaters within my own agency such as cognitive biases) and have effective monitoring and countermeasures against forms of defeat and deceit. I have a dependence on context-savviness, which is in part an external dependency on the mechanisms that furnish information about context to me, what I call the context-signalling adequacy of the environment in the next section. The reductionist move wears the problem I have in mind on its sleeve - we rely on the division of epistemic labour, that it is performed, and adequately, and that we appreciate

the extent and limitations of this. Relatedly, it requires we understand something about epistemic communities, how to find them, how to read them, and what their boundaries are. In both cases, the problem of "how do we obtain warrant for testimony" is answered by dependency on ambient externalities that help us locate the specific externalities we require. This is a description of an epistemic environmental dependence.

There is a Getterization risk with this dependence. In standard Gettier examples, peculiarities in the connections between justification and truth undermine the justification for a belief. In such cases the truth-making and justificatory conditions come apart causally, but not inferentially. These have been so significant in epistemology because almost all of our knowledge and beliefs are based on partial knowledge that falls short of certainty, and Gettier cases highlight similar riskiness in otherwise sound justificatory procedures. The testimonial Gettier cases I have in mind are those where testimony corroborates a true proposition, the contextual signals that accompany the testimony further justify belief in the testimony, but where we would not be justified in increasing our credence in the usual transitory nature of the justification-conferring properties of this encounter to future encounters.

An example of this is the persona-building (Mirsky et al., 2023) activity of disinformation accounts in social media systems. These will post accurate content and engage in normal interactions in order to obtain various signals of trust and acquire influence, then mixing in messages intended to mislead - a hybrid of the VILLAGER and

THERMOMETER example. It would be normal for this successful exercise of epistemic dependence to contribute to forming second-order evidence about the source and the media platform and our understanding of it, that justifies adoption of an unquestioning attitude, but this would be unfounded. These cases can also occur indirectly, for instance, when economic motivations for producing content are only accidentally correlated with epistemic norms. A normal search leads to a normal looking publication with accurate information, but where many other pages in the publication are not accurate, but instead random attempts to produce content that the user seeks. The increasing mediation of large language models in online search and information retrieval systems further escalates this risk. In such cases, we have accurate testimony, and it is of the right kind and type, and is apparently normal. But it was produced in ways that are not responsive to the interests and norms that govern our inquiry, and we don't know enough to even know what we should have verified in order to learn this.

We can be more precise about what it would take to meet Lackey's environment condition by examining Sosa's concept of epistemic safety (Sosa, 1999) and Pritchard's related conception of epistemic luck (Pritchard, 2006). Sosa's condition is expressed as a counterfactual, that "S would not believe that p without it being the case that p."(Sosa, 1999, p. 146). Although Sosa is ultimately concerned with skeptical problems out of scope here, the principle nicely captures problems with knowledge acquisition and justification processes which are inadequate to their environments. These must not only be sensitive to truth, but robust to patterns of failure. The implementation of some

robustness-enhancing process must be done in relation to local properties of context or environment. Barn-beliefs in Goldman's polluted landscape are not safe by Sosa's principle because I would easily believe I saw a barn when I did not. Pritchard's anti-luck principle requires that "...there is no wide class of near-by possible worlds in which S continues to believe the target proposition, and the relevant initial conditions for the formation of that belief are the same as in the actual world, and yet the belief is false." (Pritchard, 2006, p. 281). This comports well with Lackey's VILLAGE example, and Goldman's BARN - the problem is that there are very nearby worlds in which S believes P, using the same process, but where P is false. Rather than take up the task of formulating safety or anti-luck in a way that engages with the counterexamples that have arisen in subsequent scholarship, I will operate with just a minimal conception - that a safety principle is required to articulate the environmental condition, however epistemologists might agree one should be formulated.

We are in an epistemically suitable environment when we are able to engage in epistemic behaviours that are safe for the kind of inquiry we are undertaking. We tend to implement these as heuristics - as we learn to use some information space, we build up a safety strategy, we trust some source, seek some particular mark of credibility, or learn the safety practices of some other agent we find. Returning to the idea of trust as an unquestioning attitude, we can say that we have rationally engaged trust when we have reason to believe (or an absence of defeaters) that our strategy is safe, that the array of unquestioning attitudes towards environmental features is safe. Beliefs in the BARN,

VILLAGE, and THERMOMETER cases do not meet the environmental condition, and that is the general reason the beliefs in question are not justified. This gives us a new question to pursue - what kinds of environmental conditions are generally safety-conferring, and how do we know when they obtain? This question will become important when I turn to questions about the protection of epistemic environments. In the next section I'll focus more closely on what we need to satisfy the condition.

3.3 Community and context

In this section I delve into the problem of anonymous testimony as it arises in the social epistemology literature. My motivation is to show the decisive role that the context of our encounters with testimony plays in securing our everyday reliance on the epistemic division of labour. This matters to my broader project because it brings the context-furnishing adequacy of an information system into scope when analysing the susceptibility of an information system to manipulation and developing a working theory of the vulnerabilities that a cognitive security policy might wish to address.

Michaelian (Michaelian, 2010) surveys psychological literature on the detection of deceit and draws a pessimistic conclusion about our ability to detect problems with testimony which apply to both of Croce's models. If justified reliance is predicated on effective monitoring for defeaters like deception, we are not up to the task in most of the situations where we depend on testimony. On the reductionist model, where we seek

positive reasons to trust in advance of forming credence, those reasons are themselves subject to defeaters such as deception, which thus must be detected. As a result, by way of Michaelian's results we should agree that such procedures are not much better than blind trust if are we are so poor at monitoring. Similar problems arise relating to cognitive biases (Hannon, 2021), and manipulation, of the sorts described in Chapters 1 and 2. If we are poorly equipped to detect defeating conditions, merely being in a position to do so can't do much lifting in an account of justification.

Michaelian observes a gap between what we could possibly know at a given point in time when encountering testimony and what we would need to know to acquire justification, and argues we should think of the relevant epistemic properties as modal - we acquire the information from the community and context over time, and the epistemic status of the belief varies in step. We can then contain pessimism about testimony because if the credence-subvening properties are modal, determinations of generated justified belief are not settled at the moment of transmission. Like credence, it extends over time as exposure to sources of defeaters is increased, "... if the speaker were to have been dishonest, then the subject would have received other evidence indicating dishonesty (physical evidence, third-party information, etc.)" (Michaelian, 2010, p. 423). This approach entails an environmental dependence in the same way the Croce's does. We need reliable and extended access to relevant information about the speaker, the domain, and the assertion to correct for our own limitations. A cooperative epistemic environment furnishes the information we require to reliably conduct such monitoring,

and to benefit from divisions of labour and the communication of meta-evidence. On whichever account of safety or anti-luck principles we prefer, to secure such we need to reliably bridge our representation of the context to relevant truths about it.

I'll further elaborate this conception of environmental dependency by considering the epistemology of anonymous assertions. When we encounter testimony that we cannot attribute to an identifiable speaker, Goldberg (Goldberg, 2013) argues we have a diminished entitlement to justification. We ordinarily think that to assert something involves adherence to some norm, perhaps knowledge, or authority, or similar. One ought not assert that P if one does not know that P, and one is entitled to expect such norm-adherence from others who assert. Goldberg argues that however we construe norms of assertion (a matter of disagreement in epistemology literature), that these exist in turn generates a problem the hearer must solve, "...either the speaker has the relevant epistemic authority or else the assertion was unwarranted." (ibid. p. 129). As Goldberg sees matters, two kinds of interactions are involved in resolving this. The first are on the side of the audience, that it must possess "counterfactual sensitivity" to defeaters. This requires four classes of information -

1) information regarding the speaker;
2) information regarding the act of communication (or assertion) itself;
3) information regarding the content of the communicated message; and
4) information regarding the context of the communication. (ibid. p. 145)

The second kind of interaction centres on the speaker and is grounded in the mutual expectations between speaker and hearer. The speaker is aware that the audience is actively seeking these four kinds of information in the course of evaluating testimony. They appreciate that they might be questioned, that their delivery is being scrutinized, that the audience may contain experts, and so forth. Additionally, they are aware of mechanisms of what Goldberg terms "policing". Goldberg's use of this term can be traced back to earlier work on complex chains of testimony, where justification for beliefs based on such chains depends in part on the reliability of the processes that maintain their fidelity. One such process is that of “remote monitoring and policing”(Goldberg, 2011, p. 118), in which external (and usually asynchronous) intermediaries in testimonial exchanges play crucial roles in guaranteeing the integrity of the transmission and the reliability of the knowledge generated. If we think of some epistemic process as involving various distinct communities, and specialist and public communication channels transmitting information between them, we rely on the epistemic processes internal to each of these, and our own community's ability to detect when these processes fail. Such policing is part of a "system of very extensive social relations" (ibid. p.119). Sanctions include diminished trust, loss of status and opportunity, moral disapproval, alongside and more formal sanctions related to professional roles, regulation, and law.

Goldberg argues that two agents, with identical knowledge, receiving identical testimony, acquire differing justification for belief if one receives the testimony

anonymously. The difference is not one of content in the testimony, nor of competency of any kind in the hearer. The problem Goldberg diagnoses is that there are mechanisms for enforcing assertoric responsibility that are absent in the anonymous case, and that both the speaker and the hearer are aware of this. This fits nicely with Michaelian's modal view of credence - that it unfolds over time in the presence of various social monitoring processes - including Goldberg's policing practices. Speakers are aware that their assertions will be subject to this, and are thus aware that any credence their assertions might generate will be modulated by such processes - and understand they will do well to produce assertions in a manner most likely to interface well with these. The problem we face in all testimonial contexts is that we must close the gap between "... what an audience is entitled to believe on observing an assertion, to an audience's being entitled to believe (being justified in believing) what was asserted" (Goldberg, 2013, p. 140). We receive information that justifies a variety of possible beliefs, but these are not necessarily identical in content to that which is asserted by the speaker, nor what we are justified to believe.

In BYSTANDER, a weak instance of anonymous assertion, we know little, but not nothing about the speaker. We can observe their manner of delivery, apparent confidence, and other contextual clues about their reliability. Even taking into consideration Michaelian's arguments, these furnish some information relevant to the evaluation of their testimony. Goldberg argues that the mutual expectations between speaker and hearer include awareness of policing practices that generates incentives for speakers to

adhere to the norms of assertion, which might even be onerous, or practically or psychologically inconvenient. The speaker will be aware when "...asserting in a context involving few or no mechanisms available for the encouragement (or enforcement) of assertoric responsibility" (Goldberg, 2013, p. 148), might thus fail to summon the necessary resources and discipline to adhere to the norm, and audiences should be responsive to this likelihood - and maintain the gap between what is asserted and what is believed. The problem here is that the speaker knows they are anonymous, and might be less epistemically virtuous.

Ivy (Ivy, 2021) rejects both aspects of Goldberg's view. Firstly, that speaker-side interactions in Goldberg's view involve an implausible "punitive model of assertoric practice." (ibid. p. 466). Secondly, that the audiences to anonymous assertion actually possess sufficient contextual information to afford appropriate counterfactual sensitivity. I will discuss these in turn, as the motivation argument and the context argument.

First the motivation argument. Ivy argues that we should understand Goldberg's account of policing argument as punitive, and as implying a claim about motivation - that absent accountability mechanisms, speakers lack motivation to adhere to norms of assertion. It is only because of these that speakers are generally honest, and thus generally believable.

Ivy denies the motivational claim on the grounds that it underdetermines observed behaviours. People are often honest even when not exposed to punitive correction. This is certainly true, but unless one wishes to argue that the mechanisms of policing are thus redundant, it's not clear that it speaks to Goldberg's interest in policing generally - as a way of maintaining an information environment in which we can rely on heuristics to help us form true beliefs without having to verify. However, even granting this argument, it's not clear to what extent accuracy depends on motivation. Even if we intend to be inaccurate on some point, the pragmatic requirements of the communicative exchange will require accuracy on many points if the desired perlocutionary effect it to be achieved. The amount of required collateral accuracy will depend on the context - I'll need to say a lot more accurate things to convince a job interviewer of some falsehood about my skills than they will to tell me something false about the job. Communicative interactions are often deliberately constructed to weigh these pragmatic considerations against predicted motivational deficits. On this view, telling the truth is not merely a matter of motivation. We are constrained by language, by available information, by shared understandings, and the constraints of the world around us such as it pushes back against our efforts.

If motivation is less efficacious than the contextual elements that demand that much of what is communicated be scaffolded on accurate content, then we can downgrade the expected impact of motivation (no matter what we might think of the punitive view). The problem with anonymity is that the audience must do more work if the context doesn't

generate pragmatic constraints on inaccuracy. For Goldberg, the loss of incentive on the speaker to be certain creates increased burden on the audience to close the gap, one which might be unsurmountable. This is a general problem for the transmission of knowledge, not a special one for anonymous testimony, that in the absence of appropriate incentives we must adjust the division of epistemic labour. An anonymous press generates similar burdens as an unregulated one which directly impacts how much information we can get from it, and what kind of beliefs are justified by its content. Goldberg's point is that accountability mechanisms assure us this work has been done so we don't have to do it ourselves, but I don't think this view requires that the accountability be implemented in ways that attach directly to the speaker.

Moving then to the context argument - knowing facts grounded in the identity of the person making an assertion are valuable as an effective means for acquiring necessary contextual knowledge. Goldberg argues, "...publicness is central to assertion's playing the knowledge- and information-spreading role that in fact it plays"(Goldberg, 2013, p. 146), but knowing exactly whom is asserting is not central to the four classes of contextual information identified earlier. Consider an example I'll call ONLINE from Goldberg's article, where one encounters some piece of information online, from an author one does not know, and a site one as not familiar with, a case where "...you don't know who wrote this, you have no idea whether this site is monitored (and if so by whom), and so forth" (ibid p.148). If the proposition one encounters was made-up, and if this was on a matter on which we lack authority and thus have increased epistemic dependence, then Goldberg

concludes that "..counterfactual sensitivity is seriously diminished" (ibid.). It seems to me this diminishment is the result of an informational deficit that could be corrected while maintaining anonymity of the asserter - even if inconvenient and inefficient. I labour on this point because I want to highlight the extent to which we need these other channels anyways - on reductionist or anti-reductionist pictures of testimony we need to acquire meta-evidence. And for just the kinds of reasons that Goldberg's ONLINE limits our ability to be sensitive to defeaters, other features of testimonial context may as well. In the next section, I'll argue that the problem of forming appropriate credence in examples like ONLINE isn't because the statement is anonymous, or because it lacks context. It does, but because of features of the specific online system it is made in.

3.4 Trust and epistemic environmental dependence

In the previous section I argued that there is less distance between Goldberg and Ivy's argument than appears. But there is more substantive underlying disagreement - Ivy is optimistic about the informational richness of online anonymous context, where Goldberg is pessimistic. Both specifically draw attention to the unique affordances of online communication because this is where we most frequently encounter examples of anonymous speech. Ivy describes some highly anonymized online testimonial exchanges and argues that we are entitled to believe the propositions expressed, because the testimonial context is sufficiently rich (on either reductionist or anti-reductionist accounts) to be counterfactually sensitive to defeaters. Consider Ivy's example GLASS, within an information system that displays short textual messages originating from

devices that signal they are geographically proximal to the reader, with only the date, location, and self-selected username displayed. There are no affordances described for the verification of location, or the production of meta-evidence (such as how many times it was viewed, or interacted with, or rated as favourable in some way). As Ivy describes it, GLASS is anonymous and does not involve a matter where might have expertise or background knowledge, just expressing that "...Woah, someone busted the glass to the side door of Maybank Hall" (Ivy, 2021).

Note Goldberg's gap - we know someone has asserted that the door is broken, we don't know if the glass is broken. Ivy argues we can assess internal consistency, which I take to mean that we would not so easily believe that the door had been replaced with cheese, and that we have some minimal sensitivity to defeaters, "...the grammar is fine, it doesn't seem intemperate, and doesn't flout norms like using all capitals" (ibid. p. 473). On Ivy's anti-reductionist picture, we are warranted to believe the proposition. I think this assessment is missing critical contextual dependencies. There's no reason to suppose that well-formed language should be construed as a credibility signal without further contextual knowledge. Automated systems using large language models like GPT easily produce this unaided, and can replace or augment human input to such a standard, and in many cases individuals whom we would easily judge to really have epistemic authority deliberately adopt a personal style that varies from grammatical norms. Sometimes idiosyncratic style is a tell that the author is a human. At various times and places online different styles of language usage have been good indicators of credibility. Before

ubiquitous autocorrect, it took effort and knowledge to maintain a formal style. As anti-spam technology has evolved, the kinds of constructs that would evade it changed, likewise, as the requirements of search engine optimization and the economies of scale around content generation have shifted, causing the in-content signals of quality information to likewise change. On the anti-reductionist score - Ivy gives little in the way of argument that speak to the worries about gullibility, nor the impact of broader context, such as knowledge about whether Maybank Hall has a side door, whether the app is known to be deficient at verifying locations, and so forth. Warrant rests on a positive take on the motivation argument - that in general people say things that are true, and that we are generally warranted to accept things until we have reasons to doubt. On the reductionist account, Ivy again urges that the testimonial context is rich enough to furnish information for credibility assessment - which I claim is the central question that needs detailed empirical exploration.

Stepping back a little, the significance of testimony being encountered online and the significance of it being anonymous run together in both articles. For Ivy, online testimony is taken to be heterogeneous in the same ways as offline testimony - one is equally likely to encounter bad motives, bad actors, and such, but they will be as rare as an instance of BYSTANDER where the pedestrian deliberately misleads us, in part because of Ivy's motivation argument, and in part because this conforms with Ivy's observations. Goldberg shifts emphasis to affordance - that there is an epistemic cost "... associated with the privacy or secrecy that the online world affords us."(Goldberg, 2013,

p. 136). Ivy takes issue with the implication that secrecy is a problem for speaker motivation and argues that they are motivated to be accurate no matter the likelihood they can be held personally to account.

However, I think neither gives proper consideration to the extent to which online environments make much more than the speaker's identity secret. We often don't know which people, or algorithms, moderate content, or by which rules. We don't know if we each see the same content, in the same way. We don't know if most of the apparent individuals are humans, or bots, or fake accounts operated by the owners to mimic scale, or by users for some distal purpose. We don't know how reliable whatever interfaces that provide meta-evidence are - maybe "like" counts are faked, or operate with a lag, or are deliberately noisy. Are reviews real, are verification marks authentic, do some posts receive boosted distribution for hidden reasons? We don't know if we have wandered into some ecosystem niche with customs of collusive humour, or deception that we would not otherwise appreciate. Are most speech acts sincere, or do they involve ulterior motives that are not obvious? For example, the social streaming video platform Twitch gamifies engagement metrics and affords very granular segmentation of audiences, such that a streamer's content might be produced with the primary intention of increasing viewership from highly specific audience demographics. The risk that we don't understand the speaker, anonymous or not, tracks our appreciation of the speech context, which in turn depends on the available affordances and signs, their reliability, and our literacy with them. These mundane risks overlap with the more dangerous ones introduced in Chapter

1, where these same opacities create opportunities for hostile interference with our interactions in information systems.

Ivy's optimism thus strikes me as plausible only when relativized to the environment in which testimony is encountered. It's true that, as Ivy says, citing (Kenyon, 2013), that we need rich informational contexts, and that we do our best to construct and represent them and tend to accumulate a kind of folk wisdom (which I discuss as 'evidential grammars' in Chapter 5) about them. That we normally have these supports Ivy's motivation argument - in most circumstances we have a sense of the motivations at play, and most of them are epistemically good or at least benign. But to have grounds stronger than hope, we need to know something about why the other agents we encounter are present, what their interests are, what incentives and selection pressures exist, and so forth. This dispute about the epistemology of anonymous assertions highlights our reliance on context, and questions about how little we can get away with. But this is embedded in a broader problem - that of the context-signalling adequacy of the environment - the extent to which the environment in which we encounter testimony discloses enough of the right kind of information with sufficient reliability to generate justification to believe. This is especially important where the ease of acceptance significantly outweighs the labour required to establish context.

The problem of anonymous assertion and the problem of online assertion run together not because online environments contain so many anonymous assertions, and not because

there are no policing and accountability mechanisms, but because online environments tend to give limited access to meta-evidence, and because they introduce and conceal novel motivations for speech. The information channels we rely on are intermediated by interlocutors who may be pursuing agendas we don't understand or perceive. They may be designed for ends that conflict with epistemic values. This spills over into the real world - increasingly people we interact with are acting with information and motivations they have acquired online. For example, anti-vaccination organizations often deliberately launch legal actions for entirely distal reasons, not expecting success but desiring documents entered into the public record so that these can be re-deployed in online public relations efforts. Our appreciation of the people, motivations, and incentives present in our interactions are increasingly destabilized by digitization because every newly digitized interaction creates scope for unseen intermediation that requires context-savviness we may not be positioned to wield. Unlike BYSTANDER, and JOHNSON ROD, where we have stable appreciation of the context that we can rely on without systematic verification, digital information spaces always afford unexpected intermediations and novel features that they may not disclose in ways that are adequate to our epistemic reliance. We suppose no one is speaking through an earpiece to BYSTANDER, offering them a dollar if they route us past Starbucks. Such a supposition is unwarranted online, and much harder to rule out.

It would be implausible to suppose that every encounter with testimony generates a genuine epistemic dependency to the information environment, especially if the existence

of a dependency generates norms in both directions. The moral I draw from of our dependence on testimony is, firstly, the extent to which we cannot individually generate justification for belief. Our general dependence is mandatory, even if we might think that in some specific circumstances we can verify on our own. On both reductionist and anti-reductionist accounts, we engage dependencies easily. But not every engagement generates norms on the transmission side. Clearly some do - our dependence on instruments generates entitlements attached to the conduct of those involved in their manufacture, as does consultation with domain experts. I have a more general entitlement in an example like BYSTANDER that I am not maliciously mislead, and that the knowledge-norm of assertion is followed. I propose to draw the line at the boundary between reliance and trust - that the corollary of the strength condition for paternalism is a trust condition on dependence. Environmental dependencies emerge when we must bracket significant features of the testimonial context and assume that we understand their functioning and that they function properly - where we shift to Nguyen's unquestioning attitude towards features of the context that have substantial justificatory import.

For instance, even in those cases where we deliberately withhold this trust, where we might think, I don't trust this website, we yet trust that we are actually receiving content from that website. Domain names expire and are maliciously repurposed, bitsquatters exploit rare but reliable hardware errors and register domain names associated with the malformed version of genuine ones, editorial accounts are hacked, and so on. We have a

limited notion of distrust we can practically engage, and we must know more than we typically do to deploy it optimally - most of the time we trust almost all the details of our information intermediaries. Perhaps the ads we see on a webpage form part of our assessment of credibility, but perhaps we are interacting with a web platform that selects advertisements on our own browsing history, not the content we are viewing. That we understand this would be central to the assessments we might make, consciously or not, of the context of the content we encounter within it.

Earlier in section 1 I introduced a distinction between trust in agents and trust in objects and argued that what holds these together as genuine instances of trust is the combination of vulnerable reliance and expectation. In the case of the agent, we hope that they will make decisions in line with our trust, even if there emerge good reasons, even coercion, to do otherwise. In the case of objects, we hope that our decision to depend on them was well-grounded in an accurate appreciation of their properties, and an accurate forecast of how they will behave over time as we use them.

A dependence involves trust in an agent when we must rely on them, where we cannot substitute, and where a breakdown harms us - "trusting can be betrayed, or at least let down, and not just disappointed" (Baier, 1986, p. 265). I am disappointed to learn in GLASS that the door wasn't broken, perhaps I relied on the assertion as evidence for some belief, but it is not plausible that a trust has been broken. Who did I trust? But if I learn that the assertion wasn't made by any person, but was generated automatically,

tailored just to me, because the app was geolocating me and attempting re-route my walk back to campus past a business that was paying to increase foot traffic, I am let down. I trusted the affordances of the environment in which I received the testimony to provide me with contextual information to help me form appropriate credence because I must, in order to use it at all.

Sometimes trust extends to how objects are designed and maintained, and the people involved in these practices. This is an example of "proxy trust" (P. R. Lewis & Marsh, 2022) where one "...may find the artefact more trustworthy because I trust others who were involved in its journey to this current situation..." (ibid. p. 43). These agents are largely anonymous to us, and if asked why we trust them, one answer we might give is that we trust other, further anonymous agents, to have done something about it if they weren't trustworthy.

This concept of proxy trust helps illustrates a distinction along another dimension that can be made, between trust as a first-order relation between agents, and trust in a specific relationship between agents (Bellotti & Moriconi, 2020). We might take no position on some person's trustworthiness as a friend but feel secure trusting their professional work in some domain - engaging in a limited proxy trust to participate in some activity. In the case of testimony, we often adopt proxy trust in intermediaries as parts of sociotechnical systems or assemblages, where the intermediary involves ongoing interactions between human and non-human actors. When we visit Wikipedia, we delegate trust to editors and

authors, and the sources they use, but also to network layers that may or may not be altering content, and so forth. We engage here a variety of proxy trusts to intermediary agents and artifacts.

This proxy intermediary trust is complex, distributed, and difficult both practically and empirically to verify. Returning to Ivy's example, we trusted the sociotechnical system and the representations made about it, and we were let down by those with operative control of those representations and those responsible for the deviant functioning. Twitter's policy changes surrounding a symbol used to denote verified accounts in early 2023 opened a gap that, months later had yet to close, between those operating with a now outdated understanding of how it relates to the testimony they encounter, and those that understand the policy change and adjust the evidential value they attach to its presence accordingly. On learning this, one's first-order credence must be adjusted, regarding items of testimony one may have mis-evaluated, but also a set of higher-order credences must be downregulated as well, since one now possesses evidence that one's heuristics used to evaluate evidence found in the platform may become invalid without warning. One's proxy trust in Twitter as an intermediary in the relationships we have with sources of evidence has been broken. Higher-order dependencies I have to these context-furnishing properties of the environment are not direct to some specific operator or affordance, but instead they are what Goldberg (Goldberg, 2011) calls "diffuse". Goldberg uses the term to handle cases where we find defeaters in properties of the community, like in VILLAGER, where it is some fact about the environment in which

one receives testimony, not the testifier, that defeats justification. We had a diffuse dependence on the sociotechnical systems that produced Twitter's verification mark, and that dependence became unwarranted.

I'll conclude by returning to Lackey's environmental condition and reformulating it in light of the considerations I've raised. Recall, Lackey's condition is that "...the environment in which B receives A's testimony that p is suitable for the reception of reliable testimony". (Lackey, 2003, p. 716) This is supposed to block cases with the structure of VILLAGE and BARN. Both examples are self-consciously implausible - they are things that could happen, but that we don't take to be persistent issues that we actively guard against while going about our epistemic business. But I take this to be an accidental property of stable social and epistemic contexts, one that is actually up for grabs in many digital information systems, many of which have built-in tendencies to produce VILLAGE and BARN cases, which I'll discuss in the next chapters. Our dependence on a suitable environment for testimony is kind of diffuse proxy intermediary trust on the safety-relevant and context furnishing properties of the epistemic environment that justify the beliefs we form based on the information we find within it.

I began this chapter with the goal of articulating a conception of epistemic dependence on the health of the information environment that could help to make sense of the idea I introduced at the end of the second chapter, that we have a kind of public interest in the information environment. Getting clear about the nature of this interest affords a

perspective from which we can critique and improve conceptions of cognitive security premised on the improvement or protection of the information environment. In the next chapter I'll consider the problem of identifying properties of the information environment that underly this dependence and explore the ways we can assess and monitor an environment for their presence. I'll argue that, just as with our ecological interdependencies, this generates presumptive ethical concern for the environments we depend on. These are candidates for public goods that can orient cognitive security policies and strategies.

Chapter 4 - Epistemic and Ecological Community

In the previous chapter I described environmental epistemic dependence and developed a conception of this as a kind of proxy intermediary trust. The object of this trust is the epistemic or information environment - the broader constellation of artifacts and agents that directly and indirectly furnishes and provides context for evidence we encounter, and shapes how we understand these encounters. Because we cannot constantly monitor and verify for all possible failure modes, we non-optionally adopt an attitude of trust toward it in general, even when we withhold it to specific aspects or at specific times. It is the medium of testimony.

This distinguishes it from the trust we might engage with respect to specific sources of testimony and evidence. The built information environment is our interface to what (Levy & Alfano, 2020) call "cumulative culture", the collective knowledge members of a group can practically wield, but which is not the result of the epistemic agency of any one individual. To obtain knowledge from cumulative culture, we must be able to reliably assess the group, members, and practices, in just the kinds of ways I describe at the end of Chapter 3 are required to assess anonymous testimony. If we lose track of who is credible, what is important, which proxy signals have significance, we can't acquire this knowledge even when available. We rely on the stability of the epistemic environmental conditions to maintain the utility and safety of our practices and heuristics.

In this chapter, I want to develop a conceptual framework that affords a shift from the identification of unwanted content in the epistemic environment to the processes we would assess as causally primary, but also as vicious in an epistemic sense. I argue that an ecological metaphor not only offers a perspicuous vocabulary to describe this, but also a useful theoretical and methodological resource. It is no accident that the term "epistemic pollution" has been used to describe phenomena like misinformation, and I propose to take the term seriously.

I extend Leopold's argument (Leopold, 1933, 1943, 1949) that ecological interdependency generates ethical responsibility and argue that epistemic environmental dependence likewise generates a presumptive epistemic ecological value, that properties of epistemic systems that maintain environmental dependencies deserve consideration as a public good. Perhaps this means that the owners and maintainers of digital media systems have positive epistemic duties, but that is a matter for another project. Here, I just wish to posit this as a constraint on cognitive security interventions, as a kind of ecological value that could be pursued that is a bona fide public epistemic good, and thus, an appropriate subject for state-backed protection.

Here is where this chapter fits within the overall argument I am developing. In the first chapter I described the epistemic challenges of information warfare and defense, and in the second I advocated a conception of stewardship that could describe and provide justificatory conditions for the kinds of interventions that some kinds of cognitive

security defense entail. Central to this is the dual requirement that a real public epistemic good be the target of protection or improvement, and that we have grounds to believe the intervention really will be effective. The first part of this requirement is addressed in the third chapter which describes a concept of epistemic dependence on the information environment that I think is a good candidate for just this kind of genuine public epistemic good. This chapter deals with the second and explores what kinds of actions could count as the protection of this environmental dependence, by conceptualizing it as involving the health of the information environment.

4.1 Knowledge at scale

Disagreement between epistemological internalists and externalists, in the view of Dretske (Dretske, 1991), can be understood as manifesting a deeper disagreement between a top-down and a bottom-up conception of what it is to have knowledge. For the externalist, having knowledge is a matter of reliable tracking between beliefs and their objects. My cat's knowledge of where the treats are, or its belief that I am about to enter the house because it heard a garage door, counts as knowledge, even though it doesn't know much about the justificatory strength of its strategies, no more than it knows that the sound of the garage door is the sound of a garage door. In Chapter 3 I discussed Nguyen's conception of trust in artifacts as an unthinking attitude (Nguyen, 2022), which nicely captures the way we often, deliberately or otherwise, ignore a range of possible and esoteric defeaters to the value of such signals. The view is externalist just because all

that is required is the right kinds of connections to the objects of belief, and bottom-up because of its liberalism, allowing that a wide variety of human and non-human animal beliefs count as knowledge.

For the internalist, knowledge requires something extra, not just the delivery of reliable beliefs (however tested and true), and that something is delivered by the cognitive achievement of the agent. Both my cat and I have true beliefs about where the treats are, but my cat lacks the additional ingredient of meta-knowledge and is "... unable to discern the difference between correct information and misinformation or even understand the distinction between truth and deception" (Lehrer 1988, quoted in (Dretske, 1991, p. 23)). Whatever knowledge is, the hallmark is to be found in example of well-secured, justified knowledge, such as scientific practice, which carefully and systematically accounts for knowledge-breaking conditions.

Put this way, the internalist's argument is pessimistic. It takes a lot to have knowledge, frequently we don't have it, dangers lurk everywhere. Dretske counters with optimism, and argues that my cat and I are in fact quite similarly positioned in terms of our access, and lack of access, to suitable meta-knowledge, in order to pressure the top-down account to adopt a more relaxed and plausible standard for what counts as knowledge. We all depend on heuristics, proxies, and other kinds of vulnerable connections to the objects of our beliefs, and the best we can do is monitor them for reliability. There is nothing more that could be delivered to satisfy the internalist. I trust my eyes, my cat trusts the sound of

the garage door, an engineer trusts a gauge (that perhaps turns out to be malfunctioning), we trust what we read in a newspaper. We can bracket some, but not all, of the dependencies we engage. As Dretske sees things, there's nothing inherently special about any of the instruments and externalities we engage, nothing that could identify some subset as "internalist-safe".

I now want to cast doubt on both attitudes. For reasons that should be familiar from the previous chapters, I'm skeptical that any of us can wield adequate, let alone sophisticated, risk mitigation strategies without engaging in further trusts and dependencies that raise the very same challenges. The pessimistic account demands too much. Hannnon (Hannon, 2021) even worries that we can't rely on any of our mental states, raising a skeptical argument grounded in the unreliability of our own cognition, that "... one of the main lessons from the literature on human psychology is that we should not trust ourselves as inquirers" (Hannon, 2021, p. 187). But I'm also skeptical that we should suppose that reliance will naturally be attenuated when trust is undeserved. The kinds of feedback loops, their extendedness in time, and the malleability of our appreciation of success and error conditions, all complicate this picture, especially for social and political domains of discourse of the sort that frequently populates contested information environments. Another way to put the top-down position is that some belief-formation mechanisms are riskier and the bottom-up approach stamps beliefs with the seal of knowledge too cheaply. Thus, a chief virtue of the top-down view is that at least the sophisticated epistemic agent possesses a keen sense of the risks. Dretske

thinks no such sophistication is required, risks realized as failures are converted to mistrust and reduced reliance.

I think there is an even deeper foundation to the disagreement Dretske describes. In the top-down conception, false beliefs die individual deaths in the light of epistemic scrutiny. One imagines putting each item of putative knowledge under the microscope, and passing or failing the belief on the strength of scrutiny of the evidence and degree of safety. Beliefs are like weather, either it is raining, or it is not. On the bottom-up conception, beliefs are like climate. The case is decided more systemically, over time. Borrowing Nugyen's example of a climber's trust in rope, my reliance on this rope might end the moment it frays, but my trust in rope generally is sustained unless I come to appreciate some general problem with its manufacture or my understanding of its properties and suitability to my use. The first time my cat hears the garage door sound, without my appearing and offering treats, is not the end of its reliable belief, even if in fact it ought to be, for some possibly tragic reason unavailable to my cat.

The bottom-up conception is concerned with beliefs, and believers, at scale. It describes much of our everyday epistemic behaviour, which involves the development and management of a repertoire of epistemic strategies, heuristics, and external dependencies. But this adds strength to Leher's worry about misinformation, as the bottom-up view entails expanded and systemic exposure to this risk, since changes in rates of its production and prevalence will upset these strategies. Such changes are, in the

view I develop in this chapter, a central threat to the reliability of our social belief-forming mechanisms.

But in light of the conception of environmental epistemic dependence I developed in Chapter 3, even within the more constrained realm of knowledge of the top-down conception, it's practically implausible that we could detect and filter misinformation, just because there are too many possible kinds and sources, a broad range of unexpected defeaters and faulty signals, like the examples of BARN, or THERMOMETER in Chapter 3. Perhaps in some case of the top-down sort, we really can let our hypotheses die in our stead, but we are sometimes committed to die with our hypotheses. Male jewel beetles locate mates by identifying glossy, dimpled, and brown surfaces, using a perceptual strategy that is easily confused by discarded beer bottles, which nearly caused extinction of the species (Gwynne & Rentz, 1983). One feels a certain forlorn sympathy for this uncorrectable error, like the hedgehog in Derrida's (Derrida, 1991, p. 221) essay "Che cos'è la poesie", which engages its fear response, curling into a ball, sensing danger in the middle of a busy highway, as though, as Schlegel used the same metaphor, it could be "... entirely isolated from the surrounding world and be complete in itself"(Schlegel, 1798, p. 45). The same sort of disconcerting and disconsolate feeling one has upon discovering a friend consumes news exclusively from propaganda channels leaving one without even a starting point for conversation.

Recalling the notion of epistemic safety discussed in Chapter 3, where a belief is safe only if it generated by a process that would yield the belief that P only when P is the case, in such cases a novel environmental condition undermines previously safe engagement of an epistemic strategy. How do we cope with this if we are skeptical that we are individually equipped to manage and monitor on our own? On the bottom-up view, we are with the beetles and hedgehogs, and a great many of us will have to discover the danger, and experiment and adapt, maybe blindly and haphazardly, and maybe unsuccessfully. Success or failure here is a property of groups, communities, societies. It was believed for centuries that lead acetate syrup was the correct way to sweeten and preserve wine, a practice that contributed to the decline of the Roman empire, and continued well into the 17th century (Lessler, 1988) at great human cost.

I want to shift emphasis from individual cases of evidence and justification here and instead consider how we deal with these dangers as social creatures, in social systems, with engineered and ad hoc epistemic processes and practices. Perhaps a social epistemology of this broad conception of misinformation should embrace an ecological metaphor and examine the phenomena as a kind of pollution. This conception is suggested by (Sterelny, 2003, 2006), (Taraborelli, 2008) and mentioned in epistemological contexts by (Carter, 2020), and (Novaes & De Ridder, 2021). Kahan (Kahan, 2017) uses the concept in the context of expert communication environments. More extended epistemological use of the idea can be found in (Ryan, 2018), but especially (Levy, 2018, 2022). Levy engages at length with the idea of misleading claims

and signals of expertise as a kind of epistemic pollution. This erodes the reliability of markers of expertise, creating a cascading loss of environmental reliability. Ryan uses the idea to inform a remediation strategy modelled on the idea of polluter registries. Ryan and Levy's use of the concept in particular suggests that the concept of epistemic pollution can be more than just an evocative phrase to describe undesirable epistemic conditions, but to date the metaphor remains relatively unexplored, which is the task of the next section.

4.2 Ecological and epistemic pollution

Two aspects of the pollution metaphor deserve attention. Firstly, that the metaphysics of pollution involve thresholds and relations, not intrinsic properties. To use the term just to suggest that we should filter, flag, or remove epistemic pollutants risks overlooking this and treating epistemic pollution as self-disclosing. Secondly, the normative core of the concept, where designation of some item or process as polluting entails an implicit claim about the desired state of a system. Here I will examine the pollution metaphor in detail, to show connections between environmental epistemic dependence and the category of characteristically ecological harms that pollution belongs to, so that in the next section I can develop an ethical and evaluative approach to the identification of environmental pollution that is applicable to thinking about responses to epistemic pollution.

Flowering plants have visual and olfactory properties whose primary function is "...to effect sexual reproduction by attracting and manipulating pollinator behavior"(Raguso, 2008). Honeybees in turn depend on this manipulation, relying on odours to locate flowering plants, and have an advanced capacity to distinguish and memorize these odours. Researchers investigating the decline in honeybee populations have raised concerns that pollution from nitrogen oxides in diesel exhaust, even in relatively small quantities, are especially disruptive to the chemical blend of floral odours and significantly impair this detection capacity (Lusebrink et al., 2015). Bees can't find the flowers, and both flower and bee populations decline.

The concept of ecological pollution has a descriptive and an evaluative element. Descriptively, it denotes an abnormal condition with a novel causal genesis where novelty is usually cashed out in terms of human action. One definition in the context of environmental regulation is that pollution is just "the presence of matter or energy in an unusual or unintended place" (Yapp, 1972, p. 77). On the descriptive front, diesel exhaust is a pollutant just because it wouldn't exist if humans had not desired to operate diesel engines. Nutrient pollution that causes algae blooms is pollution because the nutrient concentration would be lower if not for human agricultural activity.

The evaluative component is more complex. Russell (Russell, 1974), and Springer (Springer, 1977) analyze a variety of senses of the term as it is found in discourse around environmental policy and law. These new applications of a term dating back to the 1400s

(the first documented use to refer to pollution of an ecosystem was in 1894), in regulatory contexts demanded that the implicit negative evaluative content of the word be made precise. In one early use of the term in this context, the 1909 Boundary Waters Treaty refers to pollution as a kind of property damage and requires that waters not be "...polluted on either side to the injury of health or property " (Knox, 2008, p. 2). Other expressions attempt to articulate the extent to which a key conceptual ingredient is harm, to human interests, or even to the environment itself, "unfavourable alteration of the marine environment" (NOAA, 1974), the "grave and imminent danger to coastlines" (IMO 1969). The 1961 Geneva Conference of the Council of Europe adopted this definition of water pollution, that "... it is altered in composition or condition, directly or indirectly, as a result of the activities of man, so that it is less suitable for any or all of the purposes for which it would be suitable in its natural state. (cited by (Russell, 1974, p. 177). The 1967 Outer Space Treaty, contemplating the risk of bringing extraterrestrial material back to Earth, defines pollution just as anything that risks causing "adverse changes in the environment" (cited by (Springer, 1977, p. 542).

We cannot get far with the concept in this form without taking a normative stance on what should count as adverse, as a danger, as being less suitable for normal purposes. Writing about the application of the harm principal to pollution control legislation, Feinberg (Feinberg, 1984) probes at the problem of thresholds at which some destructive action harms a common good in a way that warrants coercive state interference. Observing that, "... most of the actions and practices that are thought to be against the

public interest are such that their single occurrence causes little or no public harm" (ibid. p. 26), Feinberg describes pollution as an accumulative harm, and argues that as such, individual accountability mechanisms apply poorly. If other polluters have raised the amount of pollution to just below the harmful threshold, no special punishment is warranted for whomever drives the truck that generates the emissions that finally exceed the threshold. Entertaining the idea that we might instead focus on the most substantive polluter, not the one that crosses the threshold, Feinberg takes pains to show how different responsibility for public accumulative harm is compared to responsibility for individual criminal harms. After all, below the threshold, there was no harm at all. Feinberg concludes that to be wrongful, "... a contribution toward public accumulative harm must be a violation of an authoritative scheme of allocative priorities already in force". (ibid. p. 31) This is to say, we must have arrived, via some mechanism of evidence-informed governance, at a policy that makes explicit what dangers are considered, at what accumulations they are risked, how much risk we will undertake, and how behaviour should be governed accordingly.

In the context of the biosphere, we formulate these by integrating ecological, biological, political, economic, and ethical considerations. For example, the Montreal protocol to protect the ozone layer, adopted in 1987, is widely considered to be a successful intervention to safeguard an ecological public good. Among the reasons for its success is the extent to which is it highly nuanced, targeting supply as well as demand, and implemented in way that is sensitive to local context, and with varying requirements

for more and less developed economies, and respect for the unequal burdens of compliance. It did not just outlaw all chlorofluorocarbon emissions - instead it engaged with the complex of interests involved in production and emission. Motivating this enormous effort was an empirically informed assessment of chlorofluorocarbons as pollutants, including the thresholds at which reactions with the ozone layer threaten ecological reliance on the normal level of ultraviolet radiation that the ozone layer maintains. Intervention did not just focus on supply-side gatekeepers, complex engagement with the casual processes generating demand were built into the protocol. Efforts to protect the ecological health of the infosphere will require that we can generate the same kind of evidence and theoretically grounded assessments of what counts as dangerous pollution, that can in turn generate mitigation strategies that can pass the same kinds of assessments of feasibility, efficacy, fairness, and forecasted consequences, and which can be informed by accurate analysis of supply and demand incentives.

We can discern four dimensions, three of them thresholds, conceptually implicated in the designation of something as a pollutant:

Descriptive - A pollutant is novel, either by property or quantity, to the environment, and this novelty has a cause that is outside of usual conditions of the environment, usually because it was created by human behaviour.

Harm Threshold - At some threshold the pollution has a significant harmful effect. Harm is not a sufficient condition, a localized chemical spill might poison an area of vegetation, but this alone, absent the other conditions, might not warrant useful invocation of the concept of ecological pollution.

Interest Threshold - The harmful effect, at some level of prevalence, threatens a significant greater interest. A small amount of diesel exhaust might harm some bees, but a large amount, over time, threatens bees generally, plants generally, and the connected network of interdependencies.

Accountability Threshold - At some contributory level, it is justifiable to compel changes in human behaviour in order to reduce pollution to a level that will not cause harm to the interest at stake.

In the environmental context, these thresholds and harms are identified by ecologists, biologists, and the like, who work to recognize and react to conditions that harm ecosystems and dependents on them. The honeybee case involves an informational interaction - the presence of nitrogen oxides in the air reduces the amount of information that can be acquired by the normal perceptual and epistemic activity of the honeybee. The threshold at which nitrogen oxide is damaging to plants and honeybees has to do with the perceptual fidelity and discriminatory capacity of the honeybees, and even the extent to which the signal producers, the flowering plants, are able to switch channels, but the

threshold at which we might assess this as harmful to the ecosystem depends on broader dynamics of dependency and resiliency. Returning to Dretske's gloss on the internalism/externalism debate, the top-down conception does not apply to most organisms which can't just shift tactics for the formation of reliable belief when enabling environmental conditions suddenly change. Many generations perish, perhaps a compensatory strategy emerges, or populations migrate, or perhaps the species vanishes, and with it all that depends on it.

The Montreal protocol limited the use of chlorofluorocarbons, even if the policy and its implementation entails harms and costs, because these were outweighed by our substantial interest in the protective function of the ozone layer. To make similar arguments, for instance, to satisfy Feinberg's requirement for an authoritative scheme of allocative priorities that might in turn inform and justify a cognitive security initiative aimed to protect a public interest in safe levels of some epistemic pollutant, we need to articulate the empirical and theoretical basis for the pollution designation. We can do this by identifying normal and abnormal conditions, the socio-epistemic functions of ecosystems in the information environment, substantiable and unsustainable levels of falsehoods and misleading meta-evidential signals, notions of healthy and unhealthy functioning, and criteria that could help to identify important epistemic infrastructures and substantive public interests.

Some phenomena that we might call epistemic pollutants appear to be amenable to identification in this way. For example, deep fake videos undermine an information channel in relation to individual and systemic capacity to track accuracy. The mere existence of video evidence is not sufficient to license credence when there is a high likelihood that any given video is fake. Likelihood emerges not only from the ease of creating a deep fake, it also emerges from the gatekeeping customs of the information channel, and the incentives and culture within it. If deep fakes cross some scale threshold to be properly polluting, it is because they threaten the information channel's function role in the broader information environment.

Here the descriptive element of the pollutant designation is anchored to a veritistic property of the object, that they appear to be representations of states of affairs that are not the case. They mislead about what is the case, not just directly, in terms of content, but also in higher order terms, that the content was recorded, that this is that recording, that this information channel is entirely composed of normal evidence.

The evaluative component is not just that a deep fake video might include some specific false propositions. As Fallis (Fallis, 2021) argues, echoing a concern first articulated by Rini (Rini, 2020), the presence of deep fake videos in the infosphere reduces the amount of information we acquire by watching a video, diluting it by the probability that it does not faithfully represent its object. Worse, once present in the information environment at scale, the sorts of things we might do to counter them fail in

surprising ways. Systemic countermeasures, such as labelling suspected fakes, have been shown to disturb other equilibria and dependencies and to produce counterproductive results, such as the implication that unlabelled content is true (Pennycook, Bear, et al., 2020), or that the presence of such labels counts as meta-evidence that all content within the information system is less trustworthy (Ternovski et al., 2021). More direct attempts to filter and remove epistemic pollutants on veritistic grounds, such as fact-checks, are likewise prone to be counterproductive and to reduce trust because it makes visible editorial intervention (Bachmann & Valenzuela, 2023). Errors in such interventions further risk generating "tainted truth effect" (Freeze et al., 2021), reducing confidence in true and well-justified beliefs.

That we can articulate the normative condition on the pollution diagnostic alone provides scant guidance on how we should formulate it positively and apply to remedial efforts. One might suppose that mere factuality provides just such a basis in this case, that there just should not be deep fake videos in a healthy information system, but this requires clarification of just what sort of system, since in some discourse environments (perhaps devoted to parody or the exploration of technical skill in the production of deep fakes) their presence is interesting, entertaining, and possibly desired. Moreover, veritistic properties will also not suffice to capture all forms of pollution, as many hostile information operations involve the distribution not of false claims, but are instead focused on shifting the context with which evidence is assessed, for example, by flooding discourse environments with content that expresses affective attitudes, in the attempt to

shift the context in which social meaning is constructed, or in which factual claims are assessed, as with voter suppression campaigns based on demoralization. Moreover, some kinds of hostile disinformation are true (as discussed in Chapter 5) and others promote misleading narratives rather than pure falsehoods.

In some cases of fakes, the evaluative condition on the pollution designation can detach even further from veritistic properties. Consider a community that produces and discusses so-called fanfic, fictional episodes of a television series. The sudden introduction of fake text, that is, content produced by generative AI, radically changes information ecosystem, what was once scarce is now plentiful, and a prized skill is now automatic - there is at once too much to discuss, and not much worth discussing at all. The fake is not a pollutant on the grounds that it is untrue, it is a pollutant because it destructively disrupts the web of dependencies and functional roles within the information system.

I'll return to problems with misinformation and disinformation in Chapter 5, for now I just want to show that directives such as "no deep fake videos", or "no false claims of type X", will be poor guiding ideals for pollution remediation efforts of the sort that institutions pursuing cognitive security goals might undertake. The challenge is to connect the descriptive element to the harm and interest thresholds decisively. It's easier to see this in the connection of deep fake videos to the harm of fake news eroding democratic interest in journalistic functioning, because it makes it easier to identify

which aspects of the evidence and meta-evidence in the video are polluting, and in what contexts. As with the fanfic example, this pulls our focus from the pollutant itself towards downstream functional roles and to ecosystem effects.

The same concern arises for anonymous assertions and pseudonymous testimony online. To the extent we might diagnose these as polluting we make a judgment about the epistemic function of identity, but as I argued in Chapter 3, this is relative to the context-signalling adequacy of the environment it is found in. Fake online accounts, purporting to represent distinct real individuals but actually under the control of a single entity, are often used to astroturf narratives to give them the appearance of public support, and the construction of synthetic online identities often attempts to mirror signals of cultural and political membership to trigger affinity biases and further amplify the epistemic effect of apparent peer belief congruence. Analysis for the purposes of target selection for remedial efforts require identification of which elements of the sociotechnical system are properly polluting, and to identify mere pseudonymous speech as polluting will miss that mark, even though this is a common approach, often called for in Western countries, and implemented in authoritarian countries like China and Russia.

In this section I've undertaken some historical and conceptual analysis of the term pollution and identified the normative component that is crucial when we want to use it in the epistemic context. In Chapter 2 I wondered if there are public epistemic goods that we might pursue, even at the cost of some non-epistemic goods (recall that this is Bullock's

test for a bona fide epistemic good). The metaphor of epistemic pollution leads us to one way of uncovering such a good, when we make explicit its implicit normative content, which we can do in relation to our interest in the maintenance of epistemic environmental dependencies. Diagnosis of pollution in this context identifies adverse impacts, harms, dysfunction, and disutility in relation to this. This way of understanding epistemic pollution shifts from an individualist conception of epistemic agency, interests and harms, to one that interfaces better with social-epistemic practices and interests in a social sense, which in turn better fits the kinds of justification and motivations we would articulate for epistemic stewardship.

4.3 Epistemic ecosystems

When wolves were deliberately eliminated from the Kaibab plateau in Arizona in the 1930's, by way of government policies that encouraged their destruction, deer populations soared. One might think removing wolves thus benefitted deer, as well as the ranchers who worried for their livestock. But soon the foliage the deer depended on was depleted, and populations plummeted as the deer starved.

As Leopold (Leopold, 1943; Leopold et al., 1947) describes the first example, we should agree that the number of wolves that is good, from the point of view of deer, is not zero. Leopold marshals this and similar examples in support of an argument that our understanding of the value of ecosystems, and our duties and obligations towards them, is

grounded by interdependencies such as these. Ecologists can make reasonably well-grounded judgements about the appropriate size of the wolf population, and the required habitat protections, to ensure the wellbeing of the broader ecosystem, the deer, the foliage they depend on, and so forth. Assessments of the harmful thresholds of pollutants are can in turn be informed by this knowledge. The problem Leopold observes has since been described by ecologists as a "trophic cascade" (Hairston et al., 1960), in a paper that also argued that resource-consumers played far greater roles in ecological dynamics than was previously thought. The sort of thresholds I discussed in the context of pollution, would, in the case of predator/prey ratios, be determined in relation to the probability of setting off such a cascade. In ecological contexts the discovery of a trophic cascade is often diagnostic of a keystone species (Paine, 1969), which play outsized roles in ecosystem stability. Many cases where human activity negatively affect these at thresholds that trigger cascades turn out not to be quite evident, like the deliberate extirpation of wolves, but indirect and unexpected. When causes are discovered, such as the effect of the insecticide DDT on keystone raptor species, the designation of DDT as a pollutant is grounded in these trophic relations.

Can analogues of ecological relations can be found in the information environment? An early version of this idea can be found in McLuhan's "Laws of Media" (McLuhan & McLuhan, 1999), which describe media technologies as having four dimensions of systemic effect on the information environments they are introduced into. First "retrieval" - a new form of media brings something from a previously obsolete form, such as the

relation between radio and podcasts. Second "reversal" - as a new media technology expands it undermines some of its initial properties, as the decentralisation and openness of the early internet lead to centralization of infrastructure and management,. Thirdly "obsolescence" - as one new form of media is ascendant, another enters a period of decline and displacement, something we have seen with the internet and print media. Finally 'enhancement' - new media changes the kinds of content and behavior we produce, for example, the selfie. McLuhan argued that the mobile phone retrieved the written word as dominant form of communication, and obsoletes single-source propaganda, by affording access to broader and instant access to a plurality of media sources. We might now say, with additional decades of observation, that it also enhanced multi-channel propaganda.

We can see another use of the idea of media as an environment in an earlier set of discussions, distinct from the epistemic ecological model I am developing here, in (Postman, 1970) who first described the idea of "...the study of media as environments" . Postman developed a theory of media ecology that studied the effect of media technologies holistically, as having ecosystem effects, and as affecting the kinds of content that are reproduced, as opposed to a narrower view of media technology as content-neutral tools. However, the field of media ecology, while inspired by the same metaphor I am using here, has a largely different concern, the deterministic and quasi-deterministic relationships between media systems and content and culture. For example, (Scolari, 2012) argues that the two primary applications of the ecological analogy in the

study of media is to explain media history and development as a kind of evolution, and describe communication as a kind of environment, replacing the "...lineal perspective [...] in which the information was an arrow flying from the sender to the receiver" (ibid. p. 207) with a holistic model. One advantage of this latter view is that it can help to explain the failure of theories such as the hypodermic-needle approach to mass communication and propaganda. This supposed that "...cleverly designed stimuli would reach every individual member of the mass society via the media, that each person would perceive it in the same general manner, and that this would provoke a more or less uniform response from all."(DeFleur & Ball-Rokeach, 1989, p. 163). The idea of a media environment helped to explain the complex and indirect effects of mass communication efforts., but here I am looking in a different direction, to the idea that ecological functions and interactions in our epistemic environment help to give content to the idea that some interactions, properties, and conditions in the environment are polluting, unhealthy, ecologically destructive.

Closer to my own conception here is (Janzen et al., 2022), who develop a social-ecological model of disinformation, motivated in part by a concern which I share, that conceptualizations of causes and responses to information disorders have been overly focussed on individual interactions. They argue that a social-ecological model, one that focuses analysis on "...dynamic interactions between individuals and their environments" (ibid. p. 8) is better poised to explain the epistemic behaviour of agents by considering the information ecosystems around them. We do not just interact with propositions, we

interact with them in social contexts that change how the same message is interpreted in different demographics. For instance, they note that during the COVID-19 pandemic, efforts to recruit existing social media influencers were less effective than exposure to authoritative information and conclude "... individual-level approaches to countering information harms are limited in their potential impact and are likely to be more effective when paired with approaches targeting higher-level factors" (ibid. p. 18). Many of these factors are articulated in sociological terms, but I believe our projects to be complementary. The authors note that future exploration of the infrastructural dimensions may be valuable, and this is exactly the direction I am pursuing here - the technological substrate of these social relations, as materialized in epistemic properties of the information environment itself. Who I interact with and the contextual cues that shape my understanding of them is mediated by the sociotechnical systems that produce my information environment, which do not just transparently reveal these social relations but shape and structure them. How to effectively identify and target the higher-level factors they find to be significant, such as authoritative voices (and affordances for monitoring and maintaining the environmental dependencies of authoritative voice), are the sort of social epistemic problem I am specifically interested in.

A distinction made by Goldman between two kinds of speech regulation is helpful here. Content-specific policies target specific types or tokens of content, for example, anti-hate speech prohibitions on holocaust denial. A content-related policy "... mentions no particular content but requires for its application a reference to content"(Goldman,

1999, p. 216), for example, regulation that requires that political advertising disclose itself as such and identify its funders. Implicit in the notion of content-related regulation is that some property beyond the content of the speech is subject to epistemic normative assessment, such its mode of production and distribution. Presciently, Goldman explores this idea in the early days of the internet, that where once "...listeners would know when a speaker had a stake [...] in distorting the facts", now, in an age of "...sophisticated communication technologies and practices, clever speakers hide their stake in the persuasive success of their messages." (ibid. p. 217). Governance of the information ecosystem is thus required to protect its capacity to furnish what Goldman calls veritistic value, it's aptness for the production and discovery of truths. Access to context, including, but not limited to, speaker identity, may help mitigate this sort of worry.

Rather than adopt content-specific and often securitized outcome goals, remedial efforts are on surer footing if they pursue epistemic ecological goals, in two senses, the prevention of collateral damage, and the direct pursuit of ecosystem health rather than simplistic efforts to filter pollutants. I construe Goldman's suggestion as an ecological intervention because it must be implemented across multiple layers and actors within the system, and because its motivations relate to functions and roles that are non-local just to some particular item of content. The adequate labelling of the origin and incentives around some kinds of speech acts is directly related to the type of harm to be mitigated. The ecological metaphor suggests that it should not be surprising if further indirect measures are even more causally important. Consider the manipulation of search engine

results, perhaps to promote epistemic or cognitive security goods. Once discovered, to the extent that the search engine is central to the epistemic ecosystem, this risks a cascading loss of intermediary trust with far-ranging destructive effects on environmental dependency. It can also backfire, as Tripodi (Tripodi, 2022, p. 145) observes, when mainstream media, acting in response to a government request, acted in concert to conceal the name of a whistleblower, this caused a data void (Golebiewski & Boyd, 2019) which made it easier for false information to gain hold in the ecosystem, which in turn became highly salient due to the autocomplete feature of search query input design. In spring of 2022, Google in an effort to reduce the exploitation of data voids, began to label results for searches with unusually recent flux as "rapidly evolving topics" and added warnings to check sources, which in turn has been construed in some communities as meta-evidence for politically-valanced censorship and manipulation (Griswold, 2023). These kinds of effects will vary depending on the nature of the intervention, and the layer of the content moderation stack (Busch, 2022) that is targeted. In the next chapter I will explore at length the application of this analogy to the problem of disinformation and cognitive security.

The first conception of epistemic ecological goals for remedial effort I mentioned above was the prevention of collateral damage. I take it to be reasonable to suppose that some close analogue of the principle of least restrictive means, central to the ethical assessment of autonomy-threatening state interventions in public health ethics, should apply to cognitive security interventions. Of the range of available instruments, one ought

to select the one that will interfere with epistemic autonomy and functioning the least, while still effecting the policy goal. In public health ethics, this principle can have the result in the context of a pandemic of licensing privacy-intruding contact tracing and forced isolation when it might help avoid more restrictive measures such as general lockdowns. To make these sorts of calculations, it must be possible to foresee the effects of the measures. In the epistemic environment, our appreciation of the indirect consequences and cascading effects of interventions is in its infancy. For example, empirical analysis of the results of deplatforming sources of hate speech and disinformation shows mixed results, at times having the desired effect within the targeted platform (H. Innes & Innes, 2023; Thomas & Wahedi, 2023), but, in others, having little effect on the ecosystem-wide production of the undesired content (Ribeiro et al., 2021) and even increasing the salience of damaging properties of the content (Ali et al., 2021; Russo et al., 2023).

The second is that cognitive security goals that are first formulated in narrow, content-specific terms, such as the prevention of anti-vaccination propaganda in social networks during COVID-19, ought to be translated where possible into content-neutral ecological goals, formulated in terms of the direct pursuit of ecosystem wellbeing. For one reason, because this tends to better accord with the conditions of justification on whatever forms of coercion and exercises of power might be necessary to implement the policy. Mere debate, even intemperate and poorly informed, is not usually taken to be a form of dysfunction that demands interference, and thus in contexts where manipulation and

interference motivate the intervention, the intervention should aim for the relevant causes and be justified in relation to them. The conditions that make the environment manipulable will tend to be general, not content-specific, for example, the ease with which concerted activity can be orchestrated to appear to represent genuine consensus.

More importantly, effective measures will tend not to resemble the hypodermic model. The increasing appreciation of the difficulty in successfully enacting precision anti-disinformation campaigns suggests that on purely pragmatic grounds, a strategy promoting general resilience, such as that adopted by Baltic countries in response to Russian and Chinese information operations (Teperik et al., 2022), has better prospects of success. Whole-of-society efforts are, necessarily, whole-of-information-environment efforts, even when there are attempts to shrink and control this along national borders such as China's great firewall.

This leads back to epistemic ecological questions about which conception of health and functioning is implicated when we begin to analyze an epistemic pollutant with mitigation in mind. Perhaps what is at issue is the authenticity of its authorship or intent, or transparency about the mechanisms by which it gains distribution. Functioning signals for various forms of meta-evidence might be weakened by new processes and incentive structures. For example, I recently observed an online discussion about the release of a large language model. Evidence was almost entirely taken from academic preprints, because the review and publication cycle had not caught up to the state of the art. The

discussion devolved into argument about whether a particular paper was even a "real pre-print". Like many academic papers, it was typeset in a common style and form, and it was hosted on a known pre-print server, and the author was credentialed, but discussants worried that it was 'just a blog post formatted to look like an academic paper'. This phenomenon, alongside the growth of predatory and "fake" journals, reduces the amount of justification we can acquire with the same practice and effort prior to these developments.

This kind of change to the speed and volume of information can disrupt reliable heuristics for gathering and assessing evidence. Sometimes, we might find that the overall veritistic properties of the ecosystem remain unchanged while still finding it reasonable to think of it as becoming polluted, just because the techniques we must use to sort and assess become impracticable. This is similar to the way in which the "firehose of falsehood" (Paul and Matthews 2016) propaganda technique reduces the utility of an information environment even if none of falsehoods are ever believed. Flooding an information space with noise that's just close enough to signal that it needs to be evaluated raises the epistemic costs of using the environment, because we must consider and dismiss the high volume of false content.

The goal of this section has been to demonstrate that the ecological metaphor, applied to the information environment, is a useful extension of the epistemic pollution metaphor. Firstly, it helps us to locate the object of the normative evaluation in the diagnosis of

pollution. If fake journals are pollutants, they are because they undermine a reliable process of locating and assessing evidence. Secondly, it affords a methodology to help us look past the pollutant to its causes, because pollution, as an accumulative harm, is problematic as a productive process, not just in its specific emissions. We can then try to disentangle the broader constellation of ecosystem conditions to identify the causal factors, in just the same way we do in the environmental context. Thirdly, it helps, alongside research in this area in the social and individual psychology literature, to explain empirical results that demonstrate to the difficulty of enacting effective remediation. We should not be surprised if mitigation strategies as complex and multifaceted as the Montreal protocol are required to curb some kinds of epistemic pollutants.

This final point has an important implication - whomever designs and enacts remediation will be enacting a wide-ranging normative program, one that includes judgements not just about technical and design features of information systems, but socio-political judgements about harms, costs, and interests, and epistemic judgements about the effects of information. It would be surprising, to say the least, if such decisions, as made within a private platform, cohere with the broader public interest. Moreover, when cognitive security is performed at the margins, fitting into the affordances and opportunities of whatever happens to be the state of nature in the information ecosystem, it likewise must fit substantive normative decisions in whatever space is left by the platforms. In the next section, I'll give a fourth reason, which I think is the strongest, in

favour of the ecological metaphor, that it provides a fertile model for remediation and mitigation.

4.4 Ecological and epistemic interdependencies

An early concern within environmental philosophy was to disentangle instrumentalist conceptions of the environment from those grounded by the idea that the environment is intrinsically valuable. On an instrumentalist conception, diesel exhaust is a pollutant because it is an abnormal change to the environment that harms our interest in clean air and our interests in the presence of flowering plants. We care about the bees only because we care about our food source. Philosophers critical of instrumentalism observed that this is inadequate to describe harms to non-human animals, their ecosystems, and the biosphere itself. One branch of this critique emphasizes intrinsic value of the biosphere itself, entirely apart from human interests and existence. Such an account has limited application to my interest in ecological values of the information environment because this is a built environment whose value is ultimately totally instrumental to human interests. However, ecological systems are so complex that efforts to protect them in a limited manner, tactically oriented only to our own recognized interests, are likely to fail, which can lead us, for instrumental reasons, to adopt pragmatically less-instrumentalist views.

Leopold develops a theory of ecological value grounded in interdependence. "All ethics so far evolved rest upon a single premise: that the individual is a member of a community of interdependent parts...[t]he land ethic simply enlarges the boundaries of the community to include soils, waters, plants, and animals, or collectively: the land." (Leopold, 1949, p. 219). Millstein (Millstein, 2018, 2020) describes Leopold's argument as moving from recognition of the intrinsic value of human community, grounded in interdependence, to the recognition that ecological communities, and their relation to human communities, likewise exhibit this quality of interdependence. Leopold's view is that "...we ought to recognize that our land communities have value and that we ought to accept an ethical theory that benefits and protects them" (Millstein, 2020, p. 12). For Leopold it is important that interdependencies are non-specific, that what grounds ethical claims is interdependence itself, not particular cases, which Millstein reads as "...an appeal to consistency in our ethical beliefs" (ibid ff. 5).

It's worth lingering on the nature of the consistency implicated here. Varner (Varner, 2020) worries that if the price of admission to the ethical realm is mere interdependence, not only might this set the bar too low, but it makes a dubious equivocation between the relations between humans and those within ecosystems. From an ecological point of view, Dretske's bottom-up conception of reliable belief formation is sufficient to describe successful reliance, since stability of the population within its ecosystem is a sufficient condition for its health. But almost all significant theories of the relations between humans in their communities admits at times that individual cases are of irreducible

importance, that on some account of rights, justice, or dignity, some things are wrong, even when good for the community, just because they are wrong for one member. Similarly, our epistemic interests are often focussed on the decisive truth of a specific claim, not whether it's in some way generally true, most of the time. In times of war, for example, getting the details of an event right can be immensely consequential to the political context in which decisions are made. The repercussions of a highly publicized false story can be almost impossible to unwind. Varner argues that these considerations should lead us to recognize a disanalogy. The functioning of biotic communities is of only instrumental value to other members of the ecosystem, whereas a "... well-functioning human political community is constitutive of (or at least actively fosters) human flourishing" (Varner, 2020, p. 7)

However, the nature of our reliance on each other epistemically, by way of our dependency on our shared information environment, is also one where we have a constitutive interdependence between individual and community flourishing. We rely on each other as transmitters of information, as producers of new evidence, and as monitors that are attuned to shifting and idiosyncratic features of our interactions. We encounter each other as individuals in the information environment, but this interface is the product of our collective action. Who we encounter, and in what meta-evidential context, is determined in large part by features of the environment of testimony - this collectively constructed environment is always with us even in direct testimonial interaction. Even if Varner's criticism is right, that there's a disanalogy in a wholesale bootstrapping of the

intrinsic ethical value by way of interdependence, the nature of human epistemic interdependence is exactly of the sort that Varner is yet willing to grant. We are consistent insofar as these kinds of constitutive interdependencies, by blurring the lines of direct dependence with ineliminable diffuse dependence, generates presumptive indirect ethical value on the commons.

Leopold saw the land ethic as a product of social evolution and observes that "[c]onservation is paved with good intentions which prove to be futile, or even dangerous, because they are devoid of critical understanding either of the land or of economic land-use." (Leopold, 1949, p. 225). The advancement of what Leopold calls the ethical frontier in the face of this uncertainty leads to the recognition that self-interest in the biotic community outruns the self and can require sacrifice. This argument speaks to the test for purely epistemic interests for which we might trade non-epistemic interests developed by Bullock (Bullock, 2018b) and discussed in Chapter 2. There is a narrow pragmatic reading of Leopold's argument available here, that advancing appreciation of the subtle complexities of our interdependent relation to biotic communities should inform and temper our management efforts with a recognition that the blunt instruments we could wield might just not work, quite apart from moral objections that might apply. When we yield them carelessly or ineptly, we risk the wellbeing of others, and possibly our own. The greater we understand the complexity of ecological interactions within the biotic community, the more cautious and nuanced our stewardship efforts become, the

more our immediate interests and capacity for action should be tempered by concern for the overall wellbeing of the biotic community.

Therefore, as Leopold sees things, we should expand the ethical sphere to include ecosystems and recognition of their value independent of specific interests we have in them, because of this relationship of constitutive interdependency. This view explicitly entertains the possibility that we often can't determine precisely what these dependencies involve, what health in relation to specific interest looks like. In the last section I distinguished interest and harm thresholds in the concept of pollution. Leopold's conception of interdependence offers a way to connect these, allowing us to articulate interest in relation to harm. Diesel exhaust is a pollutant because (descriptively) it wouldn't be there without our action, and because (evaluatively) at a threshold that has a sufficiently informationally destructive effect on the bee's belief formation mechanisms, it undermines a critical interdependency between honeybees and plants, in a way that would be destructive to both. We assess our own exhaust-emitting activity as polluting, as thus as morally evaluable, because we judge that the harmful effects, amplified through these interdependencies, endanger the biotic community of which we are a part. It happens to be that we can connect this particular environmental harm directly to a human interest, but this is accidental, the fulcrum is the disturbance of ecosystem-scale interdependencies, not individual harms or human interests. In Chapter 5 I'll describe a way to identify similar fulcrums in the information environment based on Wittgenstein's concept of hinge commitments.

I described the descriptive value of the ecological metaphor in epistemic contexts in the previous section, here I turn to its prescriptive value. Epistemic interdependence, for example, as it manifests in epistemic environmental dependencies as I describe these in Chapter 3, likewise motivates and grounds our ethical relationship to the shared epistemic environment. Epistemic engineering and interventions are assessable in terms of their impacts on this. Cognitive security interventions attempting to follow the principle of least restrictive means and minimize epistemic collateral damage must consider collateral damage not just in terms of that to specific individuals, but to the infrastructures that support our epistemic dependencies.

The ecological metaphor's primary value lies here, by leading us to methods of analysis and policymaking that have arisen in the context of protection of the environment. We are unaccustomed to governing intentionality in aggregate, and while we have blunt instruments that can project power and have immediate effects on epistemic properties of the infosphere, they are imprecise, involve significant coercion, hard to control, and difficult to forecast. China's great firewall is a blunt instrument, the effort in the West to control discourse about the origins of COVID-19 a targeted one executed bluntly, both of which erode epistemic environmental dependences by generating reasons to doubt the evidence and testimony we encounter within affected parts of the information environment. Research such as (Busch, 2022; Douek, 2020, 2022) which argues for limits on the exercise of content moderation at different layers or

tiers in the digital information infrastructure should be interpreted in this light, especially when considering state-backed cognitive security initiatives. Intermediating political censorship within cloud-based word processing tools, or filtering political content at the internet backbone layer, are the sorts of interventions that, from the epistemological point of view I've developed here, can be forecasted to undermine environmental dependencies, generating meta-evidence for lowered credence across wide swaths of sources of evidence. From a tactical perspective, this generates a vulnerability to hostile attacks that expose such efforts and frame their presence as grounds for institutional mistrust, and retreat from conventional epistemic authorities to co-partisans. This is a dynamic about which I would like to say more but that will leave as a topic for future work. But briefly, in the presence of a sudden flood of information that is hard to evaluate, we tend not to just reduce the scope and scale of our beliefs, but rather, at times we re-orient them towards authorities, co-partisans, and other narrative centres of gravity that we can navigate and understand. I touch on this in Chapter 5 where I discuss trust-herding, and it is important to recognize that damage to epistemic ecosystem is not always just loss, but sometimes just sudden, opportunistic, and disruptive change.

We can turn to environmental policy to as a model of what thinking ecologically about defensive cognitive security management of the information environment might entail. Leopold can be interpreted as an early champion of a relatively new view, that of "adaptive ecosystem management", an assessment (Varner, 2020) makes, and which Norton (Norton, 2005) concurs in an expansive philosophical account of the approach. As

a model of ecological thinking that is applicable to the information environment, adaptive ecosystem management has three relevant features. Firstly, an epistemic one, that it emphasizes decision making under uncertainty, it is a "...science-based management that assumes we usually do not know enough to choose what is best to do" (ibid. xii), which aptly describes the conditions in which epistemic engineering is conducted and studied, especially in contexts such as anti-misinformation and science-communication where seemingly straightforward approaches have surprising and counterproductive effects.

Secondly, a social feature, that it seeks to intermediate community, policy, and political discourse into management decision-making, instead of viewing this as the province of the special sciences that contribute to our study of ecosystems. This is because we don't know the optimal course of action, and must take risks, make value assessments, and generate costs and even harms. We have to determine how we will understand our dependencies and how we to act on this understanding. These are questions about what matters to us, and not unlike the questions we must ask when we treat some class of information as a pollutant on the basis of its untruth, when we do so we endorse normative claims that are not self-disclosing, but which are themselves up for contestation. Adaptive management is thus well-positioned to continuously interrogate the operative normative elements when some epistemic phenomena appear to afford diagnosis as pollution. Recalling the discussion of epistemic paternalism in Chapter 2, Croce (Croce, 2018) argues that domain experts are not in the best position to make decisions about when we are justified to paternalistically interfere, and how we should

implement epistemic interventions. This is because their interests and duties are not necessarily aligned with the epistemic goods the intervention aims to promote. Instead, we should rely on what he calls "epistemic authorities", who, in addition to domain knowledge, have expertise on the epistemic needs of the subject in the context of the epistemic practice we are protecting. Kahan likewise argues that dysfunction in the communication of scientific expertise arises unless, in addition to the domain experts, "... institutions involved in the creation of policy-relevant science [...] create a capacity for protecting the science communication environment" (Kahan, 2017, p. 431), involving expertise on how science is communicated and received, in addition to the communication from scientists directly.

Thirdly, because of the need to act with incomplete information and in the process of building and maintain political consensus, it tends toward iterative and experimental approaches to management. This resembles a common pattern in the management and development of digital information platforms. These are heavily instrumented to provide analytical capabilities, have their design and functioning tweaked constantly (some platforms deploy software updates dozens of times a day), and often have features rolled out to small samples of users to run experiments (so-called split or A/B testing). The results of these iterations are analyzed in terms of their aggregate effects. This engineering paradigm, with its roots in agile software methodology, is expressed pithily in an internal slogan in the early days of Facebook, "[m]ove fast and break things"(Blodget, 2009). The methodology has been criticized as externalizing costs and

harms in favour of maximizing financial rewards, because it eschews lengthy planning and approval, and encourages decentralized labour processes, shipping features to production without exhaustive review. However, its potential as an explicitly ethical engineering methodology has also been observed (Leijnen et al., 2020; Zuber et al., 2022), and highlights the extent to which the significant matter is the selection and intermediation of stakeholders, not the engineering practice itself, which re-enforces the second feature of adaptative ecosystem management.

In this section, I've argued that the ecological metaphor is valuable because it shows a way of connecting public interests generated by epistemic interdependence to policymaking strategies such as adaptative ecosystem management. The structure of the relationship between ecological interdependence, ethical duties, and policymaking is a model for determining the ethical and policymaking consequents of epistemic interdependence. Here I've just made the case in a general sense, in the next chapter I will put the idea into action in the context of a concrete application.

4.5 Individual and ecological aspects of content moderation

In the ecological context, at times we suppose that there might be moral reasons to protect a species from extinction, even at high costs to human interest. This is grounded by a concern for populations, for ecosystems, not this jewel-beetle, but all jewel-beetles. We legislate against polluting behaviour, usually by targeting polluting processes and

systems of incentives, often directing the outputs of regulative processes towards institutions, organizations, even communities. It is a science and governance of aggregates. Conversely, most of the regulatory thinking about analogous epistemic cases is grounded in concern for individuals. Rights to publish, to speak, to have access to information, are all standardly articulated as attaching to specific persons, and restrictions, such as on hate speech, are grounded in risks of harms to persons.

Putting ecological ecosystem thinking into practice runs against the grain of current governance thinking, especially in the USA. Surveying the legal, political, and technical landscape in which the content moderation in computer networks is governed in the USA, Douek (Douek, 2022) identifies this individual focus as a major limitation. Douek advocates transition from a paradigm of thinking about the governance of content in terms of error correction and individual outcomes, to one based on systems thinking, focusing on ".. choices about design and prioritization in content moderation that set the boundaries within which downstream paradigm cases can occur" (ibid. p. 60). Instead of treating specific contested moderation cases as the object of inquiry, we should treat them as signals to help us diagnose problematic features of platforms and moderation systems. On this view, the development of notice, review, and appeal systems for individual moderation is, at best, an input to help diagnose systemic properties, not a substantive and sufficient response to moderation duties and obligations.

Speech acts are protected as though individual expressive acts each attach to a right to expression, but the problems of speech governance that emerge at the scale of the digital media platforms where content moderation policies are enacted and enforced are fundamentally systemic. It doesn't matter very much to public interests if I have false and foolish views about vaccines, and even if I air them to others. But it matters if such views occur at a scale that can disrupt a public health initiative, a problem during the COVID-19 pandemic that is perhaps the most visible recent example of democratic states grappling with a cognitive security challenge unrelated to election interference. Ordinarily, we would think that if a public view does emerge at this scale, it has prima facie legitimacy. But if we have reason to believe that the apparent or actual consensus was manipulated, if it fails to meet the epistemic norms appropriate to it and does not reflect deliberation in a sense we would reasonably think is appropriate to the content, we face a problem, especially when the view depends on broader skepticism about authoritative sources of testimony we might want to apply to the project of correction.

Content governance is not just " ...the aggregation of many (many!) individual adjudications" (ibid. p. 1) because the decisions, and the dissemination of awareness of these, restructures incentives around behaviour in the infosphere, and thus reshapes the platform's information ecology. Consequently, content moderation policy can appear to have conflicting interests, on the one hand, at times concerned with the moral and legal interest in a particular item of content produced by some individual, but also with an interest in systemic consequences and forecasted systemic effects. Douek suggests that in

large online platforms, the former should give way to the latter, that speech at scale is fundamentally different than speech in offline contexts. We should certainly be open to the idea that quite different regulatory and policy practices bear on the two interests. There can be also functional and perlocutionary differences between speech acts within different sociotechnical systems - as these are engineered to pursue a variety of non-epistemic goals, and produce and distribute speech in idiosyncratic ways.

Douek cites a US Supreme Court decision, that "... the public interest in the purity of its food is so great as to warrant the imposition of the highest standard of care on distributors ... but the constitutional guarantees of the freedom of speech and of the press stand in the way of imposing a similar requirement" (Smith v. California, 361 U.S. 147, 153–4 (1959) The starting point of cognitive security defense is the idea that sometimes, in some contexts, there are some public interests in the information environment and its functioning that in fact do rise to the very same level of public interest. But even where immediate strategic objectives might be clear, for instance, the elimination of some specific class of informational pollutant, the effects of interventions are not, nor are the effective causes easily discernable from proximal ones.

I have argued we can apply Leopold's land ethic to the epistemic environment. Our shared interest in the reliability of environmental epistemic dependencies described in the previous chapter reflects our direct and indirect epistemic interdependence on each other and the forms of epistemic engineering we create, maintain, monitor, and use. Ethical

respect for interdependencies reflects epistemic uncertainty about the consequences of action and vice versa. Not only does effective stewardship require a method to forecast the efficacy of autonomy limiting interference, but it requires that a genuine public good be delivered. Applying the ecological metaphor in the analysis of information disorders and cognitive security challenges is helpful on both fronts. It yields an ethical and empirical framework to help formulate policies, forecast its effects, and assess its ethical justification.

A weakness of this view is the extent to which it depends on the assumption that cognitive security interests will not conflict with epistemic ecological values. Certainly, some regimes actively prioritize security, at great expense to the epistemic autonomy and agency of those who inhabit their information ecosystems. However, this issue can be largely set aside as orthogonal, despite its significance. I'm specifically interested here in the cognitive security defense policies of free and democratic nations, and I suppose that however that category is best drawn, maximization of epistemic wellbeing and autonomy (however that is most perspicuously construed) of citizens is central.

The goal of the next chapter is to take the framework I've developed in the previous chapters and use it to describe a specific type of epistemic ecological value, and then apply this to a concrete cognitive security problem, that of disinformation.

Chapter 5 - The Ethics of Cognitive Security

In the last chapter I described and developed an ecological metaphor for thinking about the information environment, and with that in hand, I will return here to a promise made at the end of the first chapter. There, I introduced cognitive security and connected it to the history of information warfare, and the evolution of information and communication technologies. While accepting a securitized framing as central to the questions I am asking, I claimed that the ecological conception of the ethics of cognitive security is a better standpoint for ethical and policy analysis. This is because it generates less-securitized formulations of the aims and constraints of defensive operations that interface better with our ethical interests in the practice of cognitive security. It helps to fence off questions about the health of the information environment from politically contested questions such as the specific content and viewpoints that should be promoted, or the degree of political polarization that is politically desirable (Melki & Pickering, 2020), or the extent to which "technosocial engineering" that might involve outright manipulation (Frischmann & Selinger, 2018) is acceptable. Most importantly, it offers a grounding for more robust responses, that we should expect are more likely to be effective.

This chapter will make the case for this by applying the ecological metaphor to a specific problem of cognitive security, that of securing democratic states from hostile disinformation. I will argue that filtration and simple pollution models are inadequate to

describe the problem of disinformation, give some reasons to support the tactical and analytical advantages of the ecological approach, and describe a candidate epistemic ecological value that can aid in the ethical and epistemic analysis of potential countermeasures. I'll conclude by sketching a research agenda for future work on the problem of the ethics of cognitive security.

5.1 Defining disinformation

In this section I will defend two claims about disinformation. Firstly, that the concept is more appropriately applied to processes with intentional elements than to items of individual content. Secondly, that the truth of content is not a necessary factor in determining whether it is part of a disinformation campaign. I'll conclude by offering a definition of disinformation that is tightly focussed on the conceptual adequacy of the term in cognitive security contexts.

Disinformation is a canonical example of a cognitive security concern, identified as a threat by countries around the world. In recent years laws have been passed in almost 80 countries to curb misinformation and disinformation, alongside non-legislative policy responses such as task forces, commission, and the production of guidelines and strategies. The 2014 Russian occupation of Crimea, the 2016 US election, the 2016 UK Brexit vote, and the WHO's highly cited declaration that the COVID-19 pandemic was accompanied by an "infodemic" (Zarocostas, 2020) of mis- and dis-information have

been particularly salient inflection points in increased efforts to produce anti-disinformation policies motivated by national security concerns.

Attempts to provide empirically and philosophically informed definitions of disinformation as a category, especially in light of the extent to which true information can be used to intentionally disinform, and disinformation is amplified unintentionally, tend to end up casting a broad net. However, this risks producing a concept that is hard to apply to digital platforms that house vast amounts of heterogenous activity, much of which we would not ordinarily think is governed by particularly strict epistemic norms. It also makes it difficult to pick out the category I am interested in here, that of deliberate hostile disinformation campaigns, which I argue are best understood as intentional, and as systemic phenomena, not properties of particular items of content.

Curiously, in the literature that discusses disinformation in policy and security contexts, the term is often taken to have a straightforward definition. Usually, it's used to refer to a subset of false information presented as true (misinformation) that is produced or propagated with the intent to deceive (Guess & Lyons, 2020; Tucker et al., 2018). Operationalized definitions of disinformation tend to anchor the concept to two central features, falsehood and intention (Fetzer, 2004; Floridi, 2013; Tenove, 2020; Wardle & Derakhshan, 2017a), especially in social sciences, policy, and computing literatures oriented towards the identification and mitigation of misinformation and disinformation. Intention is usually invoked to distinguish disinformation from misinformation, where

both are false, but disinformation is false because it is intended to perform some deceitful function.

However, accounts of intent are complicated by the complex chains of transmission, remixing, and re-transmission that evidence undergoes in online information environments for example, the phenomena of “trading up the chain” (Lewis & Marwick, 2017). Disinformation is developed and refined in niche online spaces of likeminded participants and until it gains distribution in increasingly more mainstream venues, where the content becomes the subject of public attention. Some forms of disinformation take advantage of the phenomena of context collapse, where our evaluation of information is hindered by its dislocation from the context of its production (Frost-Arnold, 2021; Marwick & boyd, 2011), a condition that further complicates accounts of disinformation that tie intention to specific items of content, since the intention in the creation of the content has little to do with the intent involved in its extra-contextual, disinforming transmission and discovery. The additional category of "mal-information" (Wardle & Derakhshan, 2017a) describes true information presented in harmful and deceitful ways.

Simion (Simion, 2023) argues that intent isn't necessary to define disinformation, because examples can be found with no intent to deceive. However, this argument depends on a somewhat unorthodox terminology, where misinformation is strictly a category of false content, and where disinformation is not a subset of it. Firstly, because some disinformation is not false, a claim I agree with and will discuss shortly. Secondly,

because Simion argues that disinformation is not a kind of information at all, because information is factive, and it cannot be coherent for something to be both informative and untrue. Simion develops a notion of disinformation to accommodate this broader argument, and intent as a necessary condition would constrain it too much for this purpose. But the resulting definition, roughly, content that "undermines one's status as a knower" (ibid. p.8) while quite apt for describing what we usually think of as misinformation, is overbroad for thinking about disinformation as a cognitive security concern.

On the face of it, it seems reasonable to think that the history of some representation should have little bearing on how we evaluate it as potential disinformation, that what matters is whether its disinforming here and now. However, if there is one lesson we can draw from the literature that has arisen on the more general problem of determining the semantic content of representations, it is that accounts of meaning focussed on properties of the representation itself tend to require supplementation with information about previous use and intended function in order to deliver plausible verdicts about content. In the context of disinformation, an atomistic view is inadequate to the actual practices involved in its production for at least two reasons Firstly, questions about intent and history are central to understanding why some content is present at some scale in a media system. Recalling the discussion of pollution and thresholds in Chapter 4, that some content is dis/mis-informing is only part of what is important when we are investigating disinforming content that has become salient enough to rise to a level of harm and

potential for harm to motivate cognitive security interest. We want to know what about the system allowed it to flourish, and why at some scale threshold we find it harmful.

Secondly, as I'll describe shortly, where disinformation campaigns once anchored to very specific topics and key pieces of evidence, such as forged diplomatic letters, modern disinformation tactics tend to involve complex mixes of content that span sources and aims, many providing quite indirect contextual support to the adversaries' goal, so the disinforming nature of some particular item of content will depend on other items that may not be disinforming at all on their own, but who's intentional promulgation improves the disinforming qualities of the item in question.

My project here is focused on the category of deliberately hostile and harmful disinformation, so I am content to maintain intent as a necessary condition, to better capture all of the content (including true content) that exists at scale in some system only because of deliberate hostile action. Disinformation in this sense is better understood as a process and practice, not a property of some item of content. Perhaps there are grounds to split hairs somewhat and distinguish between the concept as it applies to describe content itself from the way it applies to behaviour that exploits features of the information environment to produce it. But it would be odd if disinformation campaigns were not defined by the fact that they produce disinformation, and I think the problem here stems from trying to construe disinformation as a concept that applies to individual items of content rather than an activity. The application of content-driven accounts tends to

expand to include most false content, when in some epistemic environments, and for some purposes, there is a high threshold of acceptable falsehood. More importantly, as I'll argue in the next section, the local truth of content is not enough to afford diagnosis of disinformation campaigns that we know are deployed.

The discussion of disinformation in the context of cognitive security often gives central attention to the phenomena of fake news. Fake news conceptually implicates the presentation, of content as news, as diagnostic. Presentation does offer a way to think about intention that isn't committed to whether it should be operant when the content is created, published, or distributed. However, as Habgood-Coote (Habgood-Coote, 2019) argues, the pragmatics of the term have undermined whatever theoretical resources it offered, owing to authoritarian usage of the term to propagandize against legitimate sources of information. As definitional of disinformation in the context of cognitive security, the term is both too broad and too narrow. It is too broad because it equivocates over different senses of "fake", and other distinctions we can make within the category, such as that between economic and political motives.

What about the first factor, that of falsehood? Here we can see why the term is often used in ways that are too narrow. Disinformation does not have to be untrue to perform an intended manipulative purpose. There are several reasons for this. Somewhat esoterically, but important all the same, the idea that online media consumers are encountering discrete propositions with simple truth values is an idealization. Most online

content is multi-modal and requires considerable context-savviness to interpret. Often what we care most about involves contesting the conditions of correctness and correspondence against which representational adequacy is assessed. Assessing content for truth, atomically, and at scale that spans contexts and audiences turns out to be quite difficult. Where human fact-checking has been used as a remedial approach, it is a laborious and contested intervention that is very difficult to implement outside a tiny fragment of content deemed worthy of intervention. As just one example of the implementation challenges, where platforms must recruit moderation assistance from populations with technical, English and native language competence, they often find the result is the selection of people with highly valenced views of local issues.

Content produced in disinformation campaigns often aims for subtle shifts of framing and context to nudge our assessments of correspondence. Pictoral representations can contain content with mixed veritistic properties, for example, a common format for disinforming content is the "evidence collage" (Krafft & Donovan, 2020), composite images that include as parts specific images or short phrases, that, to some audience in possession of the right background knowledge, to imply causal relations. A collage might include a screenshot of a police blotter, a still from a video, a dated photograph, and a quoted comment. All the parts can be authentic and true, but have the perlocutionary effect on a predicted audience to acquire a false belief.

Another example is Russia's use of micro-influencers posting content which, while ostensibly about the Night Wolves motorcycle club, is intended to transmit narrative framings that promote "... affective visions of the Russian world as a core tenet of the Russian culture that seeks to reunite all Russian speakers to strengthen Russia's positions globally" (Boichak, 2023, p. 9). Propaganda is often employed in ways that don't directly question a proposition, but just seek to shift the context of assessment, by connecting evaluation of the target matter to pre-existing disagreement or motivated reasoning. When organized efforts of this sort rise to the level of cognitive security concern, the mere removal of falsehoods-as-pollutants is not a strategy that can be employed effectively, because an accurate identification of the pollutant is not some false claim, but a complex of content that shifts collective representation of evaluative and interpretative conditions. The active creation and amplification of counternarratives is often a more effective strategy in these cases, but requires appropriate political conditions and policy instruments to enable this, since we typically have a fairly limited conception of legitimate state-sponsored persuasive messaging.

Harris (Harris, 2023), like Simion, views the lack of conceptual clarity about disinformation as counterproductive to policy and debate, and produces a positive account similar to Simion's in many respects, but with an important additional factor. Both eschew falsehood of the content as a condition on the designation, and both offer functional accounts. The core of Simon's definition is that disinformation "has a disposition to generate or increase ignorance" (Simion, 2023, p. 8), for Harris, that it

tends to cause audiences to "...form counter-normative beliefs or belief-like states or [to prevent] the audience from holding normative beliefs or belief-like states"(Harris, 2023, p. 11). Both want to zoom out from the veritistic properties of the content itself and attend to its epistemic effect.

For Harris, what is significant in explaining disinforming content is sub-doxastic effect, especially the extent to which it can generates associative states, which don't have the same sorts of epistemic norms as beliefs. Not only does this observation accord with many forms of disinformation found in the wild, it also explains the difficulty in specifying which epistemic shortcomings correctly describe the category. Sub-doxastic content can in turn bring to salience content that we might not otherwise include in our assessment of communication, for example, amplifying the weight we give to the possibility of some hidden defeater, or weighting some elements of meta-evidential context and not others. As I'll describe shortly, some propaganda strategies aim only to raise our fears of such defeaters, undercutting our grounds to form beliefs based on other information in the media system.

5.2 Disinformation and cognitive warfare

There remains important philosophical work to be done to make sense of the concepts of misinformation and disinformation that figure so prominently in a wide range of research and policy-making contexts. Here I am employing a limited sense of the term

disinformation, that I argue here is more suitable to the context of cognitive security. On this account, disinformation:

a) exists in some information system at some scale because of activity with a motivating interest that is only incidentally governed by epistemic norms appropriate to the content (some disinformation campaigns target amplification of content rather than production and thus adopt very indirect means)

b) has productive causes that can be traced to the interests of a political entity (thus most advertising would not count)

c) without regard to negative epistemic and non-epistemic consequences of the activity on the target population (it is fundamentally hostile)

d) for the purposes of changing the beliefs and behaviours of the target population in a way that the producer desires (a strategic purpose underlies the activity)

One of the ways that cognitive warfare can be distinguished from information warfare is along the axis the Harris identifies, between the overt content of communication and its associative effect. Where information warfare involves contesting the flow and content of information channels (Bernal et al., 2020), cognitive warfare is focused on its evaluation, interpretation and effects, explicitly targeting personal and sub-personal processes of interpretation and assessment (Hung & Hung, 2022). In the very small publicly available literature that makes available in English China's strategic and theoretical thinking about

information warfare, cognitive factors play a central role (Aukia, 2021; Beauchamp-Mustafaga, 2019; Cohen et al., 2021; Huang et al., 2023).

Disinformation tactics can target beliefs in one domain by changing the way we think about related beliefs in the background and context of the target one. The goal of cognitive warfare isn't to alter some particular complex of individual beliefs, it is to change the way political aggregates behave. It is epistemically dangerous because it is conducted, and can be conducted successfully, without regard to the effects on our epistemic environment more broadly - it can degrade our epistemic environment. For example, China's information operations in Taiwan during the COVID pandemic, included coordinated efforts to diminish trust in Taiwan's own vaccine, which had collateral effects of general vaccine hesitancy and trust in public health institutions (Yu & Ho, 2023).

We should think of audiences and context not just as meetings of minds, but as technologically mediated. Some audiences are formed in epistemically engineered environments that are amenable to disinformation campaigns at different rates. They may be disinformation-prone because they have structural properties favourable to the creation and sustained existence of disinformation. Simion argues that disinformation can be disinforming even when it does not conflict with existing beliefs, just because it can defeat epistemic support for evidence one has not yet encountered. I view this as the same phenomena (Begby, 2021) calls "evidential pre-emption", where disinforming content

contains, alongside first-order evidence about its propositions, meta-evidential claims that undermine "...credibility of a range of sources that would seek to contradict those propositions" (ibid. p. 8). The structure of a media system may be more or less amenable to this, for example, it might be prone to forming echo chambers (Nguyen, 2020a), it may lack stable or safe meta-evidential affordances, it might incentivize epistemically vicious behaviours in non-obvious ways by way of direct and indirect economic rewards. It might have produced evidence for mistrust by clumsily executed content moderation and cognitive security policies, for example, by implementing content filtering that both discoverable and reasonably disputable. It is possible that some kinds of media systems are, for endogenous reasons, more likely to produce exactly the kinds of content typically produced in disinformation campaigns, absent any deliberate hostile effort that would capture the activity in the definition I offered.

I'm content with this outcome and think it is consistent with the ecological view I have been developing. Intentional disinformation campaigns are just one kind of polluting process that can have this kind of epistemic effect. Where effective cognitive security policies can be enacted to improve the epistemic environment, it would be odd if it was only protective of pollution caused in one way, and not another. For this reason, I think we might be optimistic that cognitive security strategies that aim for ecosystem health can be formulated in ways that do not just advance a political interest.

Harris worries that disinformation defence requires something like an ethics of belief, but for associative states, and expresses concern that this may be a difficult undertaking. I think we can zoom out even further, from the context of belief to the evaluation of harm. Usually, the harm that motivates inquiry into disinformation is cashed out in social, political, or securitized terms. This often leads to an unproductive framing of arguments about remedial policy, contrasting values of free expression, open marketplaces of ideas and the like, with concerns about paternalistic intervention in the interests of safety, or, worse, the preservation of political or economic power. A side effect of this is to re-enforce the veritistic dimension of analysis, where the threshold of "is it demonstrably false", or "was there some specific harm" helps to wall off gray areas of protected expression from moderation policies that aim to reduce harm. But this blinds us to the explicit aims of many disinformation strategies, and the ecological damage to our information environment that both affords these efforts and that also is caused by them, and which underlies the harms we are concerned with. Just as with environmental ethics and policy, recognition of constitutive interdependencies motivates concerns for the protection of shared resources and common goods. This is very different than narrow protection from risks of some particular harm. I'll argue later in this chapter that cognitive security strategies grounded in conceptions of resilience are best understood in the sense of protection environmental epistemic goods, rather than merely the protection of existing authorities. In the next section, I'll describe a specific information ecological value we can identify that is undermined by disinformation, and that can help to orient defensive strategies.

5.3 Vulnerable evidential heuristics

I argued in the previous section that the truth of content is orthogonal to its utility for disinformation campaigns, in part because true content can be used to lead one to acquire seemingly justified false beliefs. Sometimes the disinforming function of false content is not directly related to the matter about which it is false. The "firehose of falsehood" (Paul & Matthews, 2016) propaganda strategy does not aim to advance some particular claim, but to undermine coordination of epistemic agency by inculcating epistemic helplessness (Tenove, 2020) or disorientation (Benkler et al., 2018). In these conditions, people come to believe they are unable to determine the truth of some matter, or to reliably obtain knowledge from some media environment. Rini (Rini, 2021) identifies the mechanism as the "weaponization of skepticism", where the normal and epistemically virtuous process of monitoring for grounds for testimonial skepticism is deliberately activated by generating meta-evidence for elevated of testimonial skepticism. The consequence is reduced access to safe evidence and testimony that can produce knowledge.

From the epistemic ecological perspective, we can see two distinct effects from this propaganda strategy, a pollution effect and an invasive species effect. In Chapter 4 I identified three conditions that are part of the designation of something as polluting. There is a descriptive element, which identifies the sense in which the object is novel or alien in some way. Then two thresholds, a harm threshold, which is the scale of pollution

at which there is a damaging effect, and an interest threshold, where that damage is in turn dangerous to some substantive interest in the functioning of the polluted system. I considered an example of chemicals in diesel exhaust that mimic those produced by flowering plants and relied upon by honeybees to locate them. The designation of this as a pollutant is tied to a harm threshold, of misinformation that can be tolerated, in relation to the perceptual capabilities of the honyebees, and an interest threshold, of ecosystem tolerance for reduced pollination. Above this, there is systemic breakdown of the constitutive interdependencies in the ecosystem.

Changing the scale at which defective testimony is encountered likewise breaks reliable strategies grounded in interdependence for obtaining knowledge from the environment - in this way it is destructive in a similar manner as the diesel exhaust - a way of knowing is no longer reliable. Our normal responses to expressions of disagreement or claims to expertise are confounded by increased likelihood that the signals we are using to identify these are misleading us. It can be hard to formulate the descriptive aspect of the pollution designation in a way the picks out exactly what has gone wrong, but from the point of view I am advocating, we should look to the conditions that enabled the change in scale, and the meta-evidential signals that cause our misperception, not to the content. We want to be able to distinguish a genuine upswell in disagreement with the appearance of one that has been carefully orchestrated using the affordances and vulnerabilities of an information system. This approach is congruent with

some of the conceptualizations large platforms use to describe fraudulent traffic and activity as "inauthentic", or "inorganic".

Reflection on our own experience in media systems we interact with reveals the ways in which different information ecosystems have a different set of reliable heuristics for the evaluation of evidence and testimony, an evidential grammar of sorts, where one learns to attend to various properties of the system to aid in the assessment of evidence. One learns to be savvy about the signals most likely to yield knowledge. Perhaps when certain people endorse content, or when it appears alongside specific user-interface decorations that indicate vetting, or authenticity of transmission, or when content achieves particular numerical scores, one can more reliably depend on it. Faced with information at the massive scale of modern information systems " ... everyone knows that no one can possibly keep abreast of all this information, meta-techniques, meta-strategies, meta-meta-structures, meta-meta-meta-tactics have arisen." (Dennett, 1986, p. 145). These have increasingly indirect contact with the grounds of justification and resemble the ecological webs of interdependence that motivate Leopold's ethical argument, as opposed to a linear and layered scaffolding of dependencies that one might easily reverse-engineer and verify. Every time we encounter an informational label, or a fact-check, or interact with an AI system that demurs a request or makes a recommendation, we are at the tip of an enormous iceberg of human labour, represented in the algorithms and datasets that produce these features in the material infrastructure of media systems,

which we rely on whether we mean to or not, and whether or not we can see inside and understand the functioning.

Exposed to the firehose tactic, because there is more compelling-but-false information to consider, we must resort to new meta-evidential strategies, such as relying on higher-order signals afforded by the media system, such as indicators of social proof, verification markers, follower counts, and the like. Note the vulnerability generated in this transition - a epistemic practice in some environment is undermined, but as a result new dependencies are engaged, or reliance on existing ones hastily modified. When the firehose is deliberately made visible, we gain meta-evidence that the media environment itself is not to be trusted, a generalized defeater for all the information we encounter within it. But we receive that information within it, and thus, rather than eliminate all credence, we adapt instead. When a Russian senior policy advisor admitted to election interference, he added, "... it’s even more serious than that: Russia is meddling in their brains and they don’t know what to do with their changed consciousness" (Isachenkov, 2019). Here, appearing to reveal manipulative effort multiplies the disinforming effect. When we acquire evidence that our own meta-evidential strategies are unsafe, exposed to defeat that we can't easily verify, we scramble to build replacements, and these are often offered to us within a broader disinforming campaign.

When authoritarians urge supporters to distrust traditional media, to rely on their personal information channels such as social media, the demand is not just "trust me", it

is, "trust this way of knowing". Trust-breaking is often a kind of trust-making because our essential epistemic interdependence entails a conservation of trust if we want to continue to participate in an area of discourse. We must deploy it somewhere to gain knowledge from others. In signals intelligence, a tactic called signal herding involves the sabotage of a high-security channel of communication, to cause one's adversary to switch to some other channel that one has the capacity to monitor. We can observe a similar process, that of trust herding, whereby if you can promulgate evidence that in some epistemic environment the normal evidential grammar one learns to reliably use it is subject to some new and difficult-to-monitor defeater, environmental dependencies re-configure along lines that upgrade credence to the source of this new evidence and the ecosystem around it. Alternately, one increases skepticism in the legitimacy of the discourse more broadly. Both responses can be goals of disinformation campaigns.

It takes time to learn a new reliable evidential grammar, creating opportunities for others to exploit newly forming dependencies and evidential heuristics to produce information that has a better chance of being consumed and believed at scale that it would have in in the previous, more settled and reliable state of the information environment. It should not be surprising that epistemic ecosystems polluted along one dimension tend also to polluted along others - where you find higher density of QAnon you likewise find financial scams and health misinformation (La Morgia et al., 2021; Rothschild, 2021). Conspiracy theories unrelated by subject tend to cluster at both the individual and network level (Jost et al., 2018) - and when individuals and groups that

believe in one conspiracy they are likely to believe in others (Enders et al., 2021). Psychologists have explored links between individual psychology and conspiracy belief but recent meta-analysis found no meaningful correlation (Goreis & Voracek, 2019), but preference for digital and non-traditional media has been found to predict the likelihood of conspiracy belief (Humprecht et al., 2020; Vosoughi et al., 2018).

On my account, clustering occurs because the meta-evidential properties of the information system that allow one false conspiracy theory to flourish are hospitable for others, however unrelated in terms of subject matter. They might be united instead by source, for instance particular individuals who gain influence, shared representations of evidential norms and effective heuristics, or by shared malformed epistemic dependencies. Levy (Levy, 2018) notes that in polluted epistemic environments we often find "parallel institutions", designed to mimic legitimate ones but with counter-normative aims, in one example from the paper, the "American College of Pediatricians" set up by a small group of people to promulgate conservative views, easily confused with the legitimate and much more representative body, the American Academy of Pediatrics, and acquiring parasitic credibility from this confusion. Observation of polluted epistemic environments often reveals that these develop into ecosystems of such bodies, including auxiliary institutions such as journals, and connections to other parallel institutions unconnected by topic but connected by the effective underlying motive or shared infrastructures.

Consider what we might agree is a virtuous example that is structurally analogous to a practice we often take to be vicious, that of epistemic trespassing. In an information environment that's noisy, and lacks affordances to enable the identification of domain experts, a kind of virtuous epistemic trespassing can become a reliable means of identifying them. We rely on our assessment of a source of testimony based on matters we know about, and then treat the source as authoritative across other domain on which neither we, nor they, are in a position of authority. Social networks like Twitter invite this at the level of individual accounts, where having learned of the reliability of some account as a source of information on philosophy, I elevate my credence when they vouch for the validity of a source of information on epidemiology, even when neither I nor they possesses a means to assess, because I judge that they wouldn't deceive, that they are sensitive to defeaters, and they are incentivized to be cautious lest their undermine their reputation in their home domain, which I also recognize they have reason to care about.

The reliability of this type of belief formation practice is highly sensitive to savviness with the evidence and context furnishing properties of the information system. Some digital media platforms afford this strategy at the level of specific channels, hashtags, sub-forums, publishers, and the like. Some, like Tik-Tok, de-emphasize many of the standard factors that stabilize context, such as managing the accounts one follows, in favour of a highly algorithmic feeds tuned to proxy signals of interest like videos watched to completion, social graph connections between commenters, and machine-learning

aided preference prediction (see (Rini, forthcoming) for a similar argument to this effect). Strategies used to control illegal and objectionable content production, like de-platforming, can be very successful when the originating account is a significant factor in systemic salience, such as Facebook, Twitter and Youtube (Jhaver et al., 2021). A study of vaccine misinformation on Facebook showed that the bulk of it originated from just 12 accounts (Nogara et al., 2022). But we should anticipate reduced effectiveness of deplatforming in media systems where it is properties of the content and our interaction with it, not its source, nor authority of authorship, that causes it to gain audience. It's much harder to de-platform an idea, an attitude towards evidence, than one influential account. A post-influencer disinformation paradigm will increasingly need to attend to structural properties that favour or disfavour disinformation.

Here I've described the extent to which specific beliefs we might form on the basis of information we encounter depends on heuristics we develop to help us use complex information systems. This in turn creates a dependency between this reliability and the vulnerability of the system itself to be used to operate disinformation campaigns. In the next section, I'll argue that this is a kind of trust relation.

5.4 Complex epistemic dependence and trust

In chapter 3 I discussed a functionalist approach to trust (Lewis & Marsh, 2022). In the paper, they ask a series of questions about what it's like to trust things like animals,

computers, and artificially intelligent systems. In one of their examples, when we trust a car, we engage in proxy trust, whereby we trust in a car's roadworthiness because we trust a roadworthiness test that it passed, and in turn, the test itself, those that carried it out, their instruments, and so forth. They briefly distinguish proxy trust from the concept of transitive trust, where if I trust some system A, and that system in turn trust some other system B, I have a transitive trust relationship with system B, mediated by social information I can monitor for defeating conditions, and therefore it is correct to say I trust system B. In proxy trust, I do not trust system B, I project my trust of system A on to system B. This distinction comes quickly, and the authors don't make much of it, but I think one reason it's worth making is the impossibility of transitivity in many epistemic contexts. Earlier in the paper they describe trust as accepted dependency, grounded by evidence of trustworthiness. But in epistemic contexts we generally lack available evidence and competence and resources to recognize and accept dependencies. We don't engage in transitive trust because we can't consider all the dependencies involved, and thus can't assess them. As I argued in Chapter 3, this is the case with our epistemic environmental dependencies, and I likened these to a kind of proxy trust in the intermediating sociotechnical systems that stand between us and sources of evidence and testimony.

In the last section I described the ways in which specific beliefs are vulnerable to the reliability of the evidential heuristics we use. Here I offer I more specific way to think about this, as a kind of proxy trust on an intermediary. To illustrate this, consider the

projection of our trust in our understanding of how a search engine works onto the evidential weight we give to its results. What is it like to trust a search engine? We begin to type into a search input and receive suggestions. Where did they come from? Am I nudged to accommodate my intent to what's immediately present? Manipulating autocomplete suggestions is a well-recognized influence technique (Tripodi, 2022), and search providers themselves operate human and machine subsystems to try to prevent this (Olteanu et al., 2020).

A lively arms race lurks behind this seemingly simple feature. We might know enough about this to see poor suggestions as a problem partially external to the trust we place in the search engine, but we might also re-assess our proxy trust just on the face of suggestions we deem to be of low quality. What seem to be mere interface features, things like suggestions and rankings, can have substantial effects on the beliefs we form. What about the more central systems of search? Once the internet, and search engines, became central to a broad range of cognitive projects natively, not just as convenient stand-ins for offline behaviours, their trustworthiness becomes a broad concern. When we use them, we trust our mental model of their mechanisms, that it retrieves items from a set that match our request, for instance, that they contain a particular word. The conclusions we rationally draw from the results are altered if we come to believe the system is hiding something from us. We trust that we understand the nature and composition of that set - that it is the same for all users, that items aren't arbitrarily included or excluded, that it does not contain gaps and omissions that would surprise us.

We trust that it hasn't excluded some subset of potential search space without it being knowable to us that it has done so. Does it exclude publishers that haven't paid the search engine? Content of a certain age? Content that is unpopular? Content that we have given previous evidence that we don't like, that might offend us, that we won't look at? Content that is damaging to the interests of those with operative control of the search engine?

The manipulation of search engine salience and ranking produces substantial belief and behavioural difference in the context of elections (R. Epstein & Robertson, 2015). Tripodi (Tripodi, 2022) describes at length the hostile manipulation of the affordances of search engines by American political organizations to influence public discourse, by causing them to produce desired results for particular queries, and then encouraging audiences to perform these searches, in order that one would then find this pre-ordained evidence (Tripodi, 2022). The nature of this influence tactic subverts the reliability of standardly virtuous epistemic practices. The practitioner can tout bona fide epistemic virtues "don't believe me, research this yourself", knowing that the research prompt will yield some of the planted results.

That one's search yielded no results, or some specific result, is standardly taken as evidence about the query and the content, but the reliability of this is dependent on the extent to which the proxy trust on the search engine is not misplaced. It's well documented that domestic operators of search and media platforms in authoritarian states often interfere with the performance of these systems for political or economic reasons.

Corporations that operate search products, such as Google, have adopted a range of policies to afford censorship to maintain access to authoritarian markets, such as China, including offering government access to monitor and filter, accepting lists of keywords to block, and refraining from indicating to users that their search results have been censored (something Google did in China until 2012). Democracies likewise regularly request that search engines remove content from their results (Meserve & Pemstein, 2018), and in some cases search engines add interface decorations to results to notify users that content has been redacted. Some operators of search engines and social platforms sometimes produce so-called transparency reports, documenting government requests for information, redaction, and the like, and we trust that these are complete and accurate enough so that domain experts who monitor such things are positioned to recognize conditions that might undermine the function we expect search engines to perform for us. Yet another proxy trust is implicated here - we project our trust that actors with the power to damage our reliance on search engines are not doing so onto the systems themselves, against the evidence we possess that they can, and might at times be incentivised to.

I've used this example of trust in a search engine as source of evidence to illustrate how complex the dependences engaged in proxy intermediary trust can be, and the extent to which the involve a surprisingly wide range of interests and actors, even in this brief sketch of an analysis. In the next section, I propose a way of describing the structure of this trust, in a way that can help us to assess the extent and significance of damage to it.

5.5 Hinge stability and epistemic ecosystem damage

I've argued that different kinds of media systems will have different kinds of reliable evidential strategies, and to wield these we must engage complex forms of proxy trust to intermediaries they supervene on. If this is right, there should be discernable structure and shape to this. In an ecosystem, the loss of a species or habitat leads to systemic changes such as trophic cascades which can cause complex and substantial distal effects. We can predict with some certainty the likely ecosystem consequences of different kinds of damage, for instance, by learning to recognize keystone species. We know roughly what would happen if all bees, or mosquitos, or wolves, were to vanish, what will happen to a local wetland in the presence of an invasive species, or what the consequences of an algae bloom will be. This kind of reliable ecological knowledge is an enabling condition for adaptive ecosystem management, we measure our interventions in relation to our abilities to forecast and monitor their effects.

We are in the very early days of having a similar understanding of the information environment. The extent of backfire effects when fact-checking (Swire-Thompson et al., 2020), where corrections increase, rather than decrease, belief in the false proposition, or implied truth effects (Pennycook, Bear, et al., 2020), where the presence of warning labels on some content cause people to believe unlabelled content is reliable, are unclear. They are likely to be relational to ecosystem factors such as the audiences and material

affordances. Systemic questions about whether recommender algorithms polarize, or whether various forms of deplatforming are effective, are also unclear. For instance, deplatforming often results in re-platforming in a different environment, with partial migration to extreme communities, and overall increases influence of the content that motivated censure (Horta Ribeiro et al., 2023; R. Rogers, 2020).

Here I want to explore a way of thinking about the structure of epistemic environmental trust inspired by Wittgenstein's concept of hinge beliefs. This idea features prominently in "On Certainty"(Wittgenstein et al., 1969), where Wittgenstein rejects the idea that we can ever provide fully general and rational account of our beliefs, where we can demonstrate the correctness of one belief by way of the scaffolding on which it sits, which in turn we can show to be sound. Wittgenstein writes of hinges as propositions that must be exempted from doubt, such as that there is an external world, but not because we have reason to be sure of them, but, rather, that their loss comes at exorbitant costs. "We just can't investigate everything, and for that reason we are forced to rest content with assumption. If I want the door to turn, the hinges must stay put." (ibid. §343). To discover that a hinge belief is false is catastrophic - it would be cognitively disastrous if we lack a "non-evidential right to trust [...] hinge propositions, such as that sense perception is reliable, induction is reliable, that the world isn’t a simulation created by nefarious aliens, and so forth." (Ranalli, 2020, p. 4976).

We might ask, are these hinge commitments propositions, are they beliefs, do we have reasons to accept them, and if so, what kinds of reasons? Wittgenstein did not live to finish "On Certainty", which was published posthumously from drafts and notes, but subsequent scholarship has explored such questions at length against the background of contemporary epistemology. Here I want to focus on the idea that we have some beliefs that we must take to be true because the high epistemic cost to doubting them cannot practically be entertained. On the way to knowing about something, we have no choice but to eschew skepticism about our hinge commitments. For Wittgenstein, hinge commitments cannot be rationally evaluated, and to doubt them would "drag everything with it and plunge it into chaos." (§613). Because Wittgenstein's discussion of hinges is focussed on skeptical problems, of the sort, how do we know there is an external world, only a few very general ones are considered, and this way of thinking about hinges is retained in some of the subsequent scholarship.

However, I am not alone in finding the concept to be useful when it admits more hinges, and more specific ones that can in fact be rationally evaluated. For instance, Coliva argues that to take evidence from the fossil record as proof for the age of the earth, one must suppose that the earth has existed for a long time, and that this is a hinge belief on which knowledge of geology depends (Coliva, 2015). Likewise, Ranelli argues that disagreement about hinge beliefs lie at the root of deep disagreement, such as can arise in the context of religious belief, political convictions, and conspiracy theories. Ranelli gives the example that in a community that believes in spirits, I must accept belief in

them in order that I might participate in many of my communities' cognitive projects - it is the price of admission to whole domains of inquiry and social life. This is the first aspect of the hinge that is important to my interest here - that they mark sites of vulnerability. If for some reasons we should come to believe we have been mistaken about one, it can have substantial consequence on what we take ourselves to know. In Ranelli's example, some of the person's beliefs are safe even if they come to disbelieve in spirits, but an entire category of important cognitive projects become impossible. This way of thinking about hinge commitments expands their scope to be more akin to keystone species, as fulcrums on which important stabilities depend. We do not necessarily suppose that there are just a few, and that total annihilation of all ecosystems will result from their loss, but we know that they are critical nodes in webs of interdependencies.

Ranelli's account of hinges, and the term "cognitive projects" is based in part on Wright's entitlement account (Wright, 2014), which I take on board here as well. Wright conceives of hinge commitments as the general conditions we must take to be in place when we when we endorse a particular procedure (Wright calls this an authenticity condition) as an acceptable way to answer a kind of question (Wright calls this is a cognitive project). Not all authenticity conditions are hinges, this instead arises when authenticity conditions are general and apply to many different cognitive projects. There are important questions about how to daw this line, and Ranelli's account admits more

hinges than Wright's, but these questions are beyond the present scope, where I want to use the idea of hinges to identify structures of dependence to our epistemic environments.

Hinges involve often unarticulated background beliefs we must take to be settled if we are to undertake cognitive projects. A non-hinge authenticity condition is one where "... for a given cognitive project [...] any condition doubt about which would rationally require doubt about the efficacy of the proposed method of executing the project"(Wright, 212 C.E., p. 466). Hinge conditions are just broader, common to more cognitive projects, where doubt "...about any investigation that uses some relevant apparatus or relies upon on a certain kind of evidence, or a doubt about the good standing of all previous investigations of a certain kind, or about the very subject matter of a large class of investigations, or about the propriety of their methods" (Wright, 2014, p. 216)

In Chapter 3 I described environmental epistemic dependencies as background and enabling dependencies we must undertake in order engage more specific epistemic dependencies, on particular sources of evidence, or on particular testimonies. I depend on an expert, but I also depend on the fidelity of my interface to them, on the reliability of the signals of expertise I monitor, on the contextual features I expect to warn me of potential defeaters and to offer falsifying evidence. Our reliance on environmental dependencies spans cognitive projects in just the sort of way that Wright conceives of these general, hinge authenticity conditions. I am entitled to believe that when I request a web page from Wikipedia through my web browser, that the one I receive and that my

browser displays to me is faithful to the intentions of the authors and maintainers of Wikipedia, that the broad complex of intermediary layers, processes, and practices towards which I must adopt an unquestioning attitude remain deserving of such, and are not tampered with in some hidden way, despite the fact I possess knowledge of the many ways in which this is possible and feasible. On Wright's view, my entitlement to make these assumptions does not flow from direct knowledge and evidence I possess. Rather, it flows from the impossibility of engaging in normal cognitive projects without the entitlement.

This entitlement account of hinges helps us to describe structures of epistemic environmental trust. The identification of hinges is the identification of features of the epistemic environment that we have significant shared dependencies on, where, in Wittgenstein's metaphor, these must stand still for the door of inquiry to move. When there is a broad range of cognitive projects that a community views as legitimate, that have a base of shared authenticity conditions that depend on common features of the epistemic environment, we can identify these features as central to epistemic ecosystem stability.

This offers a more perspicuous way to describe disinformation campaigns with indirect mechanisms, where the disinforming function varies from the content which serves as the vehicle for the campaign. By manipulating hinge commitments or the evidence we have to assess them, credence in dependent content can be lowered. The

result of disinformation is not mere reduction in available reliable evidence. Damage cascades from channels of evidence, second-order evidence, and meta-evidence, causing re-configuration of dependencies and changes in operative norms and evidential strategies. Rapid change brings new risks to our shared dependencies, and our ability to reliably form true beliefs from the information environment around us. Disinformation operations that seek to prepare the ground for future targeted campaigns often attempt to scaffold alternate hinges and seed doubt about existing ones. If an apparatus of narratives about what kinds of evidence are good, what signals of trust are reliable, what sources are legitimate, and similar meta-evidential evidential beliefs, has been developed and gained traction, it is easier to circulate a specific strategic falsehood at scale when some future opportunity arises, because many cues that might otherwise have inhibited credence and transmission have already been made ineffective. Disagreement about the particular content then often can become deep and irreconcilable because it includes this broader apparatus - what I'm imagining here as a hinge.

Just as we have come to recognize that we have public interests in private property and agency when it risks damage to ecosystems on which we have dependencies, we can likewise recognize public epistemic interests within privately owned information systems. These are distinct from the public interests that already are represented in regulatory constraints on the management of such platforms, which are largely aimed at reducing specific harms such as fraud and harassment, and not especially effective. Recognition of these interests can generate constraints on the kinds of cognitive security

policies a state might pursue, by requiring that they not endanger these, and assessing probably collateral damage in relation to them.

The idea that hinge commitments help us identify public epistemic goods that may deserve protection allows us to move from the mere identification of instances of specific engagements of environmental trust to the identification of social-epistemically significant epistemic systems. This offers a basis for claims about the ecological value of the implicated environmental properties. This argument and its methods demonstrate how we might inform cognitive security policy with epistemic ecological reasoning. Given some range of options to conceptualize threats and execute a defensive policy, we should prefer the one that does the least damage to operant hinge stabilities. Rather than target the specific implementation instruments of a disinformation campaign, we might prefer to improve the resilience of the information environment to the campaign. More importantly, in this light, there are stronger reasons to table when considering actions that involve state-sponsored persuasive messaging, something we might have a presumption against that and think requires elevated justification. In the next section, I'll apply the concepts of shallow and deep ecology to describe the difference in these approaches.

5.6 Shallow and deep conceptions of cognitive security

Towards the end of Section 4 I argued that our trust in search engines includes trust in the extent to which those who can manipulate it in overt ways for the purposes of

disinforming us are kept in check, or that at least we have a good picture of what kinds of manipulation to be on guard for. We often delegate this to experts who monitor such things and implement it counterfactually, supposing that if there was something important to our interests either we'd somehow be informed, or the system would be changed. This is not unlike way that cybersecurity is enacted. It's a problem we know exists, we know experts are working on it, and they modify our systems, or request that we modify our behaviour, and we tend to go along with this unquestioningly. We do this not only because we might lack the expertise to delve much deeper, but because practically there is too much to know, and at a level of detail that spans many experts. We might have a sense of what cybersecurity is, and why it matters, and that an entire industry and associated complexes of regulators, standards, and laws are mobilized in its service, but for most of us we don't notice until the dependence breaks and we face directly insecurity and its consequences. Cybersecurity is considerably more mature as a practice than cognitive security, and while cognitive security is often theorized as including cybersecurity as a concern, it operates at a level of abstraction one step removed from just the authorization of access to epistemic infrastructure, to the legitimacy of interactions with it.

There is a large literature documenting state efforts to manipulate information systems we depend on in ways that risk our epistemic dependence on them. This includes indirect tactics that don't directly interfere with the operators, such as manipulating the network layer to filter undesirable content (Griffiths, 2021), and the active production of content

(King et al., 2017) to advance favored views, or interfere with reliable functioning. The manipulation of search engines via content production exemplifies disinforming activity where it's very difficult to provide an analysis of the content alone that reveals it as disinformation. We need to analyze it in terms of its function and causes. During the COVID-19 pandemic, accounts linked to China clogged information spaces with repetitive posts containing names of major cities to prevent more specific queries from working correctly (Milmo & Davidson, 2022). A US Centre for Naval Analysis report (McBride et al., 2020) taxonomizes three variants of this sort of disinforming content production, flooding, to make a information channel unusable, hijacking, to co-opt an information channel by changing the topic, and fracturing, by using mimicry to outcompete the keywords that help people find the channel. These manipulations can rise to the level of genuine national security concerns, for example, when they influence elections or health behaviour.

Our understanding of the precise effects of these campaigns, such as we know of them, is poor. Despite agreement on the threat of disinformation, "...policymakers often appear incapable of articulating what security means in this context" (Ördén, 2022, p. 1). A recent review study notes "...our review of prior literature reveals that the extant studies examining the impacts of disinformation are mostly descriptive and atheoretical." (Arayankalam & Krishnan, 2021). Relatively little is known of the effectiveness of anti-disinformation efforts, such as they have been publicized.

One of the aims of the area of research I am advocating here, the ethics of cognitive security, is to motivate empirical study, but also to fill in this theoretical gap. The ecological approach I describe is an example of such an effort. The literature on information warfare often adopts an explicitly ecological view in characterizing the theatre of operations, in a recent review (Ronfeldt & Arquilla, 2020) this is described as the "global commons" or "noosphere", and both offensive and defensive actions ("noöpolitik") are viewed as efforts to alter it systemically, not merely to achieve particular objectives. Ronfeldt and Arquilla observe that hostile action has been effective as fragmenting the information environment and advocate the protection of the global commons as a strategic goal. They lament that "[n]o methodology exists for assessing the status of the noosphere from strategic standpoints" (Ronfeldt & Arquilla, 2020, p. 461), and while they gesture towards an environmental conception, they do not offer a methodology for analyzing it ecologically. I've offered an account of epistemic ecological health and argued that the ecological metaphor also shows a way of motivating ethical concern for our epistemic commons, in just the same way it does for our environmental commons.

We can turn again to environmental philosophy to help conceptualize two different kinds of cognitive security thinking, applying the distinction between shallow and deep ecology from Naess (Naess, 1973). For Naess, the shallow ecological view is concerned with the management of anthropogenic environmental change within limits of human tolerance for pollution and resource depletion. Shallow approaches minimize disruption

of our behaviours and economies and aim to manage and moderate externalized effects. The deep view begins with what I read as an epistemic argument - that because of the complexity of interrelations, we operate with substantial uncertainties about the effects of interferences and damage with ecological processes. Naess argues that the shallow view fails to appreciate the extent to which ecology is a limited science, and the extent of "human ignorance of biospheric relationships" (Naess, 1973, p. 97).

The constitutive interdependences we have to environmental health entail that we cannot protect specific dependencies without protecting dependence more generally. Evaluating the extent to which these claims are true reveals the extent to which this requires selection of the relevant spatial and temporal horizon, and the shallow view take the most immediate human interests as markers of this boundary, whereas the deep view rejects imagined discontinuity between human interest and biospheric interests. Likewise, we cannot easily understand and manipulate damaging human behaviour just by targeting that behaviour, deeper interrogation of the causes is required.

In this light, the shallow approach suffers from the same problem that I identified in the discussion of pollution in Chapter 4 - that our conceptions of pollution and damage contain within them implicit endorsement of some conception of health. The shallow view takes the current arrangements of human culture, economics, and politics as the norm from which environmental damage risks deviance. The deep view urges engagement with these underlying normative questions as central to any effective

response to environmental destruction, and argue that the appearance of effective, minimal, highly targeted policy options is illusory.

Deep ecology also includes within its theoretical apparatus metaphysical and moral arguments about the relationship of humans to the biosphere that is largely not transferrable to my interests here. However, I take the epistemic argument to be severable from these. Alone, it is quite congruent with the Leopoldian view of interdependence that I described in Chapter 4, where the fact of complex interdependence explains the empirical inadequacy of shallow analysis, and justifies ethical concern for the biotic community of which we are an irreducible part. I take the central distinction to be between approaches that take the current arrangements as a given, and thus as a baseline of health, and seek to fix isolated problems that flare up, with those that view the current arrangements as inseparable from the problems we describe, and thus, that openness to re-arrangement of these must be central to remedial efforts. Sessions (Sessions, 1995) argues that shallow approaches in the environmental context tend to be technocratic and anthropocentric, in the sense that they value ecosystems only as they provide value to humans.

Concerns about anthropocentrism and misanthropy have received heavy billing in criticism of deep ecology, with critics making statements such as "I happen to think humans are more important the whooping cranes" (Commoner, quoted in (Fox & Devall, 1981, p. 306)). However, the upshot of epistemic argument is that in some cases, what is

important for whooping cranes is vital for humans, and that we don't often know exactly how entangled our fates might be in the face of some particular threat. What's good for the whooping cranes is frequently what's good for us. Likewise, one might think that our freedom to connect some computers to the internet, put our own software on them, and let people use them, say, to share news, is more important than the risk that some reader might be misled. But it might not be, if, just as with the cranes, we cannot so easily articulate the conditions of my epistemic liberty and wellbeing without bringing into scope that of those around us. We are accustomed to thinking of epistemic agency individually, relating to agential capacities to access and assess evidence, to produce testimony, to speak and publish, and so forth. But it extends to coordinated epistemic engineering and the production and maintenance of the kinds of intermediating technologies I described in Chapter 4 are often implicated in environmental dependence. We might not trade my liberty to write and publish a pamphlet for some perceived increase in safety, but we might constrain my liberty to interfere en masse with the printing presses of many.

The extent to which the implementation of such a policy should be assessed as expressing a shallow or deep approach depends on the considerations that inform the policy selection and scope. In Chapter 4, I cited approvingly the association of the adaptative ecosystem management approach to ecological preservation and restoration with Leopold's land ethic, especially because this embraces uncertainty as to the appropriate types of intervention and our ability to predict effects. The approach

recommends that we make careful and tentative stewardship efforts in tight iterative cycles of monitoring and evaluation. An embrace of this sort of epistemic humility might seem directly at odds with deep approaches that are explicitly open to far-reaching structural interference with the status quo[1]. However, to give content to the idea that we should act with minimally viable methods demands that we articulate how we understand the operant notion of minimal. A policy with a deliberately minimal impact on the rights and obligations of platform owners, or which aims to narrowly eliminate a particular type of content, might have an outsized impact on users, generating substantial changes to networks of trust and ecological relationships between content creators, curators, and consumers. As in the environmental context, the deep approach forces us to articulate and defend the implicit commitments of our policy responses, and as a mode of critique, often illuminates the extent to which these commitments are not consistent with the claimed motivating interests.

If we wish to protect some forest ecosystem but endorse minimal interference with the directly involved economic actors as a prima facie constraint on policy, the approach is modest and conservative only with respect to minimizing interference with these actors and the conditions that enabled the concentration of power in their hands. It is not modest and restrained from the point of view of the forest. We can contrast this with stewardship efforts focussed directly on restoration of the ecosystem, which might begin with substantial, even overwhelming regulatory impositions on the incentives motivating

[1] I am indebted to Neil Levy for pointing out this contradiction and its significance

destructive behaviour, and then conducting careful and conservative rehabilitation efforts in the forest itself. In the environmental context, deep approaches tend to highlight root cause analysis, and to propose remedial actions that target these, and which reach deeply into the status quo, often recommending "... changes in policies [which] affect basic economic, technological, and ideological structures" (Naess, 1989)

My interest in the distinction is to highlight the extent to which the dominant forces that shape the current information environment have arisen rapidly, in pursuit of ends that are at best only loosely allied with our collective epistemic interests and dependencies, and which have disrupted social epistemic practices with much longer histories. An approach to information ecosystem stewardship that calibrates a conception of conservative and careful management to the extent to which these new platforms are altered and affected is only cautious in a narrow pragmatic sense and might have outsized and far-ranging effects on the information environment itself and entail the reproduction and sedimentation of conditions that are fundamentally at odds with its ecological health. Cautious stewardship based on identification of shared public interests, such as analysis of hinge commitments might reveal, will not necessarily appear cautious from the point of view of the property rights of those with operative control of these resources. I see the value of adopting a deep approach primarily in making visible what otherwise is pushed out of frame, encouraging openness to the possibility that conservation of the state we have arrived at is inconsistent with stewardship aligned to our collective interests that are threatened. My concern with resisting implicit framing echoes (Supran & Oreskes, 2021),

whose analyses shows that historically, large industries that exist in tension with the public good such as tobacco and oil industries have engaged in communication strategies that frame harmful effects as the result of individual choices, and response to consumer demand, which have successfully limited the scope and imaginative breadth of legislative and policy responses.

In this sense, we can see that democratic states have tended to adopt shallow approaches to anti-disinformation efforts, relying on the operators of digital platforms to manage content according to applicable regulations and their own conceptions of acceptable standards. There are many reasons for this, some principled, some with roots in the political economics of media systems, some relating to the extent to which these platforms are answerable to the legal norms of the USA, where there is a broad presumption against government interference in the publication and circulation of information. As I described in Chapter 1, even pursuing domestic public diplomacy is highly limited. During the COVID-19 pandemic, there was a marked uptick in government requests that social media platforms restrict the circulation of some kinds of disinforming content (Clegg, 2020; Gadde & Derella, 2020), often made informally, and later subject to a sweeping court injunction that such efforts amounted to unconstitutional efforts to informally regulate speech

Some of the activity that prompted this ruling is highly unhelpful to sincere efforts to improve the health of the information environment because they attempted to stifle

sincere and reasonable inquiry into the origin of the pandemic, against the evidence, for political reasons. The effect has been to further diminish political, and public will for coordinated anti-disinformation policies, on the appearance that the temptation to partisanship and the risks of error are too high. Deep approaches, such as a revival of the United States Information Agency (USIA) as a non-partisan and trusted voice to communicate competing narratives better aligned with democratic principles, encouraged in a Potomac Institute report (Pearson & Moxham, 2022), seem to have at present little chance of implementation. In the US, even a recent effort to establish an anti-disinformation task force, supported by experts (Ingram, 2020), and modelled after those of NATO-aligned countries such as the Swedish Psychological Defence Agency (Sundelius & Eldeblad, 2023), was immediately politicized, taken as evidence of cognitive insecurity and government interference and censorship, and quickly abandoned (Tollefson, 2023). Experts on disinformation themselves found themselves subject to attack and hate (Jankowicz, 2022). In the fall of 2023, The Canadian Security and Intelligence Service conducted a seemingly anodyne public education campaign to inform the public of the existence of Russian disinformation efforts relating to Russia's war against the Ukraine, a series of Soviet-themed social media posts with slogans such as "Do you know who is behind it? Disinformation is here and hides well" in Cyrillic text. This prompted swift criticism from experts who noted that most Russian disinformation does not appear to come from Russia, nor involve topics explicitly relating to Russia (Robertson & Bronskill, 2023). Public reaction was quick to equivocate

the effort with propaganda, and even as potentially motivated by partisan politics. It was an odd design choice to model the imagery in the style of historic propaganda posters.

There is now increasingly scant room is left for anti-disinformation efforts to be interpreted as suitably constrained and apolitical, which in turn weakens available policy tools, or forces them to be engaged with covertly. To the extent these are uncovered, they become further evidence for decreased trust in the information systems involved.

I view this result as evidence for the risks and inadequacies of conceptions of anti-disinformation that attempt to focus on specific content and conform to the existing technical and economic constraints on action - what we can call shallow approaches. We can think of these as hypodermic modes of intervention, which I discussed in Chapter 4, section 6, which mass communication and propaganda scholars have rejected as ineffective, and prone to systemic collateral effects. I can now elaborate that argument - the most dangerous of those collateral effects are those that weaken hinge stabilities and diminish and destabilize environmental dependencies. The opportunities for successful hypodermic interventions are minimal and unstable. That they appear to many onlookers as arbitrary and too tightly bound with particular viewpoints is not accidental.

Failed shallow anti-disinformation interventions make further remedial action more difficult. An implicit constraint on cognitive security emerges that looks quite similar to the principal of least restrictive means but operationalized as the least interference with

the ways in which we currently manage and conceptualize the information environment and our relationship to it. It imagines maximization of our individual epistemic autonomy, but narrowly, mistaking the extent to which this exists largely in relation to the successful engagement of social and environmental dependencies. The safety and reliability of these is exactly what is at issue as soon as we recognize public interest in cognitive security. Framing of discourse about cognitive security often comes to exemplify what Stanley (Stanley, 2015) calls "undermining propaganda", taking up the banner of a favored ideal, such as epistemic autonomy, in a way that functionally undermines it, for example, by eroding the will to enact effective policy to protect epistemic dependencies. Broad and poorly formulated anti-misinformation policies are especially risky in this regard. A vicious circle develops, where disinformation campaigns fracture and weaken the media environment (Arayankalam & Krishnan, 2021; Murphy, 2022), increasing vulnerability to continued disinformation attack. Our epistemic environment is engineered to serve interests and values that often have little bearing on the epistemic dependencies that supervene on them, and, just as in the context of environmental policy, recognition of these genuine public interests can, and should, licence breaches in the presumptive barriers to interference with private affairs.

An exploration of the policy implications for this way of looking at cognitive security is a large project in and of itself, and out of scope here. But I would envision bringing into scope questions about the current structures of centralization and decentralization, which today largely reflect economic relations and peculiarities in the technological and

financial contexts many platforms have evolved in. The articulations of positive epistemic duties attaching to the operation of information systems at scale might include the intercedence of new regulatory bodies, the provision of public access to significant datasets, the affordance of better meta-evidence, the preservation of archives, and so forth. If information ecosystem health is a public concern, then there must be opportunities to facilitate meaningful engagement and analysis.

5.7 The ethics of cognitive security

In section 2 I cited approvingly Harris's philosophical analysis of disinformation. In the concluding passages, Harris writes, "... effectively countering the effects of disinformation on sub-doxastic states is likely to require interventions on the broader epistemic environment that support normative associations. The democratization of control over this environment due to the emergence of social media and related developments suggests that the improvement of this environment is a project in which we may all partake". (Harris, 2023, p. 17). But in fact, control of the critical, hinge-relevant epistemic infrastructure is in fact not democratic at all, it is concentrated in the hands of a few global corporations, and exposed to hidden and hostile influence and manipulation, including at the level of state action. This structurally restricts anti-disinformation efforts to shallow forms that risk self-defeat. We know that some things are filtered, but we don't have a good window as to why, and how, and by which principles. We know governments work with platforms to help them identify what is often called "inauthentic"

content, but we lack civic representation in these ostensibly private matters of platform management, even though the sporadic need to work with governments and security services at times clearly indicates genuine public interests are at stake. What we do know often reveals serious inadequacy, such as Facebook's lack of moderators with linguistic and cultural knowledge to sufficiently identify disinformation campaigns taking place over extended periods of time that contributed to serious violence in Myanmar (Rafee, 2020), and a lack of adequate tools and policies to respond to this content when it was identified. Where state security goals are pursued narrowly and within the platform affordances which happen to be available, even well-intentioned interventions are weaponizable to amplify hostile interests and generate increased evidence for skepticism that in turn further erodes the institutions cognitive security efforts are supposed to protect. Without appreciation of the structures and interdependencies of proxy trust, for example, as hinge stabilities, there is ever-present risk of damage to critical epistemic infrastructures. Merely attempting to act in constrained, limited, hands-off ways does nothing to mitigate these risks, and sediments the conditions that generate vulnerability.

Shallow approaches to cognitive security focus on content by topic and origin, whereas deep approaches focus on causes. Murphy (Murphy, 2022) argues that adversaries have splintered the US population into "smaller, uncooperative and polarised internal tribes", existing in media enclaves, unmoored from the traditional grounds of community. The conditions for this are structural features of the digital media environment, where adversaries recognized that the "opportunity for covert foreign

influence to be influential and inserted unchallenged is now significantly higher." (ibid. 40). The security implications of the interference include diminished influence, and internal instability, including loss of trust in institutions and increased capacity to sustain demagogic political movements. These further erode the capacity for successful cognitive security defense. Murphy concludes a survey of these security challenges with an appeal for "[a]nalysis of the correlation between declining institutional trust, polarisation, the rise of social media and covert Russian government efforts" (ibid. 44).

One shape such analysis can take is that digital media platforms, as both a sociotechnical and political-economic phenomena, have destabilized a set of hinge commitments that were well-understood and in light of which reliable evidential grammars had developed. Partially they have been replaced, but with hinges that in some respects look the same, but function differently, appearing to offer continued authenticity conditions for cognitive projects that are in fact imperilled or wholly undermined.

Whatever the outcome of a such an analysis might be, it is difficult to conceive of effective shallow intervention. Fact-checks and corrections work best when embedded in broader trust-fostering practices such as explaining motivations (Lewandowsky & Van Der Linden, 2021) and proactive communication is significantly more effective than reactive communication(Nye, 1990), but there is limited public and political will to build public and political representation into private media systems. We are unaccustomed to intermediating political bodies in the epistemic sphere, we tend to presume this is an

undesirable last resort. We have a wealth of negative imaginaries for how this might work, but very little in terms of positive ones. Where a recent study of disinformation defense notes that "...government's control over a country's cyberspace plays a critical function in reducing the adverse impacts of foreign disinformation" (Arayankalam & Krishnan, 2021, p. 166), it is difficult to model a method to enact this control when we accept the premises of the shallow approach. China adopts an approach that is deep in one dimension, that of intermediation of public interests, though its conception of these is articulated largely in relation to political interests, and executed in a way quite incompatible with democratic values.

I've argued that adaptive ecosystem management is a good model for such an effort, because it takes as axiomatic the inherent uncertainty of our understanding of information ecosystems, treats these as matters of public interest, and thus incorporates domain experts and civil society in the practice of management. The environmental metaphor I've developed can demonstrate the legitimacy of this kind of effort in the information environment. A major research question for the ethics of cognitive security is to find ways to intermediate epistemic authorities, in Croce's (Croce, 2018) sense of the term, and afford civil society and political representation without smuggling in authoritarianism and epistemically counterproductive properties. I've given an account of public interest in terms of epistemic ecological values, the protection of the epistemic commons that much of our knowledge has dependencies to, and described one manifestation of this interest as the protection of epistemic environmental dependencies as hinge stabilities.

Alternative accounts of the public interest that should inform an ethics of cognitive security can be developed. Ördén argues that responses to information threats should be grounded by a concern to protect civic capacity for political judgment (Ördén, 2022). Political judgement is not a domain of narrow, political problem-solving, but depends on our ability to understand, empathize, and imagine those around us, because "...the agency of the individual is dependent on the agency of others, and therefore requires communication for its realization." (ibid. p. 385). If our communication systems cause our representations of our peers and their beliefs to be distorted and inaccurate in ways we can't see, our political judgement is impaired.

In a paper that begins, just as this dissertation does, with a discussion of information threats and cognitive security, it might seem surprising that, at the end, Ördén asks, "if we do indeed regard cyberspace as the new home of the mind, the aim of political judgment is to make this contemporary world homely" (ibid. p. 388). The idea of making the information environment more "homely" might seem oddly wistful and far removed from the sorts of efforts usually undertaken in the name of cyber and cognitive security. But I think in light of the arguments I've made, this should be a little less surprising. If the information environment partially constitutes authenticity conditions for many of our cognitive projects, including those directly related to political judgement, it cannot safely serve this function unless is managed with these interests in mind. It must be a home appropriate to the kinds of activities that take place within it. Our entitlement to assuming

the hinge commitments that are shared by many of these authenticity conditions stems in part from their centrality in community practices that we could not participate in if we had to doubt them. Unfortunately, as things stand today, we have adequate reason to doubt many of them, and the consequent loss of trust impairs the ability of political institutions to react to genuine public interests and crises.

In Ördén's conception, a primary aim of cognitive security is the generation of political legitimacy, by protecting the capacity of people and future people to develop and exercise characteristically democratic political judgement. In mine, I focus on a level below this, the broader epistemological dependencies of political judgement, and epistemic properties of the communication infrastructure through which we encounter each other. Some argue that individual judgement, knowledge and epistemic agency is not central to the success of collective political and epistemic projects, for example Brennan's theory of epistocracy (Brennan, 2016). Levy and Alfano (Levy & Alfano, 2020) argue that there is no reason to think a high degree of accuracy and epistemic virtue is require for the maintenance of the collective stores of knowledge we depend on, and in fact a great deal of knowledge-conducive practice involves individual behaviours traditionally taken to be epistemically vicious such as imitation, repetition, and reliance on authority. Yet, even in these cases, we depend on the reliability of the epistemic infrastructure involved, that connect us to reliable knowledge and practice, even if we don't participate meaningfully in their production and maintenance.

Both Ördén's and my own conception of cognitive security is capable of satisfying concerns about objectionable epistemic paternalism in the pursuit of cognitive security in ways that shallow conceptions cannot. Most importantly, many of the interventions the deep approach embraces, including positive efforts to produce authoritative content (such as the USIA model), and the intermediation of civic management where we have only narrow private management, do not pass the strength-of-substitution-of-judgement test I established as a marker of potentially paternalist acts. They depend on forms of ordinary epistemic stewardship and management that I argued in Chapter 2 require no special justification other than their efficacy and legitimacy of purpose. Deep approaches explicitly disavow filtration models of defense (which do risk paternalism), in part because these risk damaging hinge dependencies.

However, there is still scope within deep approaches for actions where some kinds of epistemic agency are constrained or influenced in ways that are strong enough that the substitution threshold might be met. AI models are increasingly shared across enormous ranges of application, and as they become ubiquitous, they can be sites for paternalistic intervention, such in the application of reinforcement learning from human feedback training phases (Ouyang et al., 2022) designed to make output conform to conceptions of safety (including not producing offensive content, or copywritten content, or content that encourages illegal behavior). Notions of what is safe, and what is allowed, are embedded into these models, which will then refuse to produce proscribed content. As things stand, the processes whereby these are articulated and enforced lack vital transparency.

Potential cognitive security interventions that take advantage of them could meet the strength of substitution tests and thus be genuinely paternalistic and require elevated justification.

I've argued that we can justify epistemically paternalist interventions when they pursue a genuine epistemic good in a way that we should reasonably expect is likely to be successful. Just as at times we willingly sacrifice other dimensions of our wellbeing, for example, economic gain, in order to protect environmental goods, so we might prioritize epistemic well-being over other types of well-being when the epistemic goods protected are likewise both self and other regarding because of our interdependent relation to them. However, these types of interventions are premised on structural conditions, such as of centralization, that we might have reason to think themselves are harmful, and private monopolies that lack adequate public representation. There is a high risk of undermining trust in the models, which as I've argued risks public interests in their functioning if they underlie sufficient dependence to count as a hinge, but also risks diminishing adoption, undermining the efficacy of the intervention.

Conclusion

At the outset of this project I set out three primary aims. A programmatic goal, to give reasons in favour of establishment of the ethics of cognitive security as field of inquiry. A theoretical goal, to articulate an ecological conception of our epistemic interdependence

on the information environment and show its applicability to the project of an ethics of cognitive security. And finally a methodological goal, to show that environmental philosophy and social epistemology has much to contribute to this undertaking.

In the first chapter, I outlined the information threats and the challenges of cognitive security, and their relevance to properties of the information environment. I raised the worry that to narrowly pursue securitized goals might conflict with our collective epistemic interests in the information environment and argued that articulating the nature of this interest could provide an alternate framing of the goals of cognitive security - protection of the information environment itself as a kind of public good. In the second chapter, I described epistemic paternalism as class of objection that can be levied against defensive actions we might take in the face of hostile information operations and epistemic environmental degradation, demanding elevated justification for autonomy limiting interference. The literature that has arisen around the concept of epistemic paternalism, especially as it relates to management of digital media systems, raises the possibility that we often run the risk of paternalism when we attempt to improve the information environment with the aim of consequently improving epistemic outcomes in those that depend on it. I argued that in fact most such efforts should not count as paternalist, instead arguing that these are forms of stewardship, and delineated a narrow category genuine paternalism.

We can therefore pursue stewardship of the information environment without worrying about the risk that this entails a form of paternalism, but to do so we still need an account of the goals such efforts should pursue. Therefore, in the third chapter, I developed a conception of epistemic environmental dependence and environmental trust that describes the indirect and collateral dependencies we engage on the way to engaging specific epistemic dependencies. As a result, we can be said to have a collective interest in the health and functioning of the information environment, the web of interdependencies we have to each other as epistemic agents, materialized in informational resources, artifacts, communication systems, and the like.

The fourth chapter develops an environmental metaphor to describes this interdependence and applies it to the problem of its protection. What kinds of reasons and methods are applicable to the stewardship of common goods that we have complex interdependencies to? I argue that when we analyze concepts such as epistemic pollution, which describe threats to this dependency, we discover normative considerations that, as in the environmental context, help to identify our ethical relationship to our common interests in the protection of ecological health. Finally, in the fifth chapter, I apply the ecological metaphor to the analysis of the cognitive security problem of disinformation campaigns, as both a theoretical and policy challenge. Here the goal is twofold, first to describe and defend an ecological analysis, but secondly, to offer this analysis as an example of the mode of engagement with cognitive security that I am advocating.

I hope by this to have shown that there is a coherent and worthwhile research project to be undertaken under the banner of the ethics of cognitive security, to articulate the public interest in a healthy information environment in a way that can generate, justify, an effective, robust, and ethical cognitive security policy. Even if we are skeptical that such policies should be pursued, they are actively being developed and implemented, and thus deserve engagement from skeptics and cautious advocates alike. In the case of the earth's environment, appreciation of the nature of ecological damage threatens what must be described as shared, communal interests, and has been widely regarded to demand political, economic, philosophical, even spiritual reflection on the origin of the risks and harms. Recognition and care for interdependence makes possible sincere environmental policy conversations across partisan divides.

We should be resistant to shallow approaches to cognitive security in large part because we should expect that analysis will show limited prospects for success, and that significant interests have been left out of frame, and that as a result, that there is limited basis for rational agreement with such policies. The dismal track record of shallow environmental policy should motivate hesitancy in adopting similar efforts to protect yet another critical common good. Where environmental policy has been most successful, is has articulated clear vision of the empirical and ethical stakes, and boldly engaged with private sector and civil society to enact root-cause solutions, as did the Montreal Protocol. The primary goal of an ethics of cognitive security is to develop the theoretical and empirical grounds that can lend confidence to undertake similarly robust analysis that

can gain sufficient consensus to support successful efforts to protect the information environment, and, even more importantly, to prevent the adoption of efforts likely to cause further harm.

References

Abrams, S. (2016). Beyond Propaganda: Soviet Active Measures in Putin's Russia. *Connections: The Quarterly Journal*, *15*(1), 5–31. https://doi.org/10.11610/Connections.15.1.01

Ahlstrom-Vij, K. (2013). *Epistemic paternalism: A defence*. Palgrave Macmillan.

Ahmad, K. (2007). Pakistan struggles to eradicate polio. *The Lancet Infectious Diseases*, *7*(4), 247.

Ali, S., Saeed, M. H., Aldreabi, E., Blackburn, J., De Cristofaro, E., Zannettou, S., & Stringhini, G. (2021). Understanding the effect of deplatforming on social networks. *Proceedings of the 13th ACM Web Science Conference 2021*, 187–195.

Allen Anderson, P. (2015). Neo-Muzak and the Business of Mood. *Critical Inquiry*, *41*(4), 811–840. https://doi.org/10.1086/681787

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

Anderson, I., Gil, S., Gibson, C., Wolf, S., Shapiro, W., Semerci, O., & Greenberg, D. M. (2021). "Just the Way You Are": Linking Music Listening on Spotify and Personality. *Social Psychological and Personality Science*, *12*(4), 561–572. https://doi.org/10.1177/1948550620923228

Arayankalam, J., & Krishnan, S. (2021). Relating foreign disinformation through social media, domestic online media fractionalization, government's control over cyberspace, and social media-induced offline violence: Insights from the agenda-building theoretical perspective. *Technological Forecasting and Social Change*, *166*, 120661. https://doi.org/10.1016/j.techfore.2021.120661

Aukia, J. (2021). China as a Hybrid Influencer: Non-state Actors as State Proxies. *The European Centre of Excellence for Countering Hybrid Threats*.

Bachmann, I., & Valenzuela, S. (2023). Studying the Downstream Effects of Fact-Checking on Social Media: Experiments on Correction Formats, Belief Accuracy, and Media Trust. *Social Media + Society*, *9*(2). https://doi.org/10.1177/20563051231179694

Baier, A. (1986). Trust and Antitrust. *Ethics*, *96*(2), 231–260.

Bakir, V. (2020). Psychological Operations in Digital Political Campaigns: Assessing Cambridge Analytica's Psychographic Profiling and Targeting. *Frontiers in Communication*, *5*, 67. https://doi.org/10.3389/fcomm.2020.00067

Baldwin-Philippi, J. (2019). Data campaigning: Between empirics and assumptions. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1437

Bandini, A., Bernal, A., & Axtell, G. (2020). Epistemic paternalism in doctor-patient relationships. *Epistemic Paternalism: Conceptions, Justifications and Implications*, 123–137.

Bauer, F., & Wilson, K. L. (2022). Reactions to China-linked Fake News: Experimental Evidence from Taiwan. *The China Quarterly*, *249*, 21–46. https://doi.org/10.1017/S030574102100134X

Beauchamp-Mustafaga, N. (2019). Cognitive domain operations: The PLA's new holistic concept for influence operations. *China Brief*, *19*(16), 24–37.

Begby, E. (2021). Evidential preemption. *Philosophy and Phenomenological Research*, *102*(3), 515–530.

Bellotti, L., & Moriconi, E. (2020). On Trust in Mathematics: Some Case Studies. In A. Fabris (Ed.), *Trust* (Vol. 54, pp. 95–109). Springer International Publishing. https://doi.org/10.1007/978-3-030-44018-3_7

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Benson, P. (1991). Autonomy and oppressive socialization. *Social Theory and Practice*, *17*(3), 385–408.

Bernal, A., Carter, C., Singh, I., Cao, K., & Madreperla, O. (2020). Cognitive warfare: An attack on truth and thought. *NATO and Johns Hopkins University: Baltimore MD, USA*.

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive consumer choice processes. *Journal of Consumer Research*, *25*(3), 187–217.

Blodget, H. (2009). Mark Zuckerberg on innovation. *Business Insider*. https://perma.cc/3FRJ-YPFV

Boghardt, T. (2009). Soviet Bloc Intelligence and Its AIDS Disinformation Campaign. *Studies in Intelligence*, *53*(4), 1–24.

Boichak, O. (2023). Mapping the Russian Political Influence Ecosystem: The Night Wolves Biker Gang. *Social Media + Society*, *9*(2). https://doi.org/10.1177/20563051231177920

BonJour, L. (1985). *The structure of empirical knowledge*. Harvard University Press.

Brantly, A. (2020). A brief history of fake: Surveying Russian disinformation from the Russian Empire through the Cold War and to the present. In *Information Warfare in the Age of Cyber Conflict* (pp. 27–41). Routledge.

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, *118*(5). https://doi.org/10.1073/pnas.2020043118

Brennan, J. (2016). *Against democracy*. Princeton University Press.

Broniatowski, D. A., Jamison, A. M., Qi, S., Alkulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *Public Health*, *108*, 1378–1384. https://doi.org/10.2105/AJPH.2018.304567

Brown, J. (2009). What's happened to anti-Americanism, and to the State Department? The Obama administration and public diplomacy: March to mid-June 2009. *Place Branding and Public Diplomacy*, *5*(3), 247–252. https://doi.org/10.1057/pb.2009.9

Buchanan, A. (1978). Medical paternalism. *Philosophy & Public Affairs*, 370–390.

Bullock, E. C. (2018a). Knowing and Not-Knowing For Your Own Good: The Limits of Epistemic Paternalism. *Journal of Applied Philosophy*, *35*(2), 433–447. https://doi.org/10.1111/japp.12220

Bullock, E. C. (2018b). Knowing and not-knowing for your own good: The limits of epistemic paternalism. *Journal of Applied Philosophy*, *35*(2), 433–447. https://doi.org/10.1111/japp.12220

Busch, C. (2022). Regulating the Expanding Content Moderation Universe: A European Perspective on Infrastructure Moderation. *UCLA JL & Tech.*, *27*, 32.

Carter, J. (2020). Trust and its significance in social epistemology. In *Oxford Handbook of Social Epistemology*. OUP Oxford.

Childress, J. F. (2020). *Public bioethics: Principles and problems*. Oxford University Press, USA.

Chodkowski, W. (2012). *Fact Sheet—The United States Information Agency* (The American Security Project).

Chomsky, N. (2003). *Necessary illusions: Thought control in democratic societies*. House of Anansi Press.

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

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811–824. https://doi.org/10.1109/TDSC.2012.75

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

Clarke, S. (2002). A definition of paternalism. *Critical Review of International Social and Political Philosophy*, *5*(1), 81–91.

Claverie, B., & Du Cluzel, F. (2022). *"Cognitive Warfare": The Advent of the Concept of "Cognitics" in the Field of Warfare*. NATO Collaboration Support Office.

Clegg, N. (2020, March 25). Combating COVID-19 Misinformation Across Our Apps. *Meta*. https://about.fb.com/news/2020/03/combating-covid-19-misinformation/

Coady, C. A. J. (1992). *Testimony: A philosophical study*. Clarendon Press.

Cohen, R. S., Beauchamp-Mustafaga, N., Cheravitch, J., Demus, A., Harold, S., Hornung, J. W., Jun, J., Schwille, M., Treyger, E., & Vest, N. (2021). *Combating Foreign Disinformation on Social Media: Study Overview and Conclusions*. RAND Corporation.

Coliva, A. (2015). *Extended rationality: A hinge epistemology*. Palgrave Macmillan.

Congiu, L., & Moscati, I. (2021). A review of nudges: Definitions, justifications, effectiveness. *Journal of Economic Surveys*. https://doi.org/10.1111/joes.12453

Corstange, D., & Marinov, N. (2012). Taking Sides in Other People's Elections: The Polarizing Effect of Foreign Intervention: TAKING SIDES IN OTHER PEOPLE'S ELECTIONS. *American Journal of Political Science*, *56*(3), 655–670. https://doi.org/10.1111/j.1540-5907.2012.00583.x

Croce, M. (2018). Epistemic Paternalism and the Service Conception of Epistemic Authority. *Metaphilosophy*, *49*(3), 305–327. https://doi.org/10.1111/meta.12294

Croce, M. (2022). On testimonial knowledge and its functions. *Synthese*, *200*(2), 141. https://doi.org/10.1007/s11229-022-03528-x

de Marneffe, P. (2006). Avoiding Paternalism. *Philosophy & Public Affairs*, *34*(1), 68–94.

Dear, K. (2021). Artificial intelligence, security, and society. In *The World Information War* (pp. 231–255). Routledge.

DeFleur, M. L., & Ball-Rokeach, S. (1989). *Theories of mass communication* (5th ed). Longman.

DeLaet, D. L. (2014). Whose Interests is the Securitization of Health Serving? In *Routledge handbook of global health security* (pp. 339–348). Routledge.

Dennett, D. C. (1986). Information, Technology, and the Virtues of Ignorance. *Daedalus*, *115*(3), 135–153.

Department of National Defence. (2014). *The future security environment 2013-2040*.

Derrida, J. (1991). *A Derrida reader: Between the blinds* (P. Kamuf, Ed.). Columbia University Press.

Douek, E. (2020). The rise of content cartels. *Knight First Amendment Institute at Columbia*.

Douek, E. (2022). Content moderation as systems thinking. *Harv. L. Rev.*, *136*, 526.

Dretske, F. (1991). Two conceptions of knowledge: Rational vs. Reliable belief. *Grazer Philosophische Studien*, *40*, 15.

Dworkin, G. (1972). Paternalism. *The Monist*, *56*(1), 64–84.

Dworkin, G. (1981). Paternalism and welfare policy. *Income Support: Conceptual and Policy Issues*.

Dworkin, G. (2015). Defining paternalism. In *New perspectives on paternalism and health care* (pp. 17–29). Springer.

Eisenstein, M. (2022). Vaccination rates are falling, and it's not just the COVID-19 vaccine that people are refusing. *Nature*, *612*(7941), S44–S46. https://doi.org/10.1038/d41586-022-04341-9

Enders, A. M., Uscinski, J. E., Seelig, M. I., Klofstad, C. A., Wuchty, S., Funchion, J. R., Murthi, M. N., Premaratne, K., & Stoler, J. (2021). The relationship between social media use and beliefs in conspiracy theories and misinformation. *Political Behavior*, 1–24.

Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, *112*(33), E4512–E4521.

Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G., & Rand, D. (2023). The social media context interferes with truth discernment. *Science Advances*, *9*(9), eabo6169. https://doi.org/10.1126/sciadv.abo6169

Fallis, D. (2021). The Epistemic Threat of Deepfakes. *Philosophy & Technology*, *34*(4), 623–643. https://doi.org/10.1007/s13347-020-00419-2

Feinberg, J. (1984). Environmental pollution & the threshold of harm. *Hastings Center Report*, 27–31.

Fetzer, J. H. (2004). Disinformation: The use of false information. *Minds and Machines*, *14*, 231–240.

Floridi, L. (2013). *The philosophy of information*. OUP Oxford.

Fogg, B. J. (2002). Persuasive technology: Using computers to change what we think and do. *Ubiquity*, *2002*(December), 2. https://doi.org/10.1145/764008.763957

Fondren, E. (2021). "We are Propagandists for Democracy": The Institute for Propaganda Analysis' Pioneering Media Literacy Efforts to Fight Disinformation (1937–1942). *American Journalism*, *38*(3), 258–291. https://doi.org/10.1080/08821127.2021.1950481

Fox, S., & Devall, B. (1981). *John Muir and his legacy: The American conservation movement*. Little Brown.

Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J. R., Olin, J., Ross, M. Q., & Szafran, J. (2021). Fake Claims of Fake News: Political Misinformation, Warnings, and the Tainted Truth Effect. *Political Behavior*, *43*(4), 1433–1465. https://doi.org/10.1007/s11109-020-09597-3

Fricker, E. (2006a). Second-hand knowledge. *Philosophy and Phenomenological Research*, *73*(3), 592–618.

Fricker, E. (2006b). Testimony and epistemic autonomy. *The Epistemology of Testimony*, 225–250.

Frischmann, B., & Selinger, E. (2018). *Re-engineering humanity*. Cambridge University Press.

Frost-Arnold, K. (2021). The Epistemic Dangers of Context. *Applied Epistemology*, 437.

Gadde, V., & Derella, M. (2020, March 16). An update on our continuity strategy during COVID-19. *Twitter Blog*. https://blog.twitter.com/en_us/topics/company/2020/An-update-on- our-continuity-strategy-during-COVID-19.html

Ganeri, J. (2018). Epistemology from a sanskritic point of view. *Epistemology for the Rest of the World*, 13–21.

Gert, B., & Culver, C. M. (1976). Paternalistic Behavior. *Philosophy & Public Affairs*, *6*(1), 45–57.

Giesea, J. (2015). It's time to embrace memetic warfare. *Defence Strategic CommunicationsDefence Strategic Communications*, *1*(1), 67–75.

Glasius, M., & Michaelsen, M. (2018). Authoritarian practices in the digital age illiberal and authoritarian practices in the digital sphere—Prologue. *International Journal of Communication*, *12*(19).

Goldberg, S. C. (2011). The Division of Epistemic Labor. *Episteme*, *8*(1), 112–125. https://doi.org/10.3366/epi.2011.0010

Goldberg, S. C. (2012). Epistemic extendedness, testimony, and the epistemology of instrument-based belief. *Philosophical Explorations*, *15*(2), 181–197. https://doi.org/10.1080/13869795.2012.670719

Goldberg, S. C. (2013). Anonymous Assertions. *Episteme*, *10*(2), 135–151. https://doi.org/10.1017/epi.2013.14

Goldberg, S. C. (2020). Epistemically engineered environments. *Synthese*, *197*(7), 2783–2802. https://doi.org/10.1007/s11229-017-1413-0

Goldman, A. I. (1976). Discrimination and perceptual knowledge. *The Journal of Philosophy*, *73*(20), 771–791.

Goldman, A. I. (1991). Epistemic Paternalism: Communication Control in Law and Society. *The Journal of Philosophy*, *88*(3), 113–131. https://doi.org/10.2307/2026984

Goldman, A. I. (1999). *Knowledge in a social world*. Oxford University Press.

Goldman, A. I. (2000). *Replies to reviews of Knowledge in a Social World*.

Goldman, A. I. (2001). Experts: Which Ones Should You Trust? *Philosophy and Phenomenological Research*, *63*(1), 85–110. https://doi.org/10.1111/j.1933-1592.2001.tb00093.x

Golebiewski, M., & Boyd, D. (2019). *Data voids: Where missing data can easily be exploited*. https://apo.org.au/node/265631

González, R. J. (2017). Hacking the citizenry?: Personality profiling, 'big data' and the election of Donald Trump. *Anthropology Today*, *33*(3), 9–12. https://doi.org/10.1111/1467-8322.12348

Goreis, A., & Voracek, M. (2019). A systematic review and meta-analysis of psychological research on conspiracy beliefs: Field characteristics, measurement instruments, and associations with personality traits. *Frontiers in Psychology*, *10*, 205.

Griffiths, J. (2021). *The great firewall of China: How to build and control an alternative version of the internet*. Bloomsbury Publishing.

Grimmelmann, J. (2017). The platform is the message. *Geo. L. Tech. Rev.*, *2*, 217.

Griswold, K. (2023, August 22). *Look How Tightly Google Controls What You Learn About Big News*. The Federalist. https://thefederalist.com/2023/08/22/look-how-google-shoos-you-away-from-the-biden-family-biz-and-other-big-news/

Grundmann, T. (2023). The Possibility of Epistemic Nudging. *Social Epistemology*, *37*(2), 208–218. https://doi.org/10.1080/02691728.2021.1945160

Guess, A. M., & Lyons, B. A. (2020). Misinformation, Disinformation, and Online Propaganda. In *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press.

Gwynne, D. T., & Rentz, D. C. (1983). Beetles on the bottle: Male buprestids mistake stubbies for females (Coleoptera). *Australian Journal of Entomology*, *22*(1), 79–80.

Habgood-Coote, J. (2019). Stop talking about fake news! *Inquiry*, *62*(9–10), 1033–1065.

Hairston, N. G., Smith, F. E., & Slobodkin, L. B. (1960). Community structure, population control, and competition. *The American Naturalist*, *94*(879), 421–425.

Hannon, M. (2021). Skepticism, Fallibilism, and Rational Evaluation. In *Skeptical invariantism reconsidered* (pp. 172–194). Routledge.

Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy*, *82*(7), 335–349.

Harris, K. R. (2023). Beyond belief: On disinformation and manipulation. *Erkenntnis*, 1–21.

Hassan, N. R., Mingers, J., & Stahl, B. (2018). Philosophy and information systems: Where are we and where should we go? *European Journal of Information Systems*, *27*(3), 263–277. https://doi.org/10.1080/0960085X.2018.1470776

Hausman, D. M. (2018). Behavioural Economics and Paternalism. *Economics and Philosophy*, *34*(1), 53–66. https://doi.org/10.1017/S0266267117000244

Hayden, M. V. (2019). *The assault on intelligence: American national security in an age of lies*. Penguin.

Hershey, P. T. (1985). A definition for paternalism. *The Journal of Medicine and Philosophy*, *10*(2), 171–182.

Hopple, G. W. (1980). *Internal and External Crisis Early Warning and Monitoring*. International Public Policy Research Corporation.

Horta Ribeiro, M., Hosseinmardi, H., West, R., & Watts, D. J. (2023). Deplatforming did not decrease Parler users' activity on fringe social media. *PNAS Nexus*, *2*(3).

Househ, M. S., Aldosari, B., Alanazi, A., Kushniruk, A. W., & Borycki, E. M. (2017). Big Data, Big Problems: A Healthcare Perspective. *ICIMTH*, 36–39.

Huang, R., Zheng, X., Shang, Y., & Xue, X. (2023). On challenges of AI to cognitive security and safety. *Security and Safety*, *2*, 2023012.

Humprecht, E., Esser, F., & Van Aelst, P. (2020). Resilience to online disinformation: A framework for cross-national comparative research. *The International Journal of Press/Politics*, *25*(3), 493–516.

Hung, T.-C., & Hung, T.-W. (2022). How China's Cognitive Warfare Works: A Frontline Perspective of Taiwan's Anti-Disinformation Wars. *Journal of Global Security Studies*, *7*(4). https://doi.org/10.1093/jogss/ogac016

Hwang, T. (2020). Dealing with Disinformation: Evaluating the Case for Amendment of Section 230 of the Communications Decency Act. *Social Media and Democracy: The State of the Field, Prospects for Reform*, 252.

Ingram, H. J. (2020). Persuade or Perish: Addressing Gaps in the US Posture to Confront Propaganda and Disinformation Threats. *Program on Extremism*.

Innes, H., & Innes, M. (2023). De-platforming disinformation: Conspiracy theories and their control. *Information, Communication & Society*, *26*(6), 1262–1280.

Innes, M., Innes, H., Roberts, C., Harmston, D., & Grinnell, D. (2021). The normalisation and domestication of digital disinformation: On the alignment and consequences of far-right and Russian State (dis)information operations and campaigns in Europe. *Journal of Cyber Policy*, *6*(1), 31–49. https://doi.org/10.1080/23738871.2021.1937252

Isachenkov, V. (2019, February 11). *Official: Russia's political system a good model for others*. AP News. https://apnews.com/article/e9757984e28b495bad5feb2bd702032e

Ivy, V. (2021). Epistemology of Anonymous Assertions. *Applied Epistemology*, 457–481.

Jankowicz, N. (2022). *How to be a woman online: Surviving abuse and harassment, and how to fight back*. Bloomsbury Publishing.

Janzen, S., Orr, C., & Terp, S.-J. (2022). Cognitive security and resilience: A social ecological model of disinformation and other harms with applications to COVID-

19 vaccine information behaviors. *ROMCIR 2022 CEUR Workshop Proceedings*, *3138*, 48–88.

Jegede, A. S. (2007). What Led to the Nigerian Boycott of the Polio Vaccination Campaign? *PLoS Medicine*, *4*(3), e73. https://doi.org/10.1371/journal.pmed.0040073

Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (2021). Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–30.

Jost, J. T., van der Linden, S., Panagopoulos, C., & Hardin, C. D. (2018). Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. *Current Opinion in Psychology*, *23*, 77–83.

Kahan, D. M. (2017, June 16). *Protecting or Polluting the Science Communication Environment?* The Oxford Handbook of the Science of Science Communication. https://doi.org/10.1093/oxfordhb/9780190497620.013.45

Kenyon, T. (2013). The informational richness of testimonial contexts. *The Philosophical Quarterly*, *63*(250), 58–80.

Kinahan, G. M. (1990). Exposing Soviet Active Measures in the 1980s: A Model for the Bush Administration? *The Journal of Social, Political, and Economic Studies*, *15*(3), 301.

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

Kleinig, J. (1983). *Paternalism*. Manchester University Press.

Knox, J. H. (2008). The boundary waters treaty: Ahead of its time, and ours. *Wayne L. Rev.*, *54*, 1591.

Kosseff, J. (2023). *Liar in a crowded theater: Freedom of speech in a world of misinformation*. Johns Hopkins University Press.

Krafft, P. M., & Donovan, J. (2020). Disinformation by Design: The Use of Evidence Collages and Platform Filtering in a Media Manipulation Campaign. *Political Communication*, *37*(2), 194–214. https://doi.org/10.1080/10584609.2019.1686094

La Morgia, M., Mei, A., Mongardini, A. M., & Wu, J. (2021). Uncovering the dark side of Telegram: Fakes, clones, scams, and conspiracy movements. *arXiv Preprint arXiv:2111.13530*.

Lackey, J. (2003). A Minimal Expression of Non-Reductionism in the Epistemology of Testimony. *Noûs*, *37*(4), 706–723.

Lackey, J. (2008). *Learning from words: Testimony as a source of knowledge*. Oxford University Press.

Le Grand, J., & New, B. (2015). *Government Paternalism: Nanny State or Helpful Friend?* (1st ed.). Princeton University Press. https://doi.org/10.23943/princeton/9780691164373.001.0001

Le Guyader, H. (2022). *Cognitive domain: A sixth domain of operations*. NATO Collaboration Support Office.

Lear, J. (2022). *Imagining the end: Mourning and ethical life*. The Belknap press of Harvard University press.

Leetaru, K., & Schrodt, P. A. (2013). Gdelt: Global data on events, location, and tone, 1979–2012. *ISA Annual Convention*, *2*(4), 1–49.

Leijnen, S., Aldewereld, H., van Belkom, R., Bijvank, R., & Ossewaarde, R. (2020). An agile framework for trustworthy AI. *NeHuAI@ ECAI*, 75–78.

Leopold, A. (1933). The conservation ethic. *Journal of Forestry*, *31*(6), 634–643.

Leopold, A. (1943). Wildlife in American culture. *The Journal of Wildlife Management*, *7*(1), 1–6.

Leopold, A. (1949). A Sand County almanac and sketches here and there Oxford University Press. *New York*.

Leopold, A., Sowls, L. K., & Spencer, D. L. (1947). A survey of over-populated deer ranges in the United States. *The Journal of Wildlife Management*, *11*(2), 162–177.

Lessler, M. A. (1988). Lead and Lead Poisoning from Antiquity to Modern Times. *The Ohio Journal of Science.*, *88*(3), 78–84.

Levy, N. (2018). Taking responsibility for health in an epistemically polluted environment. *Theoretical Medicine and Bioethics*, *39*(2), 123–141.

Levy, N. (2022). *Bad beliefs: Why they happen to good people*. Oxford University Press.

Levy, N., & Alfano, M. (2020). Knowledge from vice: Deeply social epistemology. *Mind*, *129*(515), 887–915.

Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384.

Lewis, B., & Marwick, A. E. (2017). *Media Manipulation and Disinformation Online*. https://datasociety.net/library/media-manipulation-and-disinfo-online/

Lewis, P. R., & Marsh, S. (2022). What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research*, *72*, 33–49. https://doi.org/10.1016/j.cogsys.2021.11.001

Lorenz, T. (2022, May 18). How the Biden administration let right-wing attacks derail its disinformation efforts. *Washington Post*. https://www.washingtonpost.com/technology/2022/05/18/disinformation-board-dhs-nina-jankowicz/

Lusebrink, I., Girling, R. D., Farthing, E., Newman, T. A., Jackson, C. W., & Poppy, G. M. (2015). The effects of diesel exhaust pollution on floral volatiles and the consequences for honey bee olfaction. *Journal of Chemical Ecology*, *41*, 904–912.

Mäkinen, M., & Wangu Kuira, M. (2008). Social Media and Postelection Crisis in Kenya. *The International Journal of Press/Politics*, *13*(3), 328–335. https://doi.org/10.1177/1940161208319409

Marangione, M. S. (2021). Words as Weapons: The 21st Century Information War. *Global Security & Intelligence Studies*, *6*(1).

Martin, D. A., Shapiro, J. N., & Nedashkovskaya, M. (2019). Recent trends in online foreign influence efforts. *Journal of Information Warfare*, *18*(3), 15–48.

Martinez-Martin, N., Insel, T. R., Dagum, P., Greely, H. T., & Cho, M. K. (2018). Data mining for health: Staking out the ethical territory of digital phenotyping. *Npj Digital Medicine*, *1*(1), 68. https://doi.org/10.1038/s41746-018-0075-8

Marwick, A. E., & boyd, danah. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. https://doi.org/10.1177/1461444810365313

Matthes, J., Hirsch, M., Stubenvoll, M., Binder, A., Kruikemeier, S., Lecheler, S., & Otto, L. (2022). Understanding the democratic role of perceived online political micro-targeting: Longitudinal effects on trust in democracy and political interest. *Journal of Information Technology & Politics*, *19*(4), 435–448. https://doi.org/10.1080/19331681.2021.2016542

Mazzetti, M., & Gordon, M. R. (2015, June 13). ISIS Is Winning the Social Media War, U.S. Concludes. *The New York Times*. https://www.nytimes.com/2015/06/13/world/middleeast/isis-is-winning-message-war-us-concludes.html

McBride, M. K., Gold, Z., & Stricklin, K. (2020). Social Media Bots: Implications for Special Operations Forces. *Center for Naval Analysis, September*.

McKenna, R. (2020). Persuasion and epistemic paternalism. *Epistemic Paternalism: Conceptions, Justifications, and Implications*, 91–106.

McLuhan, M., & McLuhan, E. (1999). *Laws of media: The new science* (Repr). Univ. of Toronto Press.

Meehan, D. (2020). Epistemic vice and epistemic nudging: A solution? *Epistemic Paternalism: Conceptions, Justifications and Implications*, 247–259.

Melki, M., & Pickering, A. (2020). Polarization and corruption in America. *European Economic Review*, *124*, 103397.

Meserve, S. A., & Pemstein, D. (2018). Google politics: The political determinants of Internet censorship in democracies. *Political Science Research and Methods*, *6*(2), 245–263.

Michaelian, K. (2010). In defence of gullibility: The epistemology of testimony and the psychology of deception detection. *Synthese*, *176*(3), 399–427. https://doi.org/10.1007/s11229-009-9573-1

Mill, J. S. (1998). *On Liberty and Other Essays*. Oxford University Press.

Millgram, E. (2015). *The great endarkenment: Philosophy for an age of hyperspecialization*. Oxford university press.

Mills, C. (2018). The Choice Architect's Trilemma. *Res Publica*, *24*(3), 395–414. https://doi.org/10.1007/s11158-017-9363-4

Millstein, R. L. (2018). Understanding Leopold's concept of "interdependence" for environmental ethics and conservation biology. *Philosophy of Science*, *85*(5), 1127–1139.

Millstein, R. L. (2020). Defending a Leopoldian basis for biodiversity: A response to Newman, Varner, and Linquist. *Biology & Philosophy*, *35*, 1–11.

Milmo, D., & Davidson, H. (2022, November 28). Chinese bots flood Twitter in attempt to obscure Covid protests. *The Guardian*. https://www.theguardian.com/technology/2022/nov/28/chinese-bots-flood-twitter-in-attempt-to-obscure-covid-protests

Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W., Elovici, Y., & Biggio, B. (2023). The Threat of Offensive AI to Organizations. *Computers & Security*, *124*, 103006. https://doi.org/10.1016/j.cose.2022.103006

Misak, C. J. (2004). The critical care experience: A patient's view. *American Journal of Respiratory and Critical Care Medicine*, *170*(4), 357–359.

Mitchell, G. (2004). Libertarian paternalism is an oxymoron. *Nw. UL Rev.*, *99*, 1245.

Moore, M., & Tambini, D. (2018). *Digital dominance: The power of Google, Amazon, Facebook, and Apple*. Oxford University Press.

Morozov, E. (2009, June 9). The future of "Public Diplomacy 2.0." *Foreign Policy*. https://foreignpolicy.com/2009/06/09/the-future-of-public-diplomacy-2-0/

Murphy, B. (2022). Decaying National Security and the Rise of Imagined Tribalism. *The RUSI Journal*, *166*(6–7), 32–44.

Murphy, B. (2023). In Defense of Disinformation. *Journal of Homeland Security and Emergency Management*, *0*(0). https://doi.org/10.1515/jhsem-2022-0045

Naess, A. (1973). The shallow and the deep, long-range ecology movement. A summary. *Inquiry*, *16*(1–4), 95–100.

Naess, A. (1989). The Basics of Deep Ecology. *Resurgence*, *126*(6).

Nesic, A. (2022). Cognitive Security and Emotional Warfare: The Science and Practice for Understanding, Analyzing and Preventing the Spread of Violent Extremism in the CENTCOM AOR. In *The Great Power Competition Volume 2: Contagion Effect: Strategic Competition in the COVID-19 Era* (pp. 233–248). Springer.

Nguyen, C. T. (2020a). Echo chambers and epistemic bubbles. *Episteme*, *17*(2), 141–161.

Nguyen, C. T. (2020b). *Trust as an unquestioning attitude*.

Nguyen, C. T. (2022). Trust as an Unquestioning Attitude. In T. S. Gendler, J. Hawthorne, & J. Chung (Eds.), *Oxford Studies in Epistemology Volume 7* (1st ed., pp. 214–244). Oxford University PressOxford. https://doi.org/10.1093/oso/9780192868978.003.0007

NOAA. (1974). *Report to the Congress on Ocean Dumping and Other Man-Induced Changes to Ocean Ecosystems*. National Oceanic and Atmospheric Administration.

Nogara, G., Vishnuprasad, P. S., Cardoso, F., Ayoub, O., Giordano, S., & Luceri, L. (2022). The disinformation dozen: An exploratory analysis of covid-19 disinformation proliferation on twitter. *Proceedings of the 14th ACM Web Science Conference 2022*, 348–358.

Norton, B. G. (2005). *Sustainability: A philosophy of adaptive ecosystem management*. University of Chicago Press.

Novaes, C. D., & De Ridder, J. (2021). Is fake news old news. *The Epistemology of Fake News*, 156–179.

Nye, J. S. (1990). Soft power. *Foreign Policy*, *80*, 153–171.

O'brien, S. P. (2010). Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, *12*(1), 87–104.

Olteanu, A., Diaz, F., & Kazai, G. (2020). When are search completion suggestions problematic? *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–25.

Ördén, H. (2022). Securitizing cyberspace: Protecting political judgment. *Journal of International Political Theory*, *18*(3), 375–392.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., & others. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Paine, R. T. (1969). A note on trophic complexity and community stability. *The American Naturalist*, *103*(929), 91–93.

Palca, J. (1988). US anger over accusations of trafficking in infant organs. *Nature*, *335*(6193), 754–754. https://doi.org/10.1038/335754a0

Paul, C., & Matthews, M. (2016). *The Russian Firehose of Falsehood Propaganda Model: Why It Might Work and Options to Counter It*. RAND Corporation. https://doi.org/10.7249/PE198

Pearson, C., & Moxham, J. (2022). Reclaiming the Narrative: The US and International Communications. *Potomac Institute For Policy Studies*.

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.

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

Pham, A., Rubel, A., & Castro, C. (2022). Social media, emergent manipulation, and political legitimacy. In *The Philosophy of Online Manipulation* (pp. 353–369). Routledge.

Postman, N. (1970). The reformed english curriculum. In *High school 1980: The shape of the future in American secondary education*. Pitman Pub. Corp.

Pritchard, D. (2006). Moral and epistemic luck. *Metaphilosophy*, *37*(1), 1–25.

Pritchard, D. (2013). Epistemic paternalism and epistemic value. *Philosophical Inquiries*, *1*(2), 9–37.

Rafee, A. A. (2020). Polarization on Social Media Platforms Consequences for Politics and Security. *The Digital Age, Cyber Space, and Social Media The Challenges of Security & Radicalization*, 173.

Raguso, R. A. (2008). Wake up and smell the roses: The ecology and evolution of floral scent. *Annual Review of Ecology, Evolution, and Systematics*, *39*, 549–569.

Ranalli, C. (2020). Deep disagreement and hinge epistemology. *Synthese*, *197*(11), 4975–5007.

Raz, J. (1986). *The morality of freedom*. Clarendon Press.

Reding, D. F., & Eaton, J. (2020). *Science and Technology Trends 2020-2040: Exploring the S and T Edge* (NATO Science & Technology Organization, pp. 71–73). https://apps.dtic.mil/sti/citations/AD1131124

Ribeiro, H. M., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., & West, R. (2021). Do platform migrations compromise content moderation? Evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–24.

Rimbert, P., & Halimi, S. (2022, September 1). *News we don't want to hear*. Le Monde Diplomatique. https://mondediplo.com/2022/09/08ukraine-media

Rini, R. (2017). Fake News and Partisan Epistemology. *Kennedy Institute of Ethics Journal*, *27*(2), 43–64. https://doi.org/10.1353/ken.2017.0025

Rini, R. (2020). Deepfakes and the Epistemic Backstop. *Philosophers' Imprint*, *20*(24), 1–16.

Rini, R. (2021). Weaponized Skepticism. *Political Epistemology*, 31.

Rini, R. (forthcoming). Context collapse and pop-up communities: How social media makes its own norms. In P. Connolly, S. Goldberg, & J. Saul (Eds.), *Conversations Online*. Oxford University Press.

Robbins, J. (2020). Countering Russian Disinformation. *Center for Strategic and International Studies*, *23*.

Robertson, D., & Bronskill, J. (2023, September 2). *A new CSIS ad campaign is using Soviet-style imagery to warn Canadians about disinformation | CBC News*. CBC. https://www.cbc.ca/news/politics/csis-disinformation-ads-soviet-imagery-1.6955717

Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, *35*(3), 213–229.

Rogers, Z. (2021). The Promise of Strategic Gain in the Digital Information Age. *The Cyber Defense Review*, *6*(1), 81–106.

Rogers, Z. (2020, June 18). *The End of Information Warfare?* Modern War Institute. https://mwi.westpoint.edu/end-information-warfare/

Ronfeldt, D., & Arquilla, J. (2020). The continuing promise of the noosphere and noopolitik: 20 years after. In *Routledge Handbook of Public Diplomacy* (pp. 445–480). Routledge.

Ronzaud, L., Stubbs, J., & Williams, T. (2022, November). *Same Schmitz, Different Day*. Graphika. https://graphika.com/posts/same-schmitz-different-day

Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How Accurate Are Accuracy-Nudge Interventions? A Preregistered Direct Replication of Pennycook et al. (2020). *Psychological Science*, *32*(7), 1169–1178. https://doi.org/10.1177/09567976211024535

Rothschild, M. (2021). *The storm is upon us: How QAnon became a movement, cult, and conspiracy theory of everything*. Melville House.

Russell, V. S. (1974). Pollution: Concept and definition. *Biological Conservation*, *6*(3), 157–161. https://doi.org/10.1016/0006-3207(74)90060-3

Russo, G., Verginer, L., Horta Ribeiro, M., & Casiraghi, G. (2023). Spillover of Antisocial Behavior from Fringe Platforms: The Unintended Consequences of Community Banning. *Proceedings of the International AAAI Conference on Web and Social Media*, *17*, 742–753. https://doi.org/10.1609/icwsm.v17i1.22184

Ryan, S. (2018). Epistemic environmentalism. *Journal of Philosophical Research*, *43*, 97–112.

Schlegel, F. von. (1798). *Philosophical fragments*. University of Minnesota Press.

Schreckinger, B. (2016, September 30). *Inside Trump's "cyborg" Twitter army*. POLITICO. https://www.politico.com/story/2016/09/donald-trump-twitter-army-228923

Scolari, C. A. (2012). Media Ecology: Exploring the Metaphor to Expand the Theory. *Communication Theory*, *22*(2), 204–225. https://doi.org/10.1111/j.1468-2885.2012.01404.x

Sessions, G. (1995). Deep ecology for the twenty-first century. *Boston: Shambhala*.

Shiffrin, S. V. (2000). Paternalism, unconscionability doctrine, and accommodation. *Philosophy & Public Affairs*, *29*(3), 205–250.

Shir-Raz, Y., Elisha, E., Martin, B., Ronel, N., & Guetzkow, J. (2023). Censorship and Suppression of Covid-19 Heterodoxy: Tactics and Counter-Tactics. *Minerva*, *61*(3), 407–433. https://doi.org/10.1007/s11024-022-09479-4

Simion, M. (2023). Knowledge and disinformation. *Episteme*, 1–12.

Smith, C. B. (2023). The Semantic Attack Surface: A Systems-Dynamic Model of Narrative in Cyberspace. *IEEE Transactions on Technology and Society*, *4*(2), 146–157. https://doi.org/10.1109/TTS.2022.3210782

Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical Perspectives*, *13*, 141–153.

Springer, A. L. (1977). Towards a meaningful concept of pollution in international law. *International & Comparative Law Quarterly*, *26*(3), 531–557.

Sproule, J. M. (2005). *Propaganda and democracy: The American experience of media and mass persuasion*. Cambridge University Press.

Srinivasan, A. (2020). Radical Externalism. *The Philosophical Review*, *129*(3), 395–431. https://doi.org/10.1215/00318108-8311261

Stanley, J. (2015). *How propaganda works*. Princeton University Press.

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–26. https://doi.org/10.1145/3359229

Sterelny, K. (2003). *Thought in a hostile world: The evolution of human cognition*. Blackwell.

Sterelny, K. (2006). Cognitive load and human decision, or, three ways of rolling the rock up hill. *The Innate Mind Volume 2: Culture and Cognition*, 217–233.

Sundelius, B., & Eldeblad, J. (2023). Societal Security and Total Defense. *PRISM*, *10*(2), 92–111.

Sunstein, C. R. (1992). Free Speech Now. *The University of Chicago Law Review*, *59*(1), 255–316.

Sunstein, C. R. (2005). *Laws of fear: Beyond the precautionary principle.* Cambridge University Press.

Sunstein, C. R. (2018). "Better off, as judged by themselves": A comment on evaluating nudges. *International Review of Economics*, *65*, 1–8.

Supran, G., & Oreskes, N. (2021). Rhetoric and frame analysis of ExxonMobil's climate change communications. *One Earth*, *4*(5), 696–719. https://doi.org/10.1016/j.oneear.2021.04.014

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*, *9*(3), 286–299.

Szymczak, R. (2010). Cold War Crusader: Arthur Bliss Lane and the Private Committee to Investigate the Katyn Massacre, 1949-1952. *Polish American Studies*, *67*(2), 5–33.

Tappin, B. M., Wittenberg, C., Hewitt, L., berinsky, adam, & Rand, D. G. (2022). *Quantifying the Potential Persuasive Returns to Political Microtargeting* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/dhg6k

Taraborelli, D. (2008). How the Web is changing the way we trust. *Current Issues in Computing and Philosophy*, 194–204.

Tenove, C. (2020). Protecting Democracy from Disinformation: Normative Threats and Policy Responses. *The International Journal of Press/Politics*, *25*(3), 517–537. https://doi.org/10.1177/1940161220918740

Teperik, D., Denisa-Liepniece, S., Bankauskaitė, D., & Kullamaa, K. (2022). Resilience Against Disinformation: A New Baltic Way to Follow? *International Centre for Defense and Security*, *20*.

Ternovski, J., Kalla, J., & Aronow, P. M. (2021). *Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments* [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/dta97

Terp, S., & Breuer, P. (2022). DISARM: A Framework for Analysis of Disinformation Campaigns. *2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, 1–8. https://doi.org/10.1109/CogSIMA54611.2022.9830669

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *American Economic Review*, *93*(2), 175–179.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness* (Rev. and expanded ed., with a new afterword and a new chapter). Penguin.

Thomas, D. R., & Wahedi, L. A. (2023). Disrupting hate: The effect of deplatforming hate organizations on their online audience. *Proceedings of the National Academy of Sciences*, *120*(24), e2214080120.

Tollefson, J. (2023). Disinformation researchers under investigation: What's happening and why. *Nature*, d41586-023-02195–3. https://doi.org/10.1038/d41586-023-02195-3

Tomz, M., & Weeks, J. L. P. (2020). Public Opinion and Foreign Electoral Intervention. *American Political Science Review*, *114*(3), 856–873. https://doi.org/10.1017/S0003055420000064

Tripodi, F. (2022). *The propagandists' playbook: How conservative elites manipulate search and threaten democracy*. Yale University Press.

Tucker, J., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3144139

Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232.

Varner, G. (2020). Response to Millstein. *Biology & Philosophy*, *35*, 1–8.

Verbeek, P.-P. (2015). Toward a theory of technological mediation. *Technoscience and Postphenomenology: The Manhattan Papers*, *189*.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Walker, C. (2018). What Is “Sharp Power”? *Journal of Democracy*, *29*(3), 9–23. https://doi.org/10.1353/jod.2018.0041

Wall, T. (2010). U.S. Psychological Warfare and Civilian Targeting. *Peace Review*, *22*(3), 288–294. https://doi.org/10.1080/10402659.2010.502070

Waltzman, R. (2017). *The Weaponization of Information: The Need for Cognitive Security*. Testimony presented before the Senate Armed Services Committee, Subcommittee on Cybersecurity on April 27, 2017. https://www.armed-services.senate.gov/imo/media/doc/Waltzman_04-27-17.pdf

Wanless, A., & Pamment, J. (2019). How Do You Define a Problem Like Influence? *Journal of Information Warfare*, *18*(3), 1–14.

Wardle, C., & Derakhshan, H. (2017a). *Information disorder: Toward an interdisciplinary framework for research and policy making.* Council of Europe.

Wardle, C., & Derakhshan, H. (2017b). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Vol. 27). Council of Europe Strasbourg.

Weinberg. (1968). What to Tell America: The Writers’ Quarrel in the Office of War Information. *The Journal of American History*, *55*(1), 73. https://doi.org/10.2307/1894252

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, *29*(1/2), 429–460.

Weissmann, M., Nilsson, N., Palmertz, B., & Thunholm, P. (2021). *Hybrid warfare: Security and asymmetric conflict in international relations*. Bloomsbury Academic.

Wigell, M., Mikkola, H., & Juntunen, T. (2021). Best Practices in the whole-of-society approach in countering hybrid threats. *European Parliament*.

Wittgenstein, L., Anscombe, G. E. M., & Wright, G. H. von. (1969). *On certainty*. Blackwell.

Wright, C. (212 C.E.). Replies Part IV: Warrant Transmission and Entitlement. In A. Coliva (Ed.), *Mind, Meaning, and Knowledge: Themes from the Philosophy of Crispin Wright.* Oxford University Press.

Wright, C. (2014). *On epistemic entitlement (II): Welfare state epistemology*.

Yapp, W. B. (1972). *Production, pollution, protection*. Wykeham.

Yu, M. T.-C., & Ho, K. (2023). COVID and Cognitive Warfare in Taiwan. *Journal of Asian and African Studies*, *58*(2), 249–273.

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, *395*(10225), 676.

Ziolkowski, A. (2016). Folk intuitions and the no-luck-thesis. *Episteme*, *13*(3), 343–358.

Zuber, N., Gogoll, J., Kacianka, S., Pretschner, A., & Nida-Rümelin, J. (2022). Empowered and embedded: Ethics and agile processes. *Humanities and Social Sciences Communications*, *9*(1), 1–13.