The Alan Turing Institute

Defence and Security Programme

Tackling threats to informed decision-making in democratic societies

Promoting epistemic security in a technologically-advanced world

[dstl] The Science Inside

CENTRE FOR THE STUDY OF EXISTENTIAL RISK

Tackling threats to informed decision-making in democratic societies

Promoting epistemic security in a technologically-advanced world

Authors

Elizabeth Seger[1]
Leverhulme Centre for the Future of Intelligence, University of Cambridge
Department of History and Philosophy of Science, University of Cambridge

Shahar Avin
Centre for the Study of Existential Risk, University of Cambridge

Gavin Pearson
Defence Science and Technology Laboratory, Ministry of Defence

Mark Briers
The Alan Turing Institute

Seán Ó Heigeartaigh
Centre for the Study of Existential Risk, University of Cambridge
Leverhulme Centre for the Future of Intelligence, University of Cambridge

Helena Bacon
Defence Science and Technology Laboratory, Ministry of Defence

[1] Corresponding author eas97@cam.ac.uk

Contributor list (in alphabetical order)

Henry Ajder (Head of Threat Intelligence, Deeptrace); **Claire Alderson** (Policy Officer, Ministry of Defence); **Fergus Anderson** (Principal Advisor, Ministry of Defence); **Joseph Baddeley** (Department for Digital, Culture, Media and Sport); **Craig Bakker** (Research Scientist, National Security Directorate, Pacific Northwest National Laboratory); **Carla Zoe Cremer** (Research Scholar, Future of Humanity Institute, University of Oxford); **Eric Drexler** (Senior research fellow, Future of Humanity Institute, University of Oxford); **James Eaton-Lee** (Head of Information Security, Oxfam GB); **Daniel Edwards** (Office for AI, DCMS and BEIS); **Philip Gibson** (Defence Science and Technology Laboratory); **Robert (Bob) Hobbs** (Lieutenant Colonel, British Army); **Sophia Ignatidou** (Academy Associate at Chatham House, International Security Programme); **Peter Johnson** (Professor of Computer Science and Department Head, University of Bath); **De Kai** (Professor of Computer Science and Engineering, HKUST, Distinguished Research Scholar, ICSI); **Dinos A. Kerigan-Kyrou (**Instructor NATO DEEP Partnership for Peace Consortium of Defense Academies); **Colin Roberts** (Senior Research Fellow, Crime and Security Research Institute, Cardiff University); **Darren Rockett** (Defence Science and Technology Laboratory); **Mark Round** (Principal Data Scientist, QinetiQ); **Samuel Scott** (Defence Science and Technology Laboratory); **Paul Stanley** (Head of Analysis and Quality, SVGC); **Alex Stevens** (Defence Science and Technology Laboratory); **Paul Strong** (Defence Science and Technology Laboratory); **Ann Stow** (Defence Science and Technology Laboratory); **Neil Verrall** (Defence Science and Technology Laboratory); **Adrian Weller** (Principal Research Fellow, Leverhulme Centre for the Future of Intelligence, University of Cambridge; Programme director of AI, The Alan Turing Institute)

Executive Summary

Access to reliable information is crucial to the ability of a democratic society to coordinate effective collective action, especially when responding to crises such as global pandemics, and complex challenges such as climate change. We define an *epistemically secure* society as one that reliably averts threats to the processes by which reliable information is produced, distributed, acquired and assessed within the society.

Citizens of contemporary, technologically rich societies have greater access to information than at any point in history. However, while new technologies make information more widely accessible, information abundance and other changes brought about by new technologies highlight a different set of threats and vulnerabilities in our systems of information production and exchange. We identify the following themes:

1. Adversaries and blunderers can more readily interfere with decision-making processes, through [dis/mis]information or other harmful actions than in the past.
2. Information abundance means the attention of information recipients is spread thin, making it harder to ensure essential information reaches all important parties. This leads to an attention economy in which tradeoffs are made between the truth-orientation of information and attention-grabbing strategies.
3. Insular communities that reject information that challenges their accepted views quickly emerge and persist. Strong in-group identity leads to greater polarisation between groups.
4. Information mediating and producing technologies make it more difficult to evaluate the trustworthiness of individual information sources.

Through a series of workshops we developed and analysed a set of hypothetical yet plausible crisis scenarios to explore how external threats and internal vulnerabilities to epistemic security can be mitigated in order to facilitate timely decision-making and collective action in democratic societies. Overall we observed that preserving a democratic society's epistemic security is a complex effort that sits at the interface of many knowledge domains, theoretical perspectives, value systems, and institutional responsibilities.

Consequently, challenges to epistemic security cannot be addressed as a laundry list of threats with narrowly targeted fixes. To do so may cause more harm than good because societies can suffer from multiple interconnected threats and vulnerabilities, and proposed solutions to each can have unintended first-, second-, and higher-order consequences. Epistemic threats are therefore best considered via a holistic and interdisciplinary approach that takes into account

the broader socio-technological contexts in which the threats have emerged. We developed the following recommendations to highlight areas where additional research and resources will likely have a significant impact on epistemic security in democratic societies:

1. Develop technological and institutional methods to increase the cost for adversaries and blunderers in spreading unsupported, fabricated, or false information.
2. Develop methods of helping information consumers more easily identify trustworthy information sources.
3. Explore technological and institutional methods to "signal boost" reliable decision-relevant information in an asymmetric manner. Recognize that evaluations of what constitutes reliable and decision-relevant information will most often benefit from the input of diverse communities and interest groups.
4. Develop technological and institutional methods to monitor changes in epistemic systems and to rapidly detect adversarial epistemic action during times of tension or crises.
5. Build capacity for and engage in holistic systems-mapping procedures (constructing an integrated view of social epistemic systems) and red-teaming strategies (deliberately exploring a scenario from an adversary's perspective) to help identify and analyse epistemic threats.
6. Establish working relationships with a diverse array of experts who are experienced in identifying and analysing epistemic threats and who could serve as epistemic security advisors before and during crises.
7. Invest in building and curating diverse and multidisciplinary epistemic security research groups and expert networks.

We provide a more extensive discussion of recommendation 5 in the final section of the report. We describe how holistic systems-mapping and red-teaming strategies might be implemented to better understand complex social epistemic systems and to help identify and analyze epistemic threats using examples from our workshop proceedings to illustrate.

Contents

1. Introduction

The capacity of a democratic society for timely decision-making and for organizing collective action is crucial to navigating crises and complex challenges. Examples of such crises and challenges include voting out officials who no longer serve the public interest, responding to impending natural disasters, and eliminating or halting the spread of a disease through vaccination.

Given the dispersed nature of democratic processes, the capacities for timely decision-making and collective action are easily undermined by disrupting the processes by which information is gathered, distributed, and assessed by decision-making bodies and by the public. If there is no shared belief among the actors in a community about the nature of a crisis or the efficacy of a proposed response, collective action is less likely to ensue.

We call detrimental interferences to systems of information production and dissemination *epistemic threats*. Epistemic threats include blatant censorship efforts or misinformation campaigns, the erosion of trust in expertise, the formation of insular communities, the suppression of diverse viewpoints and marginalized voices, and so on. An *epistemically secure* society is one that is robust to such threats. In this report we pay special attention to technologically-enabled epistemic threats to systems of information production and dissemination and to technologically-exacerbated vulnerabilities within those systems.

While new technologies aid in the production and dissemination of decision-guiding information, they can also enable and exacerbate threats to production and dissemination of reliable information. Consider the following hypothetical scenario drawn from our workshop proceedings (Workshop scenario 5: Xenophobic Ethnic Cleansing) in which information technologies are used to exacerbate a crisis and undermine effective response.

- A radical xenophobic group decides to turn the population of their country against a minority community of recently-arrived refugees.
- They produce a low-grade chemical weapon, and release it near a school in a poverty-stricken suburb of a major city, taking several videos of the operation.
- They edit the videos with face-swapping apps to make it appear as if a recognised figure from the refugee community was the perpetrator.
- They release the videos as "breaking news" on right-leaning social media groups and through various messaging apps. Social media posts are targeted to specific individuals

and groups and modified to most effectively agitate the audience given personality traits and predispositions inferred from social media profiles.

- The extremist group also releases messages saying "the government is going to cover this up and blame it on some scapegoat to avoid scrutiny of their reckless immigration policy".
- When officials reliably do respond to the attack and rumours, and question the validity of the videos, the radicals spread a call to "all patriotic citizens" to "carry out justice and drive out the terrorists". They provide specific times and places to gather and lists of targets (refugee shelters, businesses owned by refugees, government buildings housing refugee affairs offices, etc.)
- Different officials make rushed or contradicting statements, some calling for calm and patience while investigations are ongoing, while others promise quick action and swift resolution.
- Individuals within law enforcement who are sympathetic to the xenophobic cause break rank and publicly criticise the government for a slow and hesitant response, and (anonymously) offer information and assistance in circumventing law enforcement.
- Shocked, afraid and angry mobs rally to the call and carry out vandalism and, on some occasions, assault.

Crisis scenarios such as the one above highlight how interference with the dissemination of reliable information can compromise decision-making and collective action efforts. The same point is illustrated by the present COVID-19 pandemic and the accompanying ‘infodemic’ in which inaccurate information and the silencing of important information sources have degraded trust in health authorities and slowed public response to the crisis (Hubert et. al. 2020, WHO 2020, Reviving the US CDC, 2020). Misinformation about ineffective cures, the origins and malicious spread of COVID-19, unverified treatment discoveries, and the efficacy of face coverings have increased the difficulty of coordinating unified public response during the crisis (Jourova, 2020).

This report is long in the making, during which time the COVID-19 crisis has unfurled. Throughout the crisis we have found it useful to think about the pandemic through the lens of epistemic security we develop here. While it is unlikely the contents of this report approach a solution to managing infodemics such as that which has slowed the response to COVID-19, we hope that the recommendations we present for the promotion of epistemically secure democracies will help us be more resilient to similar events in the future.

1.1 The epistemic security workshops

In 2018, when this project started, much research had already been conducted into specific technologically-enabled threats and vulnerabilities (see section 2.2). However, less attention has been paid to how these threats ultimately affect a democratic society's capacity for collective decision-making and action, and more practically, how decision-making bodies can take steps to mitigate these threats and vulnerabilities. This gap in practical research led the authors to conduct a series of workshops in 2018 and 2019 exploring how technological challenges (epistemic threats) to the production and distribution of reliable information affect a democractic society's capacities for collective decision-making and action and how these threats can be practically addressed.[2] The workshops were held in collaboration between the Centre for the Study of Existential Risk at the University of Cambridge, the Defence Science and Technology Laboratory of the UK Ministry of Defence, and the Alan Turing Institute.

The hypothetical scenario outlined above was one of six explored in the workshops to guide our investigation. Other workshop scenarios are presented as 'blue box' examples throughout the report.

We found hypothetical crisis scenarios[3] to be useful tools for appraising threats and vulnerabilities to social epistemic systems. Factors that influence a society's epistemic security are not always obvious when life is relatively tranquil but are highlighted under the stress of a crisis. By exploring hypothetical crisis scenarios we aim to identify real threats and vulnerabilities and address them ahead of time in order to prevent a crisis or to make our response more effective.

The challenges and strategies discussed at the workshops demonstrated the need to take a holistic systems approach to epistemic security, such as systems-mapping procedures (constructing an integrated view of social epistemic systems) and red-teaming strategies (deliberately exploring a scenario from an adversary's perspective). Decision-making bodies can be thought of as distributed socio-technical information systems which operate within a broader information environment. The workshops highlighted that it is unlikely to be effective to treat technologies that exacerbate or pose new threats to epistemic security as a laundry list of independent threats each with a prescribed fix. Societies can suffer from multiple interconnected threats and vulnerabilities, and proposed solutions to each may have unintended first-, second-, and higher-order consequences.

[2] A summary of direct outputs from the workshop series is presented in Appendix 2.
[3] See Appendix 1 for a full list of hypothetical scenarios developed and discussed in the workshop series.

1.2 Report Overview

In section 2 we provide background information about epistemology and epistemic security, present a brief overview of related works, and discuss the scope and restrictions of this report. In particular, we attend to the inherently intertwined nature of value systems and information systems in a democracy and acknowledge the difficulties this presents for conducting a focussed discussion on vulnerabilities in epistemic infrastructure.

In section 3 we present concepts and categories to help analyse epistemic security as the challenge of building and maintaining a robust social epistemic infrastructure which, in turn, enables well-informed decision-making and timely collective action in a society. We explore four broad themes that can lead to epistemic vulnerabilities, linking them to areas where emerging technologies pose heightened risk or present promising opportunities. A more detailed categorisation of specific threats from emerging technologies is presented in Appendix 3, and Appendix 4 proposes a preliminary model for a quantitative analysis of threats to epistemic security.

In section 4 we present preliminary recommendations for tackling the challenges of epistemic security, highlighting the importance of avoiding narrowly-targeted quick-fixes and the need to build a diverse and robust community of experts to tackle these challenges in a context-specific manner.

Building on these recommendations, in section 5 we emphasize holistic "systems-mapping" and "red-teaming" strategies as promising methodologies for identifying, assessing and mitigating challenges of epistemic security. We describe how these strategies were used throughout our workshops to appraise hypothetical crisis scenarios like that illustrated in the introduction. Further information about the workshops and scenarios developed is presented in Appendices 1 and 2.

We conclude with final remarks in section 6.

2. Background

2.1 Why "Epistemic" Security?

Epistemology is the branch of philosophy that deals with the nature and processes of knowledge.[4] *Epistemic processes* are the processes by which information is produced, distributed, acquired and assessed by individuals and within social communities (see Figure 1).

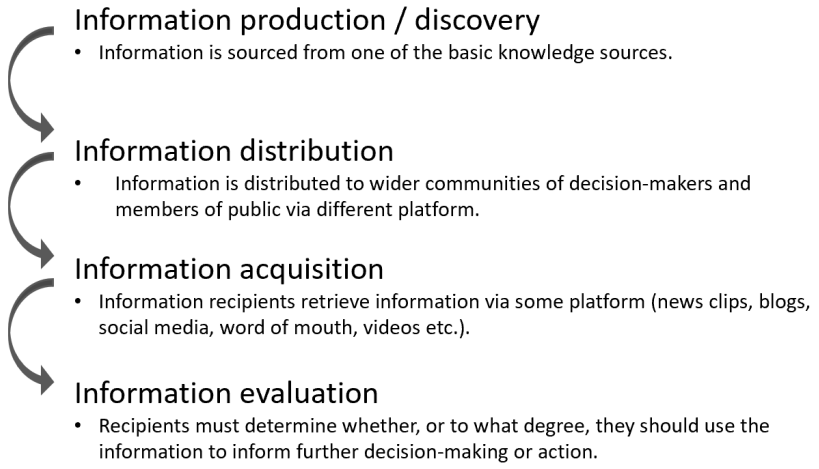


Figure 1: Epistemic processes

At a social level, epistemic processes can become highly complex. We use the term *social epistemic infrastructure* to refer to the vast collection of systems, artefacts, and actors that influence a society's epistemic processes.[5] Academic departments and think tanks, artists,

[4] Epistemologists traditionally understand knowledge as justified true beliefs, and much ink has been spilled debating the precise conditions required for a person to 'be justified' in their beliefs. However, in this report we focus less on an individual's knowledge or beliefs, and focus instead on the complex network of epistemic processes enroute to knowledge, that impact a society's ability to make decisions and to organise timely collective action, especially in response to crises and complex challenges. Critical decisions in times of crisis often must be made in the relative absence of knowledge and instead must be based on the best available information. Ideally, decision-makers will base their decisions on true information from reliable sources, but, as we outline in this report, various factors can interfere with the production and dissemination of reliable information. Therefore, for the purpose of this report we will set aside further reference to knowledge and instead speak in terms of decision-guiding information. As true information is a necessary, though not sufficient, condition for knowledge, interferences in the spread and uptake of true information are also threats to knowledge, and therefore form a good starting point to explore epistemic security.

[5] Infrastructure has also been used in epistemology of journalism to describe technologically-enabled systems that facilitate or constrain news flows (Carlson 2020). In history of science, epistemic infrastructure also refers to the collections of knowledge curated in museums, libraries, archives, zoos etc. Epistemic infrastructures organize knowledge so that it can be easily accessed and used to inform

experts, government departments and organisations (including militaries and intelligence agencies), interest groups, traditional and social media platforms, public communities, activists, corporations, financial markets, and various information and communication technologies all play overlapping roles (sometimes complimentary, sometimes antagonistic) influencing how information is produced, modified, evaluated and distributed within the society.

In general, the variety of social epistemic infrastructures has been of great benefit to human communities. Research in collective intelligence and group cognition shows that large scale collaboration between diverse groups of people is essential to scientific discovery and technological advancement (Anderson & Wagenknecht 2013; Malone 2018; Wray 2002). When factions of a community are in disagreement, a social environment of critique plays a key role in maintaining accountability and high epistemic standards for the production and dissemination of reliable information (Longino 2002; Winsberg, Huebner & Kukla 2014).

However, while large, diverse, interconnected societies are epistemically advantageous, intricate social epistemic infrastructures are also more vulnerable to epistemic threats or risks - risks of error arising anywhere in a society's processes of information acquisition, distribution and evaluation (Biddle & Kukla 2017, p.218). The more complicated a social epistemic infrastructure is, the more opportunities there are for accidental or intentional disruption to a society's epistemic processes, preventing the society from reliably yielding and dealing in true information.

As illustrated in the 'Xenophobic Ethnic Cleansing' scenario and the COVID-19 pandemic, threats to a society's production and distribution of true information can be severely detrimental to its capacity for timely and well-informed decision-making and collective action. Therefore it is important for a society to take steps to strengthen itself against epistemic threats - influencing factors that interfere with the well-functioning of a society's epistemic processes - and minimize its epistemic vulnerabilities - weak points in a social epistemic infrastructure that are most likely to succumb to epistemic threats. An epistemically secure society is one that reliably averts epistemic threats and minimizes vulnerabilities in its social epistemic infrastructures.

Accordingly, we also use the term *epistemic security* as a holistic umbrella for investigations into the processes by which societies produce, distribute, evaluate and assimilate information, and into threats that restrict access to information, or undermine our ability to evaluate information veracity or information source reliability. We are particularly concerned with how

---

further action or investigation. (Hedstrom 2005). Our use of epistemic infrastructure is broader and encapsulates both of these uses.

threats to epistemic security undermine a society's ability to make well-informed and timely decisions and to coordinate action in response to crises.

2.2 Related Work

Numerous recent scholars point to increasing evidence that the systems of information production, evaluation, and distribution that inform collective action in contemporary liberal democracies are compromised, in large part due to recent social changes and technological innovations. In *The Misinformation Age* (2019), O'Connor and Weatherall explore numerous social factors that contribute to misinformation. Both *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media* (2018) by Woolley and Howard and *Lie Machines* (2020) by Howard detail how technological advances are used by political and private actors from around the globe, to confuse and control collective attention with detrimental effects to collective decision-making and trust in democracy. In *The existential threat from cyber-enabled information warfare,* Lin (2019) describes how technologically-enabled [dis/mis]information poses an existential threat to key pillars of democracy, and in *Common-knowledge attacks on democracy,* Schneier and Farell (2018) explain that democracies are disproportionately more vulnerable to [dis/mis]information attacks than autocracies. Similar themes are explored in Runciman's (2018) *How Democracy Ends* and Pomerantsev's (2019) *This is Not Propaganda: Adventures in the War Against Reality* from a political science and cultural analysis perspective, as well as in *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics,* by Benkler, Faris and Roberts (2018). Bucher studies the effect of algorithms and informational infrastructures on social life in *If...Then: Algorithmic Power and Politics* (2018), while Vaidhyanathan in *Anti-social media - How Facebook Disconnects Us and Undermines Democracy*, (2018) focusses on one platform in particular. Hwang (2019) explores similar themes in *Maneuver and Manipulation* from a military strategy perspective.

Several government and institutional reports also investigate the influence of technology on social epistemic systems. Examples include a RAND Corporation research report titled *The Emerging Risk of Virtual Societal Warfare - Social Manipulation in a Changing Information Environment* (Mazarr et al. 2019), a NATO STRATCOM report that reviews *Government Responses to Malicious Use of Social Media* (Bradshaw, Neudert & Howard 2018), and the European Commission's (2018) report on fake news and online disinformation titled *a multi-dimensional approach to disinformation*.

A special role in enhancing the capabilities of propagandists and misinformers has been ascribed to artificial intelligence, for example in Chessen's (2017) *The MADCOM Future*,

Kertysova's (2018) *Artificial Intelligence and Disinformation,* and in the section on political security of Brundage & Avin *et al.*'s (2018) *The Malicious Use of Artificial Intelligence*.

Particular attention has been paid to synthetic media, including "deep fakes", as discussed by Chesney and Citron (2018) and in reports from Deeptrace Labs (Ajder et. al. 2019), the International Risk Governance Center (Collins 2019), and the Center for Security and Emerging Technology (Hwang 2020). The growing attention to this cluster of issues has led to a UK government inquiry into disinformation and fake news (House of Commons Digital, Culture, Media and Sport Committee, 2019) and proposed legislation on online harms, and a US congressional hearing (U.S. House of Representatives, 2019) on deep fakes and several proposed bills on the issue.

There is ongoing work on this cluster of issues at various universities including at the University of Oxford by the "Computational Propaganda" project[6], at the RAND Corporation under the heading "Truth Decay"[7], by the DARPA funded "Media Forensics" project[8], and by independent research institutions such as Data & Society[9], the Thoughtful Technology Project[10], and the Centre for Humane Technology[11]. MisinfoCon[12] also provides a publication venue and conference for related issues, and the Credibility Coalition[13] coordinates various organisations in this space.

The above review only scratches the surface of a mushrooming response to what we call challenges of epistemic security. Activity in this space, both within and outside academia, has expanded significantly during the period in which this report has been compiled. We hope our workshop findings - arrived at via a more holistic perspective and through the application of different methodologies - make a useful contribution to the discussion by situating a collection of related concerns into a broader epistemic framework and by proposing a method for working through scenarios to work towards robust policy recommendations.

2.3 Focus and limits of this report

Social epistemic infrastructures are highly complex, and therefore the ways in which they can be interfered with or manipulated are numerous - well beyond the scope of what could be

[6] https://comprop.oii.ox.ac.uk/
[7] https://www.rand.org/research/projects/truth-decay.html
[8] https://www.darpa.mil/program/media-forensics
[9] https://datasociety.net/
[10] https://thoughtfultech.org/; https://digitalfuturesociety.com/qanda/aviv-ovadya/
[11] https://www.humanetech.com/
[12] https://misinfocon.com/
[13] https://credibilitycoalition.org/

covered in a single workshop series or report. To make the scope more tractable, we have restricted our focus in two key ways.

First, we focus primarily on epistemic security in affluent, technology-rich liberal democracies in which crisis aversion requires collective decision-making and action, and on crises and challenges with a potential to cause society-wide harms. This focus implies certain assumptions and limitations. See section 2.3.1 for an extended discussion.

Second, we restrict ourselves to a descriptive analysis to how information is produced, distributed, and consumed by distributed decision-makers. We do not focus on the various ways in which decision-maker values can influence preferences and incentives and shape social epistemic infrastructures. Though individual and group values are inseparably intertwined with, and have a significant influence on, collective decision making (especially when the values held by distributed decision-makers clash) we find it useful to artificially bracket them out of our conversation for the time being. We direct readers concerned with our delineation between epistemic systems and value systems to section 2.3.2 where we expand on the limitations of our approach and explain further our decision to proceed in this way.

2.3.1 Affluent, technology-rich liberal democracies

Our focus on affluent, technology-rich liberal democracies, such as the United States of America, the United Kingdom, many members of the European Union, and many others, is motivated in part by the remarkable impact of digital technologies on information processes in those states (both positive and negative), by the essential role of distributed decision-making in these systems of governance, and by the locality of the organisations contributing to this report, all based in the UK.

For our investigation of epistemic security, this focus implies certain assumptions:

- In a democratic society power resides ideally (or at least partially and comparatively) with the citizens. Decision-making takes a distributed form. The power of citizens is expressed via direct voting, representation or pressure placed on representatives via the withdrawal of public approval of their policies. Citizens engage in a deliberation process and feed their personal decision into an aggregation mechanism. (e.g. in processes of elections or referenda).
- Democracies place restrictions on state action to guarantee various freedoms, including freedom of speech and freedom of association. These regulations encourage the development of numerous and diverse information sources.

- By their affluent state, we assume a significant portion of the citizenry is literate and educated; that information consumption and generation in large volumes plays a key role in work, leisure and social lives; and that there can be reasonable expectations from the state or between citizens about time and resources individuals have to invest in gaining more information.
- By technology-rich we assume the majority of citizens have access to reliable high-bandwidth internet, smartphones with video cameras, and that many engage with social media platforms.

We should also note that within this report we intentionally adopt a government actor's perspective to issues of epistemic security. That is, we choose to "see like a state".[14] However, it is important to acknowledge that state priorities or convenience may not always align with the interests of the populace. Entities responsible and/or incentivised to ensure epistemic security will not always be government actors - indeed, at times government powers may, intentionally or otherwise, undermine epistemic security[15] - and we acknowledge our present analysis suffers by not specifically addressing the perspective of non-government actors as well. Our focus on the state level overlooks issues that may arise at different levels of scale (e.g. on smaller scales within communities or industries or at larger scales at the regional and global levels), and does not explore in depth instances in which non-government entities are the primary guardians of epistemic security. We hope future work building on, responding to, or inspired by this report will extend our analysis beyond the current limitations we have imposed.

2.3.2 Clashing values and informed collective action

In the remainder of this report we also largely set aside the complex and important issue of how to navigate clashing values and preferences in a heterogeneous polity. We acknowledge that the separation between information and values is artificial but for the purpose of this report we find it useful to discuss information independently. Here we explain why.

It is important to acknowledge that even if the constituents of a decision-making body have equitable access to information, it may still be reasonable for the constituent individuals or groups to rationally disagree about how to proceed based on that information;[16] the values held

[14] As described by James C. Scott (1998) an entity 'sees like a state' when it develops schemes or proposes solutions primarily with a view to state convenience.
[15] For example, the September 2002 dossier published by the British Government presented an assessment of Iraq's development of weapons of mass destruction. It has been uncovered that some assessments made in the report were framed with the goal of strengthening the case for war with Iraq (Aimes, 2011). Assessments made in the dossier regarding Iraq's WMD capabilities and arsenals were subsequently found to be incorrect (Comprehensive report of, 2004).
[16] Social epistemologists who write on this topic discuss the epistemic permissibility of rational disagreement (Conee 2010, Goldman 2010, Kelp & Douven 2012).

by a recipient of information will impact the decision made at the fourth step (evaluating information) of epistemic processes as we illustrate in Figure 1. Two examples relating to public health are given below.

1. It is widely known that mass vaccination can eradicate diseases (such as polio) or provide herd immunity (to diseases like measles, mumps, and rubella, or MMR) if a large enough portion of the population is vaccinated. However, despite abundant evidence supporting vaccine safety and efficacy, a significant number of people denounce the routine vaccination of children.[17] While poor information has often been implicated in vaccine hesitancy (as we will discuss later in the report), John (2019) argues that even if a vaccine-hesitant parent were to accept the scientific evidence pertaining to vaccine safety and efficacy, they may still rationally disagree with health officers or government officials on the basis of differently held values. For instance, a parent and a public health official may both know that MMR vaccines present a very low risk of adverse health effects, however, the public health official is mainly interested in establishing herd immunity to MMR in a larger community while the parent is primarily concerned with protecting their individual child's health. Furthermore, given their different goals, they may have a different risk tolerance. John argues that if the parent has reason to believe that the health official holds different goals and values, then even though they both acknowledge the same risks and benefits of vaccination, the parent may rationally refuse to vaccinate their child on the health official's advice.

2. Furman (2020) points out that in the 2013-2016 Ebola crisis in West Africa there was high penetration of scientific communication through affected regions. However, in some instances, community members also had reason to believe that the values held by scientists and medical workers did not align with their own and therefore had reason to resist expert advice. For instance, it was common practice that bodies of the deceased were not returned to families because they remained highly infectious and community burials were a common point of new infection. However, proper community burials were linked to beliefs about the fate of the deceased and the fate of communities (e.g. crop failure) if not performed. With their reasoning embedded in a different framework of values, community members often had good reasons to resist the efforts of health workers by performing secret burials or refusing to hand over contact tracing information until proper burial arrangements had been made.

[17] The adoption of vaccination campaigns around the world is a complex issue, involving cases of unsafe vaccines (Arkin 2019), fake vaccine campaigns used to cover up intelligence work (Kennedy 2017), religious beliefs (Navin 2017; Wombwell et al. 2014), and many others. However, in affluent democracies a recent movement against MMR vaccines seems to relate more to communities that refuse to vaccinate children despite abundant scientific evidence pertaining to vaccine safety and efficacy.

Overall, it is important to appreciate that even when high quality information is equitably available to and consumed by decision-makers and actors, misalignment of values means rational deliberation can still result in reasonable disagreement over a course of action.

Box 2.3.2
Clashing values in a health crisis: An example from workshop scenario 1 & COVID-19

In workshop scenario 1 a health crisis of an unknown source sweeps across the world. Official national organisations seek to reassure their populations, provide informed advice and guidance to keep their populations safe, and take action to address the crisis. The populations want to know how to stay safe. They seek information and guidance tailored to their situation:

Assume that all population members acquire true and reliable information about the health crisis. This outcome is very unlikely for reasons described throughout the rest of this report, but we here want to highlight that equally well-informed individuals may still rationally disagree on the proper course of action in response to a crisis due to value prioritisation.

The recent COVID-19 crisis provides a prime example. Complete social lockdown focuses on reducing threat to life of a vulnerable subpopulation from COVID-19 by reducing the spread of the virus, but lockdown measures can also increase risk to life from other causes (e.g. depression and suicide rates may increase due, for example, to job loss or social isolation) and threaten economic prosperity. Individuals can balance these risks differently and may have a hard time reconciling their different roles and values. For example one individual may be an owner of a small business, a school governor and care provider for older relatives and be torn between economic and familial interests. Different people with different value priorities (or the same individual with multiple value priorities) may be well-informed but still have rational grounds for disagreeing about appropriate guidelines for controlling COVID-19 .

The disruptive influence of unaligned values on timely decision-making and effective collective action in a democratic society will always be a challenge, but this must not only be looked upon negatively. Social epistemologists point out that exposure to diverse viewpoints and critique is central to encouraging epistemic modesty, careful deliberation, and well-considered judgments

(Holst & Molander 2017; Sunstein & Hastie 2015). Toward this end, modern communication technologies might be expected to bolster a society's potential for sound decision-making by facilitating the free exchange of opinions and the public critique of information and information sources.

However, many of the benefits of disagreement from clashing values is lost if the differing constituents also base their arguments on false information. While value (mis)alignment between actors has a strong influence on collective decision-making, in this report we focus on the prerequisite of reliable and timely information being available to all constituents. This leads us to focus on how emerging communication technologies can facilitate the malicious or accidental manipulation of information, overwhelm user capacities to process and filter information, asymmetrically emphasize minority extremist viewpoints and suppress marginalized voices, and make it more difficult for information recipients to evaluate the trustworthiness of information sources.

3. Analysing Epistemic Security

Through the development and analysis of epistemic security scenarios, we have found common themes that help frame and tease apart the myriad challenges in this space. In this section we draw on existing literature to provide an initial framework for analysing epistemic systems and vulnerabilities. In section 4 we outline how this framework can be used in practice to analyse a specific scenario of concern, building on the methodologies used in the workshops.

Section 3.1 first looks at the pathway from information production to informed collective action and at the four main knowledge sources discussed in classic epistemology: *experience, memory, reason,* and *testimony*. These provide a useful framing for the impact of technology on the supply-side and distribution of decision-guiding information. In Appendix 3 we show how these also provide a useful framing for novel epistemic vulnerabilities introduced by new technologies.

We then look at four key aspects of epistemic vulnerabilities: *adversaries and blunderers, attention scarcity and bounded rationality, insular communities and group polarisation, and fabrication and erosion of trust.* While these headings all interact with each other in important ways, they provide useful perspectives with which to analyse potential and unfolding crisis scenarios. Each helps to draw out important details and nuances that matter for effective intervention. In appendix 3 we show how these aspects can help frame suggested epistemic security interventions.

3.1 From information production to informed collective action

In the scenarios developed in the workshop, significant social harms emerged from either misinformed collective action (e.g. in scenario 5 - xenophobic ethnic cleansing) or from a failure to coordinate informed collective action (e.g. in scenario 1 - global health crisis).

To identify the source of disruption in light of failed collective action (real-world or hypothetical), we find it conceptually useful to break down the steps in the epistemic processes leading from information production to informed collective action, as outlined in Fig. 2.

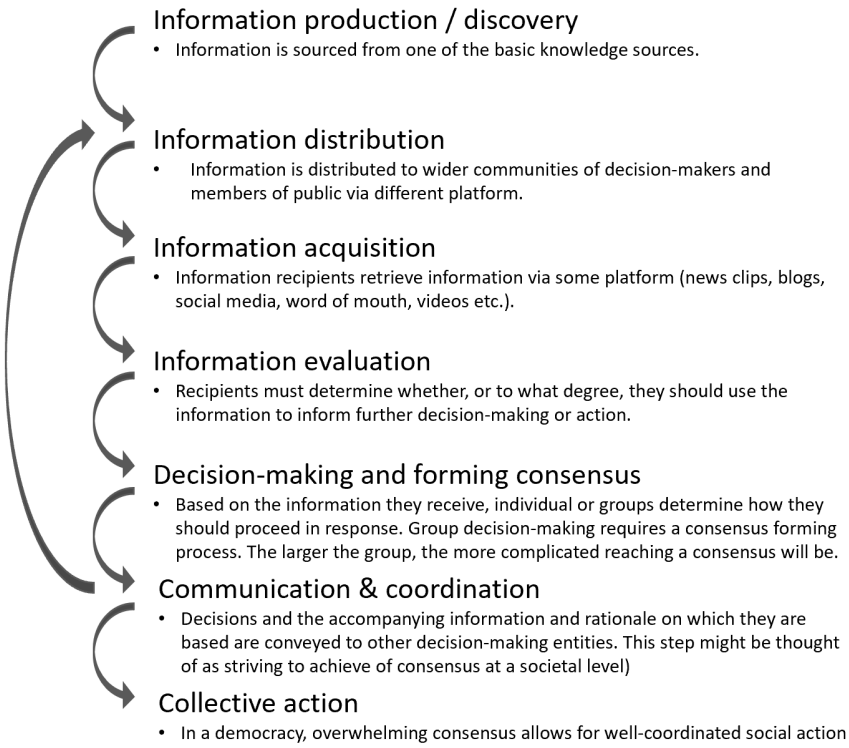


Figure 2: Epistemic process feeds collective decision making and coordinated action

This depiction of an admittedly simplified step-wise process makes clear this report's assumption that access to true information is a key driver in informed collective action through the crucial step of coordinated decision-making. In an epistemically secure society, if all decision-makers - official decision-making bodies and members of society - have access to the same true information , then it would seem that informed collective action should ensue. (*This assumes of course that we set aside the influence of conflicting values that will be present in any heterogeneous population* - see section 2.3.2) For example, if all relevant decision-making bodies were to have access to the same reliable information about trends and causes of climate change, then deliberative processes should lead to an agreement that significant steps must be

taken to mitigate the crisis. But given the limited degree of agreed societal action on climate change (or other real world examples that our scenarios draw on), we infer that the failure to coordinate and act implies that something goes wrong in the *information → decision → action* process which suggests that our society is not epistemically secure. The stepwise process allows us to zoom in at particular steps where epistemic vulnerabilities may be present.

We start our analysis with information production. In contemporary technology-rich societies information is abundant. To emphasise the abundance of information production and to elucidate the role that technology plays in enabling this abundance, we find it useful to refer to the four main sources of knowledge presented in traditional epistemology: experience, memory, reason and testimony. Contemporary societies contain numerous specialised roles and institutions for creating and sharing information via these conduits, and modern technology plays a significant part in mediating and improving human access to each knowledge source. In philosophical literature such technologies are often referred to as *epistemic enhancers* as they augment or improve human perceptual and cognitive capabilities.[18]

Experience

People gain knowledge from experience by observing the world for themselves. Recent advances in technology have improved our ability to gather and retain experiential knowledge. There are now more people alive than ever before, and many of them are equipped with devices that can capture experiences directly in accurate and persistent formats (e.g. mobile phone video recordings). There are also a very large number of devices that can capture records of events without human presence, including satellites, CCTV cameras, Internet of Things, and computer and network logs.

Memory

Closely related to experiential knowledge is knowledge from memory. People can file away knowledge they have acquired via experience (or the other knowledge sources) to draw upon at a later time. Digital technology enables the creation and curation of extremely large repositories of "memories" that can be easily and rapidly retrieved with perfect fidelity.

[18] Term coined by Paul Humphreys (2004) in *Extending Ourselves: Computational Science, Empiricism, and Scientific Method.*

Reason

Knowledge from reason is acquired by deducing truths about the world from a more basic set of information such as physical laws or well-established facts. For example, scientists, analysts and other "knowledge workers" reason about information to generate new insights about the world. Specialised languages and technologies help extend the ways that such reasoning can be explicated and shared with others, such as advanced data visualization tools that leverage our ability to digitise vision and sound to produce conceptual representations. In addition, we also have algorithmic and automated approaches to information distillation, pattern extraction and anomaly detection, enabling machines to derive knowledge with little human input.

Testimony

Testimonial knowledge is knowledge acquired from the reports of other people. Every time a person consults a book, searches the web, or attends a lecture, she acquires knowledge via testimony. Without testimony, each new generation would have to relearn all the facts known to previous generations, and human collaboration on a scale that builds civilization and sends rockets to the moon would be impossible. Modern information technologies make testimonial knowledge easily accessible to vast populations by connecting more people at higher speeds than ever before and by providing individuals with abundant sources of information. For example, one of the world's largest companies, Google Inc has as its mission statement "to organize the world's information and make it universally accessible and useful", and many of its products (e.g. Search, Books, Scholar, Sites, YouTube) increase the access of uncensored internet users to testimonial knowledge sources.

However, while information-mediating technologies enable the mass production and distribution of information to broad audiences, there remains a significant challenge in minimizing the production of unreliable information, in differentiating between reliable information from trustworthy sources (which, if true, would constitute knowledge) and unreliable information from untrustworthy sources, and in eliciting decision-guiding information equitably from diverse communities. In particular, there is a deficit of robust and epistemically valid pathways for historically marginalised voices to be expressed and attended to.[19]

[19] AIDS activism in the 1980's United States provides a classic illustration of the extreme difficulty with which epistemic "outsiders" gain credibility within epistemic communities with established social norms and institutional structures (Epstein 1995).

As we describe throughout the following section, modern technologies can exacerbate challenges to epistemic security in various ways along the pathway from knowledge production to decision-making and action. In the next section we describe different sources of vulnerability in an epistemic system, each providing a useful perspective to understand what is going wrong in the scenarios we explored.

3.2 Themes of vulnerabilities in social epistemic infrastructures

As life becomes more complex, as society becomes more interconnected, and as decisions become more distributed (as our social epistemic infrastructures become more complicated), it becomes increasingly difficult to ensure that reliable information is uniformly available and readily accessible throughout society. Sources of epistemic threats and vulnerabilities - from state-run disinformation campaigns to cognitive biases of elected decision makers - are too numerous to list. (Indeed, awareness of the many such threats and vulnerabilities is rapidly evolving as is highlighted in related works - section 2.2). Here we consider four themes of technologically-exacerbated vulnerabilities and threats that present challenges to the maintenance of epistemically secure societies:

- Action by adversaries and blunderers
- Attention scarcity and bounded rationality
- Insular communities and group polarization
- Fabrication and erosion of trust

We note that these themes do not constitute mutually exclusive groupings of epistemic vulnerabilities and threats; they interact and feedback in numerous ways. Nonetheless, in the analysis of the scenarios developed in the workshops, we have found it useful to tease out these four themes and work through them in turn, as each highlights and emphasises different aspects of a social epistemic infrastructure and its technologically influenced dynamics.

3.2.1 Adversaries and Blunderers

Decision-guiding information is frequently modified or manipulated by third parties. When information manipulation is intended to deceive, mislead or confuse, it is considered an ‘adversarial attack’ against knowledge acquisition and distribution processes. For instance, a political actor or foreign power may modify or fabricate news stories or misrepresent a politician’s comments in order to undermine a presidential campaign. We label the instigating actors of such attacks as *adversaries.*

On the other hand, actions that bring information recipients to false or poorly supported beliefs can also be well-intentioned or accidental. For example, a vaccine researcher wary of side effects and distrustful of medical authority might make a well-meaning but slightly alarmist comment during an interview, which could then be picked up and spread by an online community of parents, instigating a widespread anti-vaccination campaign. We label the instigators of such accidental or well-intended interferences as *blunderers*.

Mal-intended adversarial actions meant to disrupt and/or hijack knowledge-acquisition and decision-making processes and the accidental or well-intended interference of blunderers can not be strictly delineated. The moral distinction between warranted and unwarranted information manipulation is fuzzy at best, and so there is a continuum from epistemic adversaries to blunderers who engage in a spectrum of activities (Figure 3). For the purpose of systematically analysing epistemic security we find it less useful to focus on the intention behind epistemic system interference, and highlight instead the possible negative consequences of certain activities on our epistemic systems.

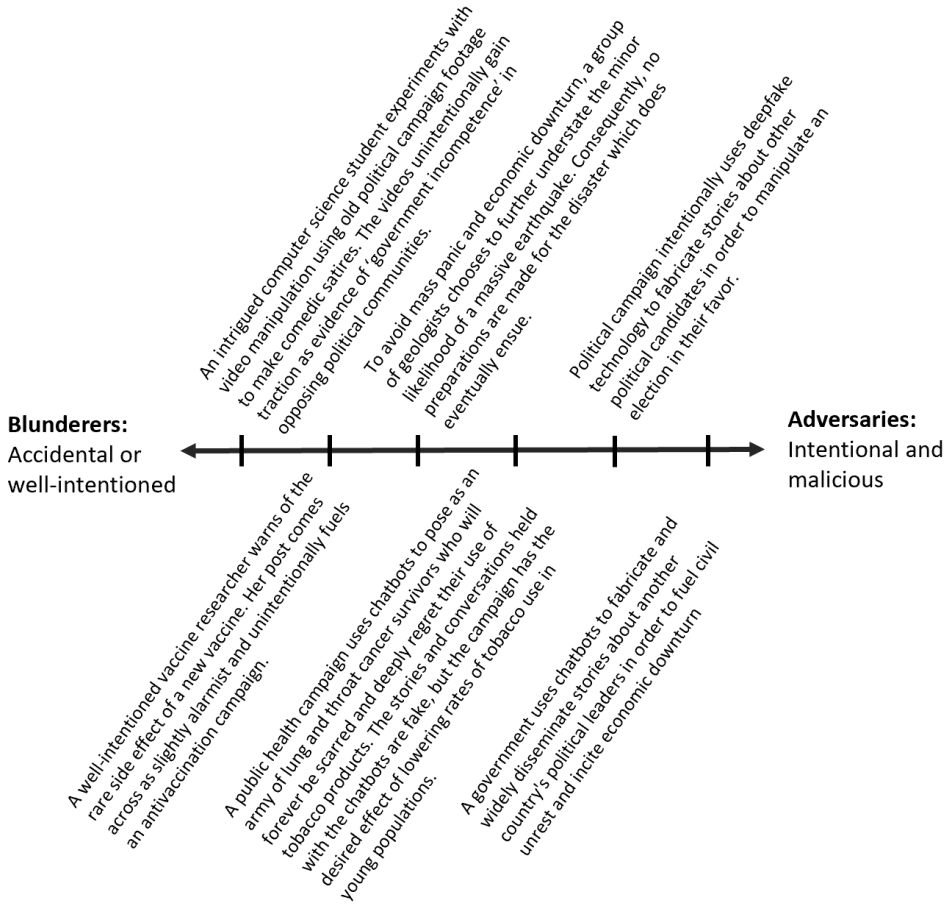


Figure 3: The spectrum from epistemic blunderers to adversaries

As many of the works mentioned in the introduction highlight, the actions of adversaries and blunderers in our heavily technologically mediated system of knowledge production and dissemination makes our current epistemic situation precarious. Whether we are threatened by state-sponsored disinformation operations, corporate propaganda aiming to sow doubt, criminals seeking to extort or exploit, curious hackers keen to explore new technologies, or merely unwitting social media users, it has arguably never been easier to intentionally or accidentally undermine the functioning of our social epistemic infrastructure. Several factors enabled or exacerbated by new technologies contribute to the increased threat of adversarial or accidental influence to our epistemic systems:

- Global connectivity: New communication technologies and platforms allow (mis/dis)information to travel more widely and quickly than ever before.
- Lack of accountability: The global community is often unable to meaningfully attribute or prosecute malicious acts against epistemic security - and is therefore unable to deter such malicious acts - when these are carried out through digital or autonomous means.
- Social visibility: The social network structure and epistemic norms of online communities, and the digital behaviours of individuals (including through their membership of multiple online communities) produce highly visible traces which make these groups and their individual members more susceptible to effective targeting with specific content (e.g. advertising, political opinions, but also disinformation).
- Automation: The ability to automate, or leverage existing automation of, cognitively demanding activities such as targeting (compare informer networks to facebook's targeted ads), experimenting (compare focus groups to automated A/B testing), and content crafting (consider recent remarkable developments in image, audio and text synthesis) significantly reduces the cost and allows a massive scaling up of the production of (mis/dis)information, the targeting of this information at the most receptive individuals, and distribution of targeted and tailored messages to massive global audiences.
- Financial Incentives: The technical ability of digital platforms to measure, in real-time, interactions between individual users and individual content items such as a single news story or advertisement, has enabled the creation of business models around specific engagement metrics (for example click-through rate or CTR), which in many cases has led to the prioritisation of these metrics. Divisive, controversial or otherwise emotionally gripping content performs well on such metrics, not only by maximising user engagement but also through increasing the likelihood of subsequent sharing, propagating the message on the network and increasing engagement for other users. This dynamic creates a perverse financial incentive for digital platforms to limit their restriction of (mis/dis)information.

Furthermore, in liberal democracies all of these factors operate against a legal and political background that limits the ability of the state to restrict speech except in very narrow domains. Consequently, new technologies increase the threat posed by adversarial actors or blunderers to our systems of knowledge acquisition and information exchange. Adversaries and blunderers may generate fake content outright, or they may merely selectively repackage truthful anecdotes to generate misinformation. They may either intend to cause harm or act selfishly, or they may believe themselves to be acting in society's best interest. Regardless of the exact method or motivation, the study of ongoing or hypothetical future crisis scenarios should include careful consideration of adversarial efforts and blunderer interference. Based on our understanding of current trends, these threats are likely to become increasingly destructive to our epistemic systems as information technologies evolve.

Box 3.2.1 a
Adversaries engage in political character assassination (Workshop scenario 2)

In this scenario, a politically motivated group undertakes a long term strategy to undermine future political leaders of a rival government. A group of 'potential future leaders' is identified and a whole 'fake history' is created for each individual as they progress through their careers. The plan is that once these individuals become influential, some of the fake facts about their early lives can be used to manipulate their actions either overtly or subliminally. This influence could be achieved in a number of ways:

- Traditional blackmail either for money, as a way of financing terrorist acts, or to make them behave in a way that they would not otherwise have done (voting for example);
- Simply distracting their attention at a political or economic crisis point so that chaos ensues;
- Releasing what would by then be authenticatable historic information, in order to undermine their credibility and force a resignation or other desired outcome.

Such elaborate character assassination planning is enabled by the availability of cheap data storage, and the technology to allow photographs and documents to be believably tampered with. The ability and foresight to create fake documents at the appropriate time (e.g. a picture of someone dealing drugs created and stored when they were 20 years old and held for use

in 20 or 30 years' time) adds to the authenticity of the spurious information, as does the creation of a holistic story involving multiple event and associates. This opens up a variety of opportunities which could enable the subversion of the political process, or create enough confusion to reduce the ability to deal effectively with a set of simultaneous crises.

Box 3.2.1 b
Blunders cause economic collapse (Workshop scenario 4)

In this hypothetical scenario, financial professionals make profit-driven decisions which are quite lucrative in the short term, but in the long run have the unintentional consequence of leading to financial system failure and economic collapse:

A highly competitive culture within the financial industry, and its focus on short-term gain, leads the industry to create narrow AI code that lacks proper validation and verification and which makes short-term gain decisions. Financial professionals, driven by profit, encourage the short-term gain culture. They also begin to advise investors to rely on crypto-currencies. The expectation that commodity has value underpins the international financial system. As a result dependence on crypto-currencies, which are not linked to any tangible asset nor supported by any national government, threatens that system. International regulation by the financial authorities, designed to keep a financial system based on commodities with value stable, fails to adapt sufficiently quickly to the new models of financial dealing driven by the use of crypto-currency. Before the authorities can enact international controls on these new models, investors suddenly lose faith in a specific crypto-currency when its technical underpinning is compromised. They attempt to realise the capital it represented in large numbers. The crypto-currency cannot deliver and this leads to the disintegration of other financial models linked to it. As investors have traded real commodities using crypto-currency, this undermines the currency of countries which invested real money in the production of those commodities. Traditional financial models therefore also begin to break down, leading to a failure of most parts of the financial system and economic collapse.

3.2.2 Attention Scarcity and Bounded Rationality

When a decision maker is confronted by a quantity of information that is greater than her cognitive capacities can handle, she experiences an information overload; she finds herself

overwhelmed by an excessive amount of information. This is not a new phenomenon. As Ann Blair (2012) points out, as early as the thirteenth century, scholars have complained that "the multitude of books, the shortness of time and the slipperiness of memory" can overwhelm the human intellect. However in more recent times the emergence of information technologies, particularly internet services, have exacerbated the issue, making larger quantities of additional information more easily accessible to larger populations of decision-makers than ever before (Roetzel 2019).

In his book *Designing Organisations for an Information Rich World,* Herbert Simon (1971) points out that in an abundant information environment the most limiting factor to human information processing is the human capacity for *attention*.[20] As Simon explains, humans are serial processors that can effectively attend to only one item at any given time, and their time is limited. Consequently, the rapid increase in available information and information channels has led to the emergence of a competitive "attention economy" in which information-providing organisations must compete for the limited attention of their audiences.

As such, information providers are incentivised to invest heavily in strategies to make their information products more attention grabbing. For information creators, a competitive attention economy exerts a pressure to create content that is sensational, that can be rapidly consumed, and that is emotively charged, that relates to a person's identity or group affiliation, and that affirms previously held beliefs.

While these pressures have always existed in the media environment, there are now much better tools available to content creators to attract and measure audience engagement. For example, digital platforms focus on developing intuitive interfaces that facilitate rapid engagement with content (e.g. by making information and options easy to find). If it takes a user too much time or effort to locate information that holds their interest, they are likely to switch platforms. Similarly, search engines prioritise and filter recommendations by relevance, and content aggregators such as news or social media feeds prioritise items to target user interest. Such prioritisation and filtering is known to rely on many factors, including past user behaviour (on and off the platform), the user's social networks (as expressed by "connecting" to other users) and group membership, stated demographic information, and expressed preferences. The same tools and strategies used to deliver engaging content are also used to deliver engaging advertisements. As advertisements often provide the main revenue source for such platforms, there is a strong incentive to invest in and perfect attention-grabbing methods.

[20] For Simon an information rich world is one characterised by computer-based information processing and xerox copying machines, but we may now add a host of internet-enabled communication technologies.

Attention-grabbing strategies employed by information creators and on information distribution platforms can be epistemically problematic for several reasons.[21]

First, the methods employed by information distributors to prioritise attention-grabbing content are not reliable indicators of the truth value or contextual usefulness of the information being promoted. The methods employed by information creators and distributors to attract attention - such as content personalization and targeting techniques and visual presentation - are *truth-neutral*; they help perpetuate information irrespective of whether it is true or false, informative or misleading.

Second, the commitment of resources to truth-oriented activities such as fact checking and double sourcing are disincentivised. Because the attention economy is fiercely competitive, information creators and distributors are incentivized to commit more resources to attention grabbing strategies. In turn, fewer resources can be committed to employing strategies that aim to ensure the distribution of reliable information. In other words, in a competitive attention economy there is an implicit penalty for information creators and distributors who wish to engage in the perpetuation of reliable information. The quality of prevalent information is therefore likely to decrease.

Finally, the competitive pressures of attention scarcity also implies that methods for drawing consumer attention must evolve rapidly for information providers to achieve and maintain an edge over competitors. In turn, a quickly changing information environment may leave consumers behind in terms of their ability to discern which information sources are most trustworthy, undermining efforts by consumers to prioritise truth-oriented information sources. We speak more about trust in technologically-mediated information sources in section 3.3.4.

While on the information supply and distribution side there is fierce competition for scarce attention. To further the challenge, on the information consumption side there is also growing empirical evidence that humans are not fully rational about what information they choose to consume. Various decision heuristics and biases affect what information people attend to

[21] Information processing technologies may also help in reducing the cognitive effort required to filter out irrelevant information. For example, Landhuis (2016) describes how academic communities curated on social networks like Facebook and Twitter and searchable and preference-learning literature databases like Google Scholar and Academia can help individuals significantly narrow the pool of publications to which they lend their attention, a necessity given that academic databases are experiencing exponential yearly growth (http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html). However, while significant technological advancements have been made in computer-based summarisation, filtering, and information network curation, these advances still lag behind the growth of information databases and the growing demand for high-quality, context-appropriate information. For instance, Landhuis notes that despite the aid of technological advances in academic resource curation, academics spend 6-8 hours a week curating their own information resources. In many cases, an overly burdensome and cognitively demanding amount of information processing must still be conducted by human decision-makers, and the challenge of information overload persists.

(Kahneman 2011). In particular, humans are naturally prone to confirmation bias; they are more willing to attend to, and more likely to believe, information that confirms their pre-held beliefs (Klayman 1995) or which they have seen before (Kahneman 2011). Attention grabbing strategies such as the targeting of user interests on social media sites and advertisements consequently encourage the development of filter bubbles and echo chambers in which individuals consume information that primarily serves to bolster their existing opinions and beliefs. This information is not necessarily true or well-evidenced (See section 3.3.3 for an extended discussion on echo chambers and opinion polarisation).

Furthermore, not only are humans biased in their allocation of scarce attention, they also imperfectly rationalize about the information they do consume. They suffer from what Simon (1957) calls *bounded rationality.* Perfect rationality in decision-making requires that a person has access to complete information about the options available to her, has perfect foresight about the consequences of those options, and the cognitive capacity to process this information and optimize her decision accordingly. Rarely, if ever, do any of these conditions obtain, but they define a more well-reasoned individual (though her rationality is bounded) as one who more closely approaches the ideal criteria for perfect rationality. Given this ideal, it is clear why a competitive attention landscape is problematic: it limits a decision-maker's access to information regarding available options which in turn limits her ability to reason well about those options. The goal of attention grabbing strategies are not to inform a decision-maker about the pros and cons of all her options, but to focus her gaze in one direction or another.

Overall, our competitive and technologically-enabled attention economy presents less than ideal conditions for well-reasoned decision-making. Information providers prioritize eye-catching appeal over veracity, diversity, and informativity, and human decision-makers are overwhelmed by information curated with little regard to its orientation to truth.

3.2.3 Insular Communities and Group Polarisation

The communities we belong to and identify with have a strong influence over our belief formation and knowledge acquisition processes. As described in Kusch's (2002) *Knowledge by Agreement,* this is true even within the most "scientific" and "rational" communities.

In the context of epistemic security, community norms play a role in setting shared commitments to beliefs or belief formation processes, the bundling of beliefs, the willingness to trust members who share similar beliefs (or mistrust those who hold differing beliefs), and the role of influencers and network topology in belief formation in groups (O'Connor and Weatherall 2019; Urban 2019).

Community thinking and group collaboration can be of great epistemic benefit to society. No individual holds all the knowledge or working capacity to achieve great organisational feats. Knowledge transfer via testimony in combination with the organisational structure provided by a community - such as productive divisions of labour or the necessary outsourcing of tasks - enables the greatest of human achievements (Anderson & Wagenknecht 2013; Malone 2018; Wray 2002).

Information acquisition from another person (through *testimony*) often relies on evaluating the communities to which that person belongs. Established community norms and structures help to ensure the trustworthiness of individual community members as sources of information. For example, Heather Douglas (2017) argues that we have good reason to trust that individual scientists are reliable sources of information if those scientists are members of a 'well-functioning epistemic community'. Following on the work of social epistemologist Helen Longino (2002), Douglas considers a well-functioning epistemic community to be one in which the following conditions hold:

- there are platforms and avenues for dissenting opinions and criticism to be voiced,
- criticism is taken seriously,
- diverse viewpoints are considered equitably.

If a scientist is a member of a community that follows these epistemic codes of conduct, Douglas suggests, then it is reasonable to believe that the scientist is also held to these same standards. Accordingly, we may trust an individual scientist to be a reliable source of information in her field of expertise if the community of scientists to which she belongs is a well-functioning epistemic community.

However, community-mediated reasoning can also have negative epistemic consequences. Like individuals, insular groups suffer from conformity bias which inhibits the consideration of external viewpoints and new evidence that run counter to the group's preconceived beliefs. In turn, conformity bias taken to the extreme may result in community polarisation or radicalisation in which groups that initially differ slightly in opinion on a particular issue move to strongly disagree (Sunstein 2007). This change is largely a response to cultural or ideological differences between groups, and each community ignores or neglects outside viewpoints to maintain an identity independent of "the other" (Whitaker, 2018). The issue is exacerbated because extremely large, highly-visible, or well-connected epistemic communities (e.g. the activist wing of a political movement, the followers of a popular brand, or the members of a majority religion) also attract a disproportionate amount of attention. Consequently the viewpoints presented by these groups are often assigned more epistemic weight even though the size of a community or the community's ability to disseminate information publicly is not a clear indication of the community's epistemic rigour or the truth-value of the community's

claims. It is all too easy for outsized importance to be placed on the opinions of mass movements on the one hand, or on the opinion of minority radical groups with high connectivity and effective strategies for drawing attention. However, neither the number of supporters nor the loudness of their support make a claim more true.

Emerging technologies amplify the epistemic processes of communities. In well-functioning epistemic communities, technology can help identify diverse voices and increase the evidence and rigour with which they are evaluated. However, in poorly-functioning epistemic communities technology can lead to silencing, groupthink, radicalisation and polarisation. The dynamics by which these take place are discussed below.

To start on a positive note, a general trend of technology has been to provide more ways in which to create and manage communities (Shadbolt et al. 2019). Communities require an infrastructure with which to communicate, and the emergence of new communication technologies facilitates the formation of new, larger, and more complex yet efficient epistemic communities and inter-/intra- community collaborations. For examples, global connectivity has allowed the creation of one-to-many and many-to-many communications networks connecting billions of individuals in different parts of the world (e.g. social networks), topic-based-tagging has enabled the fluid formation of topic-based communities, and video streaming platforms have enabled many more individuals to build communities around their unique content and personas (Day, 2015). In general, technology companies have been encouraging the creation of communities on their platforms.[22]

If used well, information technologies and social media platforms in particular can help build and maintain epistemically well-functioning communities that encourage the expression of diverse viewpoints and that are receptive to criticism. For example, new information technologies provide platforms for minority opinions and critical viewpoints to be easily voiced, which, if integrated with and scrutinised alongside mainstream opinions in a rigorous manner, lead to better decision-making. Furthermore, many online community platforms provide tools that allow community moderators to manage group membership and content, which can be used to promote positive epistemic norms such as rigour and openness. For example, reputation or “karma” systems are used to reward or punish members with greater or lesser

[22] Digital platforms do not enable community formation merely out of goodwill. Users prefer to be on the same platform as their friends and colleagues, or other people they consider to be "part of their community". This is an example of the *network effect,* whereby the value of a product or service increases with the number of users or participants. This in large part explains the dominance of a few platforms in every niche, and the competition between platforms to appeal to communities that are not well-served by existing alternatives. It also partially explains why users are often reluctant to abandon a platform even when there are serious misgivings about its practices and impacts, unless such abandonment can be coordinated *en masse*.

visibility and moderation powers according to their alignment with community standards.[23] At their core, such systems are important for maintaining the quality of discussion and reliability of information created and shared by a community. An example of a system that enforces epistemic standards is the peer review process used by the majority of academic communities to reward individuals who adhere to community defined standards of good methodology, objectivity, and academic rigour with the publication of their work (greater visibility) while minimizing the visibility of statements that fail to meet standards of rigour or are suspect due to a conflict of interests.[24]

However, while new technologies can and do facilitate the establishment and maintenance of diverse and reliable epistemic communities, they also enable (and sometimes incentivise) epistemically detrimental community practices. For instance, while social media platforms can help give a voice to critical, underrepresented, or marginalised views, such platforms can also be flooded (by adversaries or blunderers) with poorly-reasoned radical opinions, held by few, to make them seem much more prevalent and widely accepted than they actually are. As Asch (1951) observed, individuals are much more likely to accept and echo majority community views, even if they personally hold evidence that contradicts those views. Similarly, karma systems that can help enforce good epistemic standards can also be hijacked to ostracize and silence individuals deemed 'unwelcome' or 'undesirable' for non-epistemic reasons, diminishing the diversity of views in a community and undermining its quality of reasoning. It is important to keep these knock-on effects in mind.

Because the tools provided by social media platforms are agnostic to the epistemic standards promoted by each community, and due to confirmation bias and other psychological and social tendencies, the communities that use such platforms to engage in critical rigorous discussion and consider minority views fairly are the exception rather than the norm. Rather, such platforms enable communities to follow their biases and develop into insular echo chambers for opinions that align with the pre-held beliefs of the members or moderators. Such insular communities are largely deaf, and at times aggressively opposed, to criticism or opposition. As a result, individuals embedded in these communities will often be unaware of or unable to access information which is important to good decision-making and informed collective action.

---

[23] For example, a user that contributes content that other users find interesting or insightful may gain the power to remove offensive or unhelpful content posted by other users, close off discussions, or to block users that violate a site's policy.

[24] It should be noted that the academic peer review system is far from perfect, and the critique and continuous development of the system is part of what makes certain scientific communities well-functioning epistemically. In particular, digital technologies have enabled new ways for disseminating and responding to scientific findings (in particular via electronic preprints), and the formation of norms and standards regarding the appropriate use of such technologies is ongoing.

Box 3.2.3
Xenophobic Polarisation (Workshop scenario 5)

In this scenario a far right xenophobic faction wishes to force the departure of a specific ethnic community and pursues its goal by spreading false information about the ethic group:

The far right faction's controlling group decides on a strategy of implicating the ethnic community in a chemical or biological attack. The aim is to turn public opinion against the ethnic community so much that extreme violence against the ethnic group will be considered justifiable by elements of the population. First, the far right faction uses AI-enabled technologies to identify (target) individual members of society who are most likely to sympathize with the group's xenophobic inclinations. The radical group then stages the chemical/biological attack, and uses AI-enabled communication technology to push messages at speed and scale to likely sympathizers claiming that the specific ethnic community is responsible for the attack. The sympathisers continue to share the radical group's messages among themselves and with other like-minded individuals. Elements of the targeted population then mobilise, arm themselves and use violence to drive out the ethnic community. The far right faction has achieved its objective.

Because technology has made participation in online communities easier than ever, because the majority of online communities fail to meet the rigorous standards of well-run epistemic communities (as characterised by Longino and Douglas), and because these communities and their poor epistemic practices are persistent, the overall impact of technology on community epistemology has created significant challenges to epistemic security. It is important to note, however, that the creation of epistemic echo chambers on individual platforms does not necessarily preclude individuals from accessing diverse sources of information across a range of platforms and media. Information technology has enabled a very rich media ecosystem that allows people to easily join numerous communities, which may all contribute information and a variety of perspectives to a person's decision-making processes.[25] However, the challenge still remains, individuals have to be self-motivated to overcome their confirmation biases, and then actively challenge the biases of their communities and information sources, while maintaining a rich and balanced "information diet". While this may work for some, and indeed leaves some individuals and groups better informed than ever before, *being well-informed is often a privilege*

[25] http://www.ox.ac.uk/news/2018-02-21-social-media-and-internet-not-cause-political-polarisation-new-research-suggests

of time and resources that cannot be afforded by many, even in affluent and technology-rich societies.

3.2.4 Fabrication and Erosion of Trust

Evaluating the trustworthiness or reliability of an information source is often central to justifying the use of information as the basis for further decision-making and action. This is particularly the case when acquiring information from reports provided by other people (i.e. from *testimony*) Because the information recipient does not observe, rationally derive, or remember a piece of testimonial information for herself, she cannot directly verify the veracity of the source's claim. Instead she relies on indirect indications of speaker trustworthiness that are more easily accessible - she swaps cautious, evidence-based evaluations of information source reliability for quick heuristic evaluations.

As explained in Onora O'Neill's (2018) *Linking Trust to Trustworthiness*, decisions about who to trust draw on three components:

- evaluating the source's reliability
- evaluating the competence of the source
- evaluating the source's honesty or sincerity

Each of the three could be evaluated based on different aspects of the source's past track record. However, track records are not always available or easily accessible, and they do not necessarily guarantee future good performance. Instead, information recipients may look for signs of competence or expertise. For instance, a person might look to the source's credentials or certifications or ask whether the source is a member of a community that holds its members to certain standards of epistemic conduct. Alternatively, a person can look for indications of biases, intentions, or underlying motivations which may drive an information source to be (dis)honest or (in)sincere in her communications. For example, physicians who accept significant amounts of research funding from a pharmaceutical company may be more likely to push specific medication to treat a malady, and may be less trustworthy than physicians who are not in receipt of such funds.

This quick heuristic approach to evaluating the trustworthiness of sources of human testimony is imperfect but useful. Trust is necessary to the success of any instance of communication, collaboration, or delegation. Without a willingness to trust one another and to believe information provided by others, society would cease to function. Therefore, despite its fallibility, the quick heuristic strategy for evaluating the trustworthiness of testimonial speakers has generally served humans well in guiding our daily decision-making processes and informational exchanges.

However, our indirect methods of evaluating the trustworthiness of human testimony are sometimes ill-adapted in a technology-rich environment, leaving them vulnerable to manipulation. Relevant technologies include testimony-mediating technologies such as search engines and recommender systems on social media platforms, and arguably testimony-generating technologies such as natural language generation systems and other synthetic media sources that generate and modify information by automated means.

Such information-mediating technologies help to make testimonial knowledge more accessible and widely distributed. However, these technologies also make it easier to manipulate user trust by making it easier to hijack the indirect proxies for trustworthiness that people use to decide whether or not to rely upon a source of information. For example, machine learning-enabled natural language processing (NLP) and natural language understanding (NLU) systems may be used to identify and mimic speech patterns or vocal tones that users respond to positively in order to gain user trust regardless of the truth-value of the speech content. On this topic Matt Chessen (2017) warns that in the near future NLP/NLU-enabled systems will produce content indistinguishable from that produced by humans and will much more efficiently target receptive readers with tailored content. If these systems are being used to sway public opinion or drive decision-makers to specific action, the upset parties will likely develop their own NLP/NLU-enabled adversaries to push back. Chessen cautions that an arms race to control information will result in a future infosphere dominated by “machines talking to machines” and human observers will struggle to know where to turn to inform their decisions.

Regina Rini (2019) makes a similar point with regard to our reliance on photographic media. Humans are instinctively inclined to accept photographs or videos as perceptual evidence of the content depicted as if the viewer had directly observed the event for herself. However, since the development of deepfake technologies, Rini argues that consumers of visual media must suppress their instinct to believe what they see, and instead scrutinize the providers of visual media as one would a source of spoken or written testimony.

In addition to their ability to hijack trust mechanisms to spread (mis/dis)information, new technologies can also be used to undermine trust in individuals and groups. For instance, social media platforms can be leveraged to draw large amounts of attention to specific acts or statements that undermine a speaker's trustworthiness. Since it is inevitable that at some point in time all individuals or communities will make a poor decision or suffer some moral or ethical failing, drawing attention to such failings is an efficient way of quickly undermining trust in the person or group in question. As trust is slow to build but quick to destroy, drawing attention to such failings is an efficient way of quickly undermining trust in the person or group in question. In this way, social media consumers can be swayed to mistrust those who may actually be the

most reliable sources of information or the most careful decision-makers to whom one might defer.

Box 3.2.4
Epistemic Babble (Workshop scenario 6) & the erosion of trust in experts

In this scenario the ability for the general population to tell the difference between truth and fiction (presented as truth) is lost largely due to the widespread use of information mediating technologies:

Although information is easily available, people routinely purport to be other than themselves on electronic media and this goes undetected, so people cannot tell whether the information they are receiving is reliable or not. For example, social media has allowed people to put forward spurious views and for them to be accepted along with the views of true experts as being of equal value. This in turn has led to expertise being devalued and people no longer respecting or accepting the views of educated and knowledgeable people, or accepting authority in any way.

Additionally the education system relies on digital technologies to radically reduce the number of real teachers (without adequate testing of the change) and this results in pupils not developing the ability to apply critical thinking to the information that is presented to them, without the guidance of an adult.

The result of this 'Epistemic Babble' is that there is an environment of 'knowledge' and belief that could be easily manipulated.

3.3 The costs of informed collective action

In section 3.2 we consider four themes of technologically-exacerbated vulnerabilities and threats that present challenges to the maintenance of epistemically secure societies:

- Action by adversaries and blunderers
- Attention scarcity and bounded rationality
- Insular communities and group polarization
- Fabrication and erosion of trust

To better understand the combined effects of technologies on epistemic security, and subsequently on a society's capacity for timely decision-making and collective action, we find it

useful to think in terms of the *costs to information distributors* (government actors, experts, researchers, adversaries, blunderers, etc.) *and consumers* of dealing in reliable information. These costs are broadly construed as the expenditure of scarce resources (money, time, attention etc.) required to produce and distribute information. The various technologies described in section 3.2 drive these costs up and down in different ways which can make it either more or less difficult for adversaries and blunderers to interfere with an epistemic system, and more or less difficult for guardians of epistemic security to intervene.

One category of costs that has received much attention relates to information generation and distribution. **The costs associated with information production and dissemination have dropped significantly with the deployment of modern information producing and mediating technologies, leading to information abundance**. Lower costs for producing and distributing information have made reliable information more easily accessible to a broad audience and have also allowed diverse viewpoints, including those of minority and/or marginalised communities, to be more easily communicated to a wider population. As discussed above, diverse epistemic environments characterized by open communication and critique encourage the development of truth-oriented (or epistemically well-functioning (Longino 2002)) and epistemically secure societies.

However, **since many information distribution channels are truth-neutral, the cost for adversaries and blunderers to generate and distribute information which undermines informed collective action has also dropped**. The costs of information distribution have dropped significantly with any-to-all digital communication platforms, and costs of information generation are falling with the increasing capability and diffusion of synthetic media generation technologies. Furthermore, in democratic societies adversaries and blunderers are afforded some protection by freedom of speech legislation and norms which limits government capacity to intervene in social epistemic systems.[26]

The fall of costs for producing [mis/dis]information have highlighted the importance of another category of costs: of evaluating the reliability of information sources and of preferentially attending to reliable sources. **As adversaries and blunderers find it increasingly easy to gain attention (e.g. by appealing to community motivated reasoning, by appealing to cognitive biases, or by fabricating the appearance of trustworthiness), the costs of identifying reliable information sources go up for information consumers.**

[26]While the boundaries set upon governments by freedom of speech legislation prevents the censorship of extremist and unverified content, it is important to note that these rules do not require that such content be given visibility or reach. Additionally, the US first amendment law does not apply to private US social media companies which means they are free to decide what content is allowed and amplified on their platforms (Gershan, 2020). Innovations in information technology have contributed to a decrease in the costs for adversaries, thus the volume of adverse activity can increase.

To complicate matters further, it is not enough for most decision-makers to attend to reliable information sources most of the time: for informed collective action, they need to coordinate about which sources and topics they attend to in order to create a broad enough basis of agreed-upon reliable information around that particular topic. **With increasing fragmentation and polarisation of epistemic communities as enabled by many online platforms, the costs of information coordination also goes up.**

Furthermore, the challenge of coordination is often in tension with the challenge of providing a platform for all voices. This tension can be misused to intentionally delay or undermine consensus by continuously demanding more information, deliberation, and participation before action is taken. However, premature closure of discussions also has a harmful effect on society. Finding the right balance between open discussion and effective coordination is at the heart of epistemic security and needs to be addressed in a context-based manner that appraises the factors we list in this report, among others.

These ideas regarding the costs to information distributors and consumers of dealing in reliable information are further developed into a preliminary model for understanding the costs of maintaining epistemic security in Appendix 4. We note that while these rising costs pose challenges for informed collective action, they largely stem from broad improvements in our epistemic systems (it is still easier than ever to acquire reliable decision-guiding information): they should not be seen as trends to be reversed, but rather as tradeoffs that should be carefully addressed.

3.4 Summary of Definitions

To conclude this section, we summarise the key terms presented in the first half of this report. Readers interested in learning more about how specific technologies can be used to undermine or bolster epistemic security and collective decision-making processes should look to Appendix 3.

> **Decision-making entities** are individuals or groups who gather information about the world, and use that information to make decisions about how to act in order to achieve some goal. In group settings decision-making is more complicated as it requires the cooperation of many individuals in some consensus forming process, and coordination in acting on group decisions. Collective decision-making and action is further complicated when individuals do not have access to the same information on which to

base their opinions or when their end goals are informed by different values or incentives.

Knowledge/Information sources are the places from which a person or group can acquire information on which to base their reasoning and decisions. In traditional epistemology the four basic sources of knowledge are experience, memory, reason, and testimony, each of which can be influenced by information mediating technologies.

Social Epistemic Infrastructure (Social Epistemic Systems) refers to the collection of systems, processes, and actors that influence how knowledge is produced, distributed, acquired, modified, and evaluated within a society. Timely group decision-making and collective action are reliant upon the strength and efficiency of a society's epistemic infrastructure.

Epistemic communities are groups whose members share and enforce norms of knowledge production, distribution, acquisition, modification and evaluation. Such norms could be institutionalised, e.g. through a process like peer review, they could be codified, e.g. through a list of behaviours that can result in being banned from an online forum, or they could be more subtle, e.g. a tendency to engage more with certain topics, perspectives or sources while ignoring others. Communities that facilitate the contributions of diverse viewpoints and that encourage responsiveness to open critique are more likely to be producers of reliable information.

Epistemic security ensures that a community's processes of knowledge production, acquisition, distribution, and coordination are robust to adversarial (or accidental) influence. Epistemically secure environments foster efficient and well-informed group decision-making which helps decision-makers to better achieve their individual and collective goals.

Epistemic threats are factors that interfere with the well-functioning of a society's epistemic processes.

Epistemic vulnerabilities are weak points in a social epistemic infrastructure that are most likely to succumb to epistemic threats.

Adversaries interfere with processes of information production, acquisition, evaluation, and distribution in a society. In turn, adversarial action adversely affects the ability of decision-makers to make well-informed decisions that lead to timely and effective

collective action. Adversaries are considered to be individuals or groups who intentionally seek to undermine epistemic processes.

Blunderers are individuals or groups that, like adversaries, interfere with information processes, but unlike adversaries, their interference is either unintentional or intentional but well-meaning. There is a spectrum of actors between adversaries (intentional, malicious) and blunderers (unintentional or intentional but well-meaning).

Trust in a source of information is the degree to which a recipient is willing to take the information from the source to be true without independently verifying it.

Cost to information distributors (government actors, experts, researchers, adversaries, blunderers, etc.) is the expenditure of scarce resources (money, time, attention etc.) required to produce and distribute information. The various technologies described in section 3.2 drive these costs up and down in different ways which can make it more or less difficult for adversaries and blunderers to interfere with an epistemic system and more or less difficult for guardians of epistemic security to intervene. Overall, modern technology has decreased costs for adversaries and blunderers allowing the volume of adverse activity to increase.

The following terms are central to the final section of this report which deals with addressing epistemic vulnerabilities and adversarial (or accidental) action in order to maintain or restore a society's epistemic security:

Epistemic interventions are actions taken or policies implemented by a regulatory body to preserve the epistemic security of a society and thereby preserve the society's capacity for timely and well-informed collective decision making.

Systems-oriented views of epistemic security require stepping back to view all the actors, influences, and stakeholders within a social epistemic infrastructure as a whole. As we will emphasise in the next section, appropriate and effective epistemic interventions can be difficult to identify, requiring a system-oriented view.

Higher-order effects of epistemic interventions include their impacts beyond the immediate intended outcomes, for example a creative response by adversaries to new legislation. Many interventions, though well-intended, can have both positive and negative second- and third-order effects, requiring a systemic and risk-aware perspective when interventions are designed and before they are deployed.

4. Preliminary recommendations for appraising and maintaining epistemic security

Information-mediating technologies make it easier for a diverse array of information producers to generate informational content and distribute it to a wide audience. However, these technologies have also lowered the costs for adversaries and blunderers to create and distribute misinformation which ends up undermining informed collective action. Various epistemic vulnerabilities, including attention scarcity, bounded rationality, fragmentation and polarisation of epistemic communities, and the fragility of trust, both challenge informed collective action and assist the effectiveness of actions by adversaries and blunderers.

Overall, innovations in information technology have contributed to a decrease in the costs for adversaries, and thus the volume of adverse activity can continue to increase. Accordingly, to advise government actors (or other guardians of epistemic security - it can not always be assumed that a government is interested in maintaining epistemic security instead of undermining it) in preserving a society's ability to organise timely and well-informed collective action. We present the following recommendations to highlight areas where additional research and resources will likely have a significant impact on epistemic security in democratic societies:

1. **Develop technological or institutional methods to increase the cost for adversaries and blunderers in spreading unsupported, fabricated, or false information.** For example, penalties could be instituted for the knowing dissemination of false or misleading information or fines given to information organizations that do not undertake minimum fact-checking procedures. Such penalties do not have to be centralised or tangible: censure and condemnation from leaders and respected communities also has a role to play.

2. **Develop methods of helping information consumers more easily identify trustworthy information sources.** For example, information organizations and platforms could be certified as an *epistemically responsible information source*. Building on Boaz Miller and Isaac Record's (2013) explication of epistemic responsibility, an *epistemically responsible information source* would be one that has done all that it 'practicably can' to distribute true and well-founded information, where practicability - being possible in practice - is constrained by the information provider's ethical and social circumstances and by its technological resources.

Explore technological or institutional methods to "signal boost" reliable decision-relevant information in an asymmetric manner. Signal boosting reliable decision-relevant information requires engaging head-on with the various societally held views on what makes information true and useful, and developing and adapting tools and methods that work for, within, and across diverse communities.

To begin, it is possible to draw on existing practices such as scientific replication[27], journalistic fact-checking[28], legal evidential thresholds[29], and analytical quality assurance[30] that are employed by professional communities that have long been concerned with issues related to the production and consideration of true information.

Furthermore, in order for democratic authorities to serve the public interest, and in order to facilitate widespread collective action in response to crises, it is necessary to elicit input from socially and culturally diverse communities as well. In particular, historically marginalized groups in a society may have good reason to doubt information offered by authorities.[31] Therefore, it is important for diverse communities be given a voice when considering what constitutes a reliable information source and decision-relevant information.

3. **Develop technological or institutional methods to monitor changes in the epistemic ecosystem and to rapidly detect adversarial epistemic action during times of tension or crises.** While intervening in epistemic systems is fraught with unintended consequences, at the very least strategies should be developed for monitoring

[27] Philosophers of science have much to say about the efficacy of result replication as a method of verifying scientific results. Inter alia see Collins (1992), *Changing order: Replication and Induction in Scientific Practice;* Giles (2006) *The trouble with replication;* and Zwaan et. al (2018) *Making replication mainstream.*

[28] "Fact-checking has a traditional meaning in journalism that relates to internal procedures for verifying facts prior to publication, as well as a newer sense denoting stories that publicly evaluate the truth of statements from politicians, journalists, or other public figures" (Graves & Amazeen 2019). Also see Graves (2016).

[29] For example, the Crown Prosecution Service (CPS) in England and Wales lays out a set of general principles for prosecuting criminals which includes a threshold test for determining whether sufficient grounds of non-speculative evidence have been acquired for lawful prosecution. https://www.cps.gov.uk/publication/code-crown-prosecutors

[30] For example, the Aqua and Magenta books provide guidance on producing evidence based analysis during the design, implementation, delivery and review stages of policy making. https://www.gov.uk/government/publications/the-aqua-book-guidance-on-producing-quality-analysis-for-government.

[31] For example, historical and continued mistreatment of African Americans in medical trials (e.g. the Tuskegee syphilis study and the common exclusion of African Americans from clinical studies) and in clinical settings (e.g. the withholdment of pain medication) as well as sustained racial disparities in access to healthcare has understandably fostered widespread mistrust among African American's in medical institutions and practitioners (Washington 2006; Scharff 2010).

emerging information technologies and platforms, forecasting their impact on informed collective action, and monitoring emerging claims and narratives that could undermine collective action in times of crises.

However, until such significant technological and institutional advances are made, indeed, even once they are made, we must accept that coordinated collective action will be very difficult to achieve. Rigorously identifying and assessing potential interventions and threats to a complex social epistemic infrastructure is a time consuming and labour intensive process. Therefore, it is important to prioritize epistemic security efforts in order to minimize harm from epistemic threats. Toward this end we present the following recommendations:

4. **Build capacity to engage in holistic systems-mapping procedures (constructing an integrated view of social epistemic systems) and red-teaming strategies (deliberately exploring a scenario from an adversary's perspective) to help identify and analyse epistemic threats.** As we describe in the following section, systems-mapping procedures and red-teaming strategies help to provide more accurate overviews of social epistemic infrastructures and their epistemic vulnerabilities and strengths and to identify and analyze epistemic threats to society.

   Holistic overviews of social epistemic infrastructures are important because complex epistemic systems most often suffer from multiple overlapping epistemic threats and vulnerabilities such that a solution to one might exacerbate another. Many interventions will have second-, third-, and higher-level effects that could be unintentionally detrimental to epistemic security. For instance, attempting to discredit information spread by an extremist group through public education campaigns may make the extremist opinions seem more widely held than they are. The greater visibility of extremist perspectives could lend the group greater credibility in the eyes of observers limited by time and attention in their capacity to investigate the issue further. This does not mean that a public education campaign is an unwise intervention, simply that guardians of epistemic security must be careful to consider and prepare for possible knock-on effects.

   Overall, epistemic threats and vulnerabilities (of technological origin or otherwise) should not be addressed as a list of independent problems with prescribed fixes. See Table 3.2 in Appendix 3 for further examples.

   Finally, holistic system-mapping and red-team strategies also provide useful tools for identifying leverage points for effective intervention to mitigate threats and to lessen vulnerabilities. These capacities and exercises are particularly relevant when public

cooperation is required to achieve beneficial outcomes, e.g. in public health, crime prevention and environmental protection.

In section 5 we describe how holistic systems-mapping and red-teaming strategies might be implemented using an example from our workshop proceedings (workshop scenario 5 - Xenophobic Ethic Cleansing) to illustrate.

5. **Establish working relationships with a diverse array of experts who are experienced in identifying and analysing epistemic threats and who could serve as epistemic security advisors before and during crises.** In a crisis it is important that a democratic society can deploy people skilled in the kinds of techniques for appraising epistemic threats and vulnerabilities described in this paper. Such experts do exist and are currently distributed throughout various disciplines and professions (government and non-government) and employ different strategies for identifying and dealing with epistemic threats.

   For example, responsible journalists and journalism agencies engage in internal fact checking procedures to counter the spread of misinformation, and external fact checking organizations within universities and independent research centers engage in activities to encourage fact-based public discourse and promote accurate beliefs among the public and policy makers (Graves & Amazeen 2019). Psychologists investigate vulnerabilities in the processes by which individuals choose to consume information and form beliefs (Kahneman 2011, Klayman 1995), information security experts are trained in methodologies to prevent unauthorized use, disclosure, or alteration to private or sensitive information (von Solms & van Niekerk 2013), and public health experts are well familiar with the challenges of coordinating mass collective action to address public health crises ranging from disease eradication by vaccination to the current COVID-19 pandemic (WHO 2020).

   It is important to draw on a diversity of viewpoints when assembling a community of epistemic security experts in order to attend to the wide variety of epistemic threats and vulnerabilities that face a heterogeneous society.

6. **Invest in building and curating multidisciplinary epistemic security research groups and expert networks.** Epistemic security experts are embedded within separate and diverse professions and often have limited capacity to respond to (or to help to preemptively mitigate) epistemic threats. We recommend establishing dedicated

programs and institutions to train additional epistemic security experts and to bring together a diverse selection of epistemic security experts previously trained in other disciplines. As a long term goal, it may also be wise to establish epistemic security as its own discipline and for (government) decision-making bodies to employ dedicated epistemic security experts. Along these lines, UK Research and Innovation has recently issued a call to establish a UK Research Centre of Excellence for Protecting Citizens Online with the goal of exploring "holistic approaches to the development of privacy-enhancing technologies" and building an "interdisciplinary community [that provides] a clear single engagement point with enough critical mass to engage with government, industry and citizens."[32] Such efforts should be emulated globally.

5. Systems-mapping and 'red team' approaches to identifying and assessing epistemic threats and vulnerabilities

In this final section we describe practical techniques to help in identifying threats to a society's capacity for organizing timely and well-informed collective action and for fortifying epistemic security. These techniques include systems-mapping methods for the appraisal of epistemic vulnerabilities and identification of potential interventions (section 5.1) and red-teaming strategies for appraising potential interventions (section 5.2).

Systems-mapping generally describes a process by which the interactions between and within social epistemic systems are visually mapped. This is to provide a holistic overview of the system's infrastructures and constituent actors and to help identify epistemic vulnerabilities to the system. Systems-mapping can also help identify interventions on a system that might bolster the society's epistemic security.

Red-teaming broadly refers to the practice of identifying flaws, weaknesses, and failure points in a proposal by taking an opposing stance in order to rigorously challenge it. The term is derived from a simulated adversarial attack in which a "red team" tries to undermine a "blue team" with the intention of helping the blue team identify its weaknesses and failure points. Red teaming helps blue teams overcome biased viewpoints and broaden their solutions search.

These techniques build upon one another to help practitioners acquire greater understanding of the influences and vulnerabilities in a society's social epistemic infrastructure. Overall, they

[32] Research Centre of Excellence in Protecting Citizens Online

highlight the value of thinking about threats to epistemic security as systemic challenges, and of treating proposed solutions with initial skepticism.

We used systems-mapping methods and red-teaming strategies in our workshops to identify and evaluate epistemic vulnerabilites and interventions for six hypothetical scenarios - global health crisis, leader character assassination, state fake news, economic collapse, xenophobic ethnic cleansing, and epistemic babble - and we found the process to be useful from a state regulator's perspective. We provide running example from scenario 5 (xenophobic ethnic cleansing) in the supplementary boxes. Readers can find a full description of all six workshopped scenarios in Appendix 2.

5.1 Systems-mapping for the appraisal of epistemic vulnerabilities and interventions

Epistemic infrastructures can be incredibly complex; they are composed of a wide variety of actors with different goals and interests who are organized into complex and technologically-enabled networks of information processing and exchange. Developing an understanding of a society's epistemic infrastructure is central to identifying epistemic vulnerabilities, understanding how different threats and vulnerabilities interact with one another, and for mitigating epistemic threats to the system while avoiding negative second-, third- or higher-order consequences.

That being said, it is important to acknowledge that given the complexity of social epistemic infrastructures it is very challenging, and often impracticable, to fully understand any given system and all the relevant internal and external influences thereon. Therefore, the goal of a systems-based appraisal is to prioritize epistemic security efforts in order to minimize harms - both from epistemic threats and from adverse effects of interventions.[33]

We identify several steps that can help build a systems map that provides an understandable and useful characterization of a social epistemic system.

1. **Identifying the (potential) crisis**

Identifying a crisis (e.g. terrorist attack) or complex challenge (e.g. epistemic babble) provides a tractable starting point for building an understanding of the actors, factors, and interactions that may lead up to harmful outcomes. In the workshops we chose to focus on hypothetical

[33] A complementary approach is to apply complexity science to the epistemic system, which can highlight feedback mechanisms and sources of stability or instability, and propose systemic interventions. See for example the application of complexity science to concerns around "democratic backsliding" (Eliassi-Rad et al, 2020), which shares many concerns with, and provides a good complement to, the current report.

worst-case scenarios to broaden our scope and test the limits of our frameworks for thinking about epistemic security. However, we envision practitioners going through this assessment for potential crises in their area of responsibility (e.g. assessing epistemic threats that may take place during a major cyber-attack) or during ongoing crises, in which case the focus will be determined by the circumstances (e.g. monitoring epistemic security during a pandemic or a financial recession).

2. **Identifying relevant stakeholder and actor communities**

Various constituent actors in a social epistemic system may include official decision-making bodies, interest groups, public communities, institutions, technology and media platforms, and adversaries (or blunderers) with different goals. To keep a systems-based analysis tractable it is important to prioritize actors whose decisions hold outsized power with respect to achieving or preventing the outcome of concern. For example, it might be especially important to understand the epistemic security situation of police workers, health workers, environmental protection workers, or financial decision makers, depending on the nature of the crisis or dynamic.

3. **Identifying community vulnerabilities (and vulnerable communities)**

For the actors and other communities identified as relevant to the scenario under consideration, it is useful to analyse the different vulnerabilities of those communities, e.g. in terms of attention scarcity, fragility of trust, community motivated reasoning, cognitive bias, susceptibility to adversarial (or blunderer) interference, or other vulnerabilities. For example, children, working adults, and seniors might be differently vulnerable to misinformation and disinformation, due to different levels of media literacy, attention scarcity, misplaced trust or poor heuristics for establishing trust. Also, communities that have been historically mistreated or marginalized by authoritative decision-makers may have good reason to be distrustful of the information and decisions passed on by those authorities.[34]

4. **Evaluating how technologies can be used to either perpetuate or diminish epistemic vulnerabilities**

Within each community various technologies can exacerbate threats to epistemic security or enable (or hinder) activities that may bolster epistemic security and the organization of collective action. To help identify these technologies and the potential epistemic threats they pose, relevant questions to consider include:

[34] See footnote 30.

- What technologies are central to the society's systems of information dissemination and exchange?
- What media sources are available and how are they consumed by different communities?
- What search or filtering mechanisms exist and how are they used or misused?
- What role do algorithmic recommendations play in shaping information flows and network connections?
- Who has control over the shape and behaviour of communications networks? Who in the group has the ability and/or responsibility to contribute new information, moderate content, or control group membership? How are they technologically enabled?
- How might adversaries gain access to communication networks, and what could they do with the information?
- What technologies perpetuate the dissemination of reliable information?
- Are there technologies that can be used to minimize the influence of adversarial actors or blunderers?

Appendix 3 provides an initial list of epistemic threats posed by specific emerging technologies, and some proposed solutions for each, which may be used as a tool to help in identifying and assessing technology-based epistemic risks and vulnerabilities.

5. Mapping the socio-technological epistemic system

Visually mapping out the interaction within a technologically-enabled social epistemic system provides a useful tool for identifying further epistemic threats and vulnerabilities and for identifying points for effective intervention on the system. We suggest using a causal map to visualize the flow of authoritative and malign content in relevant epistemic networks. When building the map, it helps to keep the following factors in mind:

- **Information sources:** How does information get into the system (see knowledge sources (section 3.1) for inspiration). If the information source is testimonial, do you know anything about the source's interests, goals, or other characteristics that might influence their behaviour?
- **Information distribution:** How is information spread or amplified, and who has control over or might otherwise influence the spread of information?
- **Attenuation and evolution of content**: Do messages retain their content as they spread? How might informational content be lost, changed, or manipulated over time?
- **Counter-messaging**: Do any of the constituent actors engage in efforts to spread 'alternative-narratives' to counter the messages being spread by another group? For example, a human rights organization might flood social media platforms with stories

about atrocities being committed against refugees to counter the messages being perpetuated by a xenophobic extremist group.

- **Character and behaviour of information recipients**: Do information recipients ever become information sources or amplifiers? Do they change their beliefs and actions based on the information they receive? Are information recipients likely to trust information distributed by authorities? What factors might be underpinning or undermining this trust?
- **Technological influence**: How are different technologies used to amplify, undermine, or otherwise influence these factors?

Boxes 5.1a and 5.1b provide an example from workshop scenario 5. See Appendix 2 for additional examples.[35]

Box 5.1 a
Systems-mapping scenario 5 - xenophobic ethnic cleansing

To recap, scenario 5 describes a society in which a radical xenophobic group wishes to turn a country's population against a minority refugee community. First, the radical group launches its campaign with a low-grade chemical attack on a school and second, they use video alterering technologies and targeted social media blasts to pin blame on the minority community and to rally public animosity towards the refugees. Note the second phase of the radical group's plan constitutes an adversarial attack against the society's epistemic infrastructure; the group widely distributes false information with the intention of manipulating public and policy-maker opinion and behavior toward the refugees.

We began by identifying the harm that the adversarial actor wished to achieve (public animosity towards refugees). Doing so leads naturally to listing the most relevant actors for appraisal and then mapping out their interactions (technologically-enabled or otherwise) based on our appraisal of each group's epistemic vulnerabilities and technological capabilities.

Actors
● Radical xenophobic group

[35] This simple causal map provides only an initial sketch of the relevant system. For a more in-depth analysis, it is possible to go further and develop a hierarchical causal map, for example by focusing, initially, on the actions of the malicious actors. Such hierarchical causal maps can provide a richer picture of how specific actions give rise to the harm and of how actions by different actors interact.

- The public, whose opinions, decisions, and actions the radical group aims to influence
- Public authorities

Other Communities/Groups/Audiences

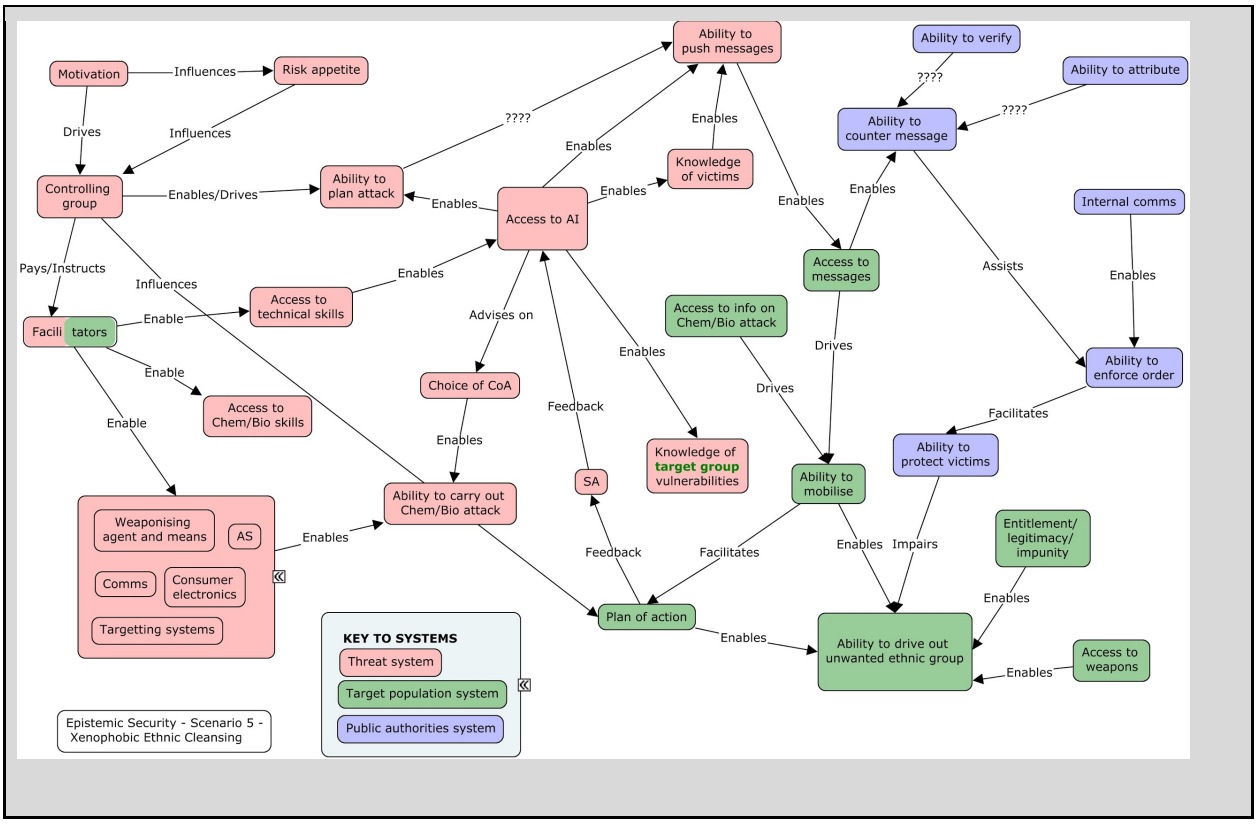
- Minority refugee community - not identified as an actor as it is not the direct target of the radical groups adversarial effort.

A more complete analysis of such a scenario might result in the addition of the minority refugee community as an actor. Similarly it might break down one or more of the above actors into a more granular set of actors. The workshop scenario analysis was bounded by time and effort and so we prioritized the three actors above as key.

The resulting causal map of systems is depicted below.

Systems Map

Threat system - the radical xenophobic group
Target population system - the public, whose opinions, decisions, and actions the radical group aims to influence
Public Authorities system - policy-makers, government officials, and groups enabled by the officials (e.g. police force, military, health workers etc.) to counteract the radical group's efforts and to prevent further adversarial attack. If the radical group is successful public authorities may also be hijacked to carry out the radical group's wishes.



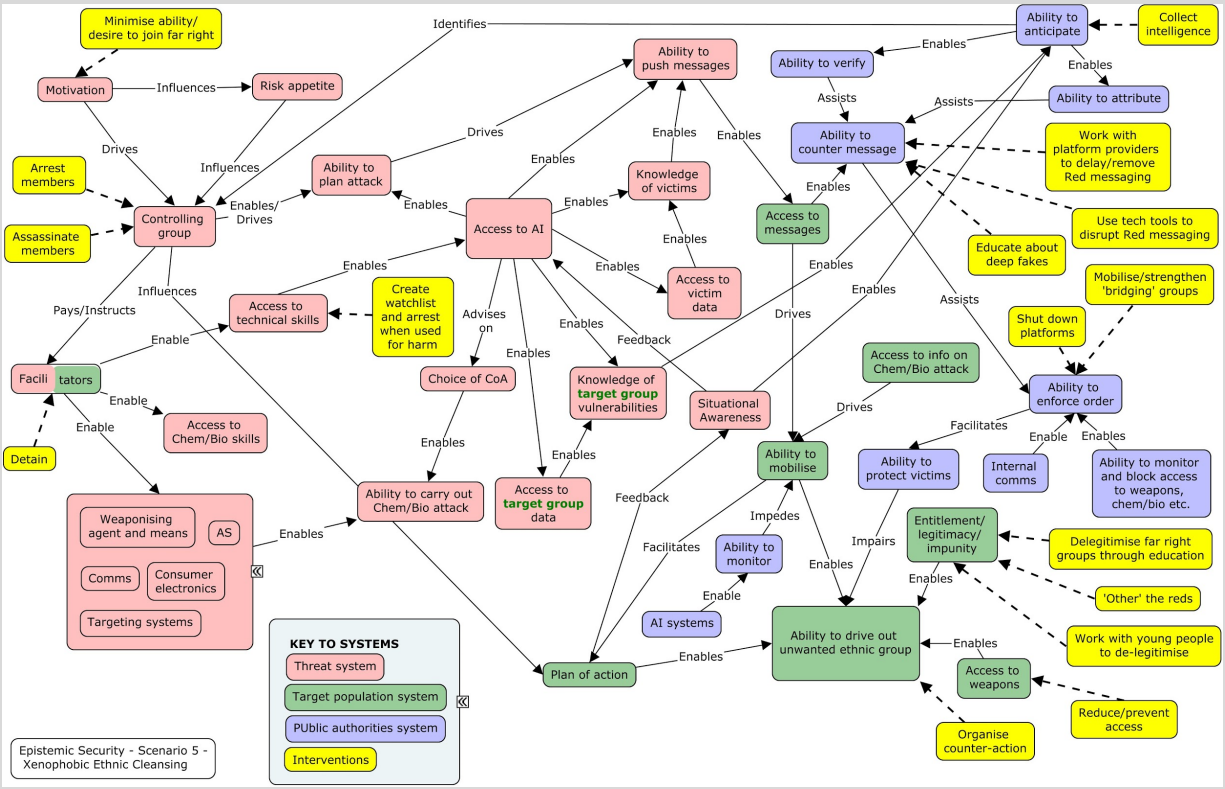
Once the epistemic systems and vulnerabilities implicated in a crisis or dynamic have been mapped, it is possible to identify and begin evaluating interventions that can bolster epistemic security and mitigate threats. A more comprehensive causal map will depict a greater number of potential intervention points and will make it easier to evaluate which intervention points will likely yield the most effective interventions.

However, it is not necessary to hold off on intervention identification until systems-mapping is ‘complete’. Contemplating different interventions brings to light new ways in which different actors interact and influence one another. Map building is an iterative process that benefits from switching back and forth between system assessment and intervention identification.

Box 5.1 b

Adding potential interventions.

Here we add potential interventions to the system map sketched in section 3.1.4 . Note that when compared with the original map, a number of enhancements can be seen, for example the addition of a set of blue boxes in the bottom right. The main change has been the identification of a larger number of possible interventions. A comparison with the original scenario system map will highlight that a number of items in the original map have been re-coded as interventions in this revised version.



The interventions in yellow have not yet been appraised for their feasibility (including ethical, legal and financial feasibility) or the efficacy of their impact. See section 5.2.

5.2 Red-Teaming

When assessing potential interventions, it is crucial to consider the impacts of the interventions and to assess the various ways in which an intervention could backfire. Epistemic interventions often exist in a delicate balance of tradeoffs (e.g. between freedom of speech and prevention of harm). Therefore they may easily exacerbate epistemic vulnerabilities (e.g. erosion of trust in state-provided information) if not well-vetted. We recommend coupling the systems-based mapping presented in sections 5.1 with "red team" strategies for testing the feasibility, efficacy, robustness, and potential unintended consequences of different interventions. Recall, red-teaming broadly refers to the practice of identifying flaws, weaknesses, and failure points in a proposed intervention to an epistemic system by taking an opposing stance in order to rigorously challenge it.

A well-functioning red team will be a diverse group of individuals composed of a variety of stakeholders and of experts from a range of relevant disciplines. It is a standard tenet of social epistemology that diverse viewpoints in a decision-making group encourage careful deliberation (Holst & Molander 2017; Sunstein & Hastie 2015) and that the examination of critiques and dissenting opinions yields more well-considered judgments (Longino 2002). In particular, diverse representation of stakeholders on red teams ensures that the values and interests that motivate different actors are accounted for in the assessment. Understanding the values motivating red team and blue team actors, and understanding the coherence of values within those teams, will provide significant insight into each group's capacity for timely collective action.

Furthermore, incorporating diverse stakeholder viewpoints ensures that the interests of communities most likely to be affected are not overlooked in the assessment. For instance, if an epistemic security-defending entity, or 'blue team' (government or otherwise), is developing strategies to mitigate epistemic threats posed to social media users by online echo chambers, social media users should be involved in 'red-teaming' the authority's proposed strategies to look for flaws or unintended adverse consequences.[36] Similarly, if a minority group is being villainized by adversarial actors, members of the minority group ought to be included on the task force to address the threat.

[36] We attempted to follow this advice in our workshops, drawing on expertise from academics across a range of fields (media studies, psychology, technology risk assessment), from security experts and engineers in corporates, NGOs and government labs, from national and international defence communities, and from technology regulators. However, we acknowledge that even this mix can and should be greatly expanded, especially when dealing with ongoing or imminent crises.

In this section we describe three red-teaming strategies - **S**trengths, **W**eaknesses, **O**pportunities and **T**hreats (SWOT) charts, pre-mortems, and futures wheels - which we found useful for evaluating potential interventions to aid in maintaining or restoring epistemic security in a society. We used these techniques because (i) the SWOT forces consideration of an intervention from the perspective of a particular actor, (ii) pre-mortem forces consideration of intervention failure, and (iii) the futures wheel forces consideration of potential 2nd and 3rd order intervention effects. The different techniques support the 'red team' in taking a number of different views and provide a lightweight structure (or handrail) to the 'red team' analysis.

Box 5.2 provides an example of each red-teaming strategy with respect to a single intervention proposed for scenario 5 - implementing a public education campaign to delegitimize the rhetoric of far right extremist groups. Each proposed intervention or different combination of interventions should undergo its own red team evaluation. If multiple interventions are identified that are likely to be robust and effective, red-teaming strategies should be employed to evaluate the interventions when used in concert; it is possible that additional consequences emerge from the simultaneous employment of multiple interventions.

SWOT

When considering the impacts of interventions, it is important to recognise that the intervening actor (e.g. the state or blue team) is not the only actor operating in/on the system; Internal system dynamics, blunderers, or a creative adversary might act to foil, re-direct or co-opt the intervention. The SWOT technique helps to evaluate how well different actors are equipped to respond to (or to initiate) an intervention.

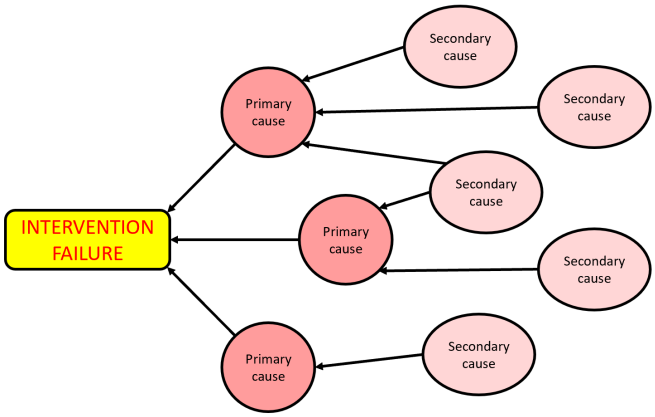
Filling out a SWOT helps to identify the **S**trengths, **W**eaknesses, **O**pportunities and **T**hreats that pertain to specific actors (e.g red teams and blue teams) and to evaluate the likely efficacy of a potential intervention given those factors. Strengths and weaknesses are internal factors or actor characteristics that will help or hinder the actor in achieving its objectives. These may include team composition, past experience, or access to physical, financial, or technological resources. On the other hand, opportunities and threats are external influencing factors such as economic trends, cultural change, new legislation, the adoption of a certain technology by another group, and national or international events.

STRENGTHS Which of our characteristics will help us achieve our objectives?	WEAKNESSES Which of our characteristics will work against the achievement of our objectives?
OPPORTUNITIES What events or developments might help us achieve our objectives?	THREATS What events or developments might work against the achievement of our objectives?

Generally speaking, the more an actor enjoys alignment of strengths and opportunity, the better equipped it is to make more aggressive or risky moves. For instance, an adversary that enjoys alignment of strengths and opportunity will be more robust to intervention and more likely to make more aggressive attacks. Similarly, blue teams that enjoy greater strengths and opportunities are more well-equipped to take offensive action against adversarial influence. Inversely, blue teams plagued with a greater proportion of weaknesses and threats may be wise to act more defensively or to implement more defensive interventions.

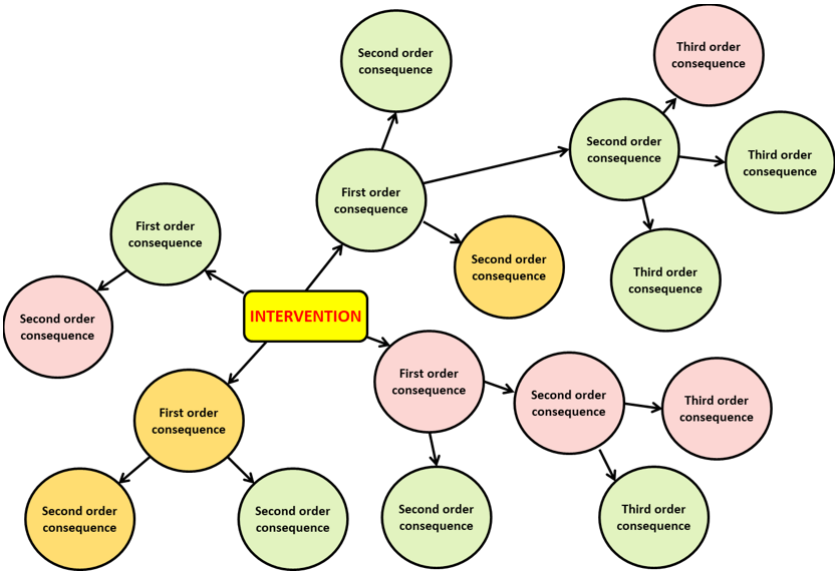
Pre-mortem

The pre-mortem red-teaming technique starts by assuming that an intervention has failed and then works backwards to construct causal pathways that could lead to such a failure. With the possible pathways to failure sketched out, various modifications to the intervention structure or enforcement are proposed to prevent the same failure when the intervention is actually rolled out. Conducting a pre-mortem may illustrate that a particular intervention has too many potential paths to failure to be addressed for a timely and responsible implementation.



Futures Wheel

The futures wheel red-teaming technique is the inverse of a pre-mortem. It starts with a potential intervention and works outwards to identify possible first-, second- and higher-order consequences of an intervention. First the initial consequences of an intervention (first-order impacts) are traced out. Both probable impacts and low-probability/high-impact consequences are included and identified as positive (green), negative (red), or neutral or uncertain (yellow) outcomes. Then, by considering the first-order impacts as interventions in and of themselves, a second level of consequences are added (thus mapping out second-order impacts). This process can continue for third- and higher-order impacts as long as new consequences are identified that are considered robust or particularly important (e.g. because they could lead to unacceptable risk). Once multiple orders of potential intervention consequences are identified, it is easier to pick out interventions that are more likely to yield positive and neutral outcomes. These interventions should be explored further.



Box 5.2

Red-teaming scenario 5 - xenophobic ethnic cleansing - Education Campaign

<u>SWOT</u> - evaluating actor strengths, weaknesses, opportunities and threats

In the workshop we applied the SWOT analysis to the interventions 'work with platforms providers to delay/remove Red messaging' and 'Use tech tools to disrupt Red messaging'. Both interventions were concerned with Red messaging and concerned a third party.

We filled out SWOT charts for relevant actors: the guardians of epistemic security (the blue team), and the xenophobic extremist group (the red team).

Blue Team

STRENGTHS **Which of our characteristics will help us achieve our objectives?** • *Legitimacy and trust* • *Leverage over media companies* • *Ability to use the combined assets of the State* • *Access to legitimate authority / authority-linked messaging media, => increased ability to intervene* • *Legitimate access to signals intel enables detection* • *Established technical capabilities (eg. GCHQ)* • *Access to good technical skills* • *May stem mainstream awareness/publicity of far right ideas and allow alternative ideas space to become established* • *Govt data of population gives good audience insight and reach* • *Buying power and scale*	**WEAKNESSES** **Which of our characteristics will work against the achievement of our objectives?** • *Govt systems not dynamic or agile enough to have tempo* • *Too risk-averse* • *Lack of trust among subset of population* • *Foolish rhetoric by authority-related figures* • *Difficult to attract relevant talent (better pay in industry)* • *We may lag behind the cutting edge in our tech tools* • *Have to monitor and recognise Red message before we can disrupt* • *Difficult legal landscape: free speech, IP => could stop govt intervention* • *No access to encrypted channels / legal constraints* • *The state is a slow, ponderous organisation which takes a long time to make decisions far too late*
OPPORTUNITIES **What events or developments might help us achieve our objectives?** • *Boost innovation and creativity in tech sector* • *Can infiltrate groups to improve counter-messaging* • *We could insert our own messages* • *Could change Red message to suit our own agenda* • *Partnership with messaging / media platforms* • *(If handled well) good PR opportunity for both govt and tech firms* • *Tracking messaging can lead back to ringleaders and arrests* • *Ability to legislate* • *Ability to spend vast amounts of money* • *Ability to work with international partners*	**THREATS** **What events or developments might work against the achievement of our objectives?** • *Authorities are seen as 'Big Brother' by population* • *Builds persecution complex (potential martyrdom of those being blocked – anti-establishment sentiment). May then embolden 'underdog' campaigns* • *Drives underground far right messages and ideas – harder to track and establish network maps* • *May lead to a false perception that there is no problem / presence from general populace / politicians and in turn less resources are allocated* • *Our tech tools are hacked* • *Other crisis emerges to pull limited capabilities away* • *Far right and other threat actors will act much more quickly, dynamically and secretively than us* • *Far right moves its messages into other fora*

***Red Team** - Xenophobic radical group*

STRENGTHS **Which of our characteristics will help us achieve our objectives?**	WEAKNESSES **Which of our characteristics will work against the achievement of our objectives?**
• *Ability to say 'unacceptable' things* • *Ability to use illegal or illegitimate/unethical methods* • *The bomber always gets through* • *Ability to take advantage of an opportunity / crisis* • *Can hop between platforms and media types to avoid detection* • *Validates our strategy as a potent force – we must be effective if they are disrupting us* • *Agility and distribution* • *Can evolve faster than the state* • *We can manipulate and utilise online very well. We can scare people* • *Ability to tap into underlying population discontent and distrust*	• *Hard to know when info security has been compromised* • *Cannot see what Law Enforcement is doing* • *Far right does not have access to the tech expertise needed to counter authorities* • *Location of servers to avoid government interception and prosecution (cost)* • *Fewer / less sophisticated resources and channels* • *Very hard to control our direction and strategy (group is disparate)* • *Limited depth of capability, capacity*
OPPORTUNITIES **What events or developments might help us achieve our objectives?**	THREATS **What events or developments might work against the achievement of our objectives?**
• *False flag messaging, creating multiple dilemmas for the authorities* • *Deliberately trigger your response to flood platform with topic, waste government resources* • *Move to new or encrypted platforms and open new opportunities for exploitation and reach* • *Impact new audiences through other platforms* • *Diverse networks enable a 'whack-a-mole' system – your followers can find you in places where your presence is not yet known to authorities – constantly avoiding detection.* • *Go offline to avoid tech surveillance and disruption – word of mouth and leaflets* • *"We are being silenced! Back our fight against the [ethnic minority]!" – new recruits* • *Develop codes and covert ways of communicating ideas* • *Take advantage of <u>every</u> crime committed by 'the Other' and saying: "We told you so"* • *Make use of criminal tools and networks* • *If government is attempting to disrupt, use decoys and deception to undermine their credibility* • *Encourage lone wolf attacks based on less central organisation across key platforms* • *Can cast any counter-message as 'government propaganda' or '[ethnic group] propaganda'* • *Can 'immunise' population by pre-warning of counter-messages*	• *Cannot trust the online systems we need to keep advantage. Offline slower* • *Infiltration by the authorities* • *Friends or family may rat out* • *Defections / 'ratting out' when faced with prison sentences* • *Volunteers drift away through fear of government online counter surveillance and interdiction* • *Members get worn down or become complacent, exposing infosec risk* • *Own group members might be influenced by counter-campaign* • *Competition with other FRE groups – need to be seen as 'the figurehead'* • *Far right loses its audience as authorities shut it down*

The Blue SWOT shows a concern for speed of response and a consistency of action which is balanced; whilst the Red SWOT shows strengths and opportunities around agility and speed of action. A comparison between the charts highlights how Red opportunity aligns with Blue weakness.

<u>Premortem</u> - describing how an intervention might fail

In the workshop we quickly applied the pre-mortem approach across the set of interventions to identify what might be the outcome of a failure. Time prevented creating a causal map for

these failures; but it served the important function of getting the group comfortable with expressing why an intervention might fail and thus looking for potential weaknesses in the ideas.

There were 3 interventions related to 'delegitimising the far-right group' so we grouped them together as below.

Interventions:

- *Delegitimise far right groups through education*
- *Work with young people to delegitimise*
- *'Other' the Reds*

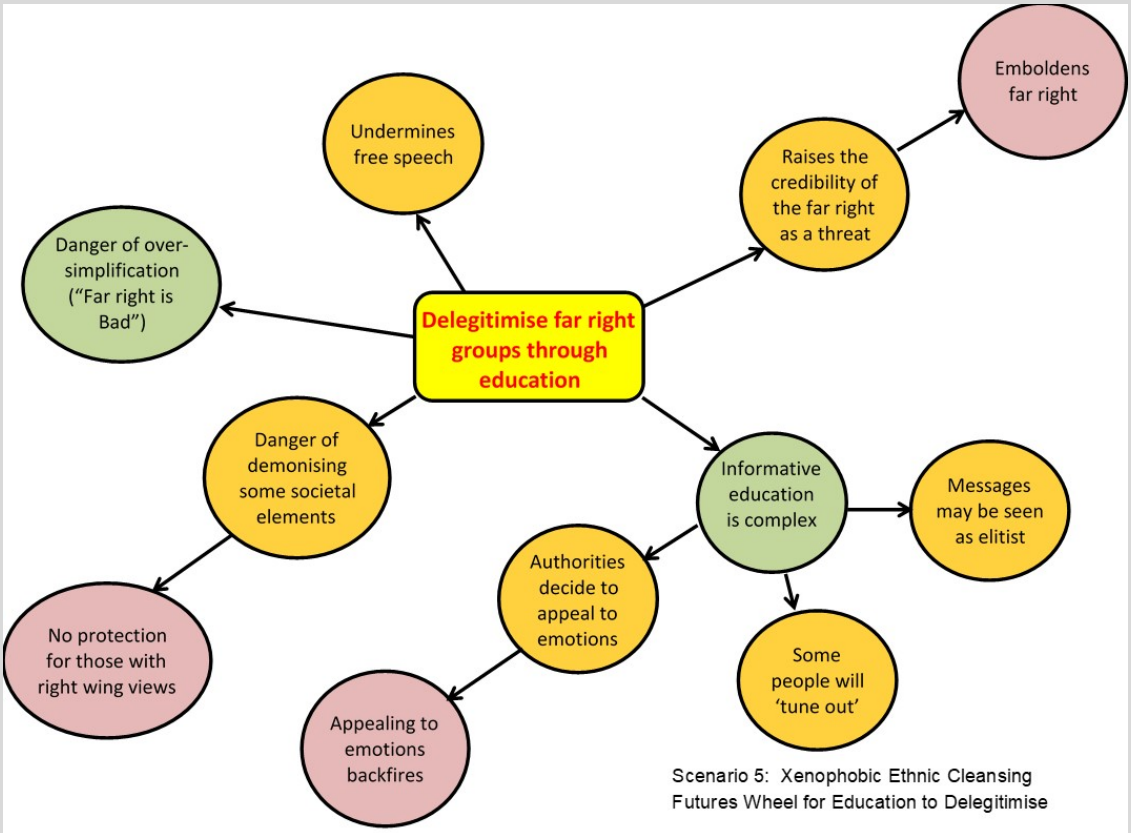
Failures:

- *Attempt to delegitimise far right backfires, delegitimising the government and bringing about the election of a far right government.*
- *'Young people' may not be the primary target audience of Red. Likely to reach older, disadvantaged segments of the population.*
- *'Othering' generates sympathy for right wing – Tommy Robinson 2.0.*

The pre-mortem identified a set of potential failures; further analysis of the causes can identify risks associated with the intervention. These risks can then be analysed and, where feasible, mitigation plans developed.

Futures Wheel - mapping intervention consequences

We started with the intervention (delegitimize far right groups through education) and mapped out potential first-, second-, and higher-order consequences.



The futures wheel identifies potential negative (red) consequences. It may be possible that there is no reasonable modification, and the negative consequence has to be acknowledged as a possible risk.

6. Conclusion

Maintaining a society's epistemic security is a multifaceted challenge. Well-informed and timely group decision-making requires knowledge drawn from a variety of sources, including experience, memory, reason, and most of all, a diverse assortment of testimony. Our social epistemic infrastructure - the collection of systems, processes, and actors that influence how knowledge is produced, distributed, acquired, and modified within a society - are susceptible to a variety of threats and vulnerabilities. These include attention scarcity, conflicting community values, the fragility of trust, and adversarial action. Technologies embedded in our social epistemic infrastructures are intentionally and accidentally used to both pacify and exacerbate threats to epistemic security, but we identify a trend whereby recent technological advances have led to a net increase in costs for establishing informed collective action.

This report presents insights drawn from a series of workshops conducted to explore challenges to collective knowledge acquisition and knowledge-based decision-making. We explain that because many narrowly targeted fixes to current epistemic threats often risk detrimental second-, third-, and higher order consequences, epistemic threats should not be viewed as a laundry list of challenges accompanied by a list of prescribed fixes. Rather, more holistic approaches to identifying and analyzing threats to epistemic security, such as "red team" and "systems-mapping" approaches that employ diverse teams of experts and interest group representatives, are necessary to accurately characterise a social epistemic infrastructure and its epistemic vulnerabilities and strengths. We suggest that such holistic appraisals, coupled with long-term investment in technological and institutional solutions, are likely to present effective strategies for evaluating and mitigating threats to a democracy's epistemic security. This security is critical to a society's ability to organize collective action on the basis of timely and reliable information in a technologically advanced world.

Acknowledgements

We would like to thank the following people their time and effort in providing internal reviews of multiple drafts of this report:
Haydn Belfield (Centre for the Study of Existential Risk, University of Cambridge); **Julian Huppert** (Intellectual Forum, University of Cambridge); **Clarissa Rios Rojas** (Centre for the Study of Existential Risk, University of Cambridge); **Jess Whittlestone** (Leverhulme Centre for the Future of Intelligence, University of Cambridge) and to the following people for providing valuable feedback on the report at different stages of the process:
Jaime Sevilla (Centre for the Study of Existential Risk, University of Cambridge; University of Aberdeen); **Sam Weiss Evans** (Program on Science, Technology, and Society, Harvard University)

Bibliography

Aimes, C. (2011). Memo reveals intelligence chief's bid to fuel fears of Iraqi WMDs. *The Guardian.* Retrieved from https://www.theguardian.com/uk/2011/jun/26/intelligence-chief-iraqi-wmds

Ajder, H., Patrini, G. & Cullen, L. (2019). *The state of deepfakes: Landscape, threats, and impact.* Deeptrace. Retrieved from https://sensity.ai/mapping-the-deepfake-landscape/

Anderson, E. (2011). Democracy, public policy, and lay assessments of scientific testimony. *Episteme, 8*(2), 144-164. doi:10.3366/epi.2011.0013

Anderson, H., & Wagenknecht, S. (2013). Epistemic dependence in interdisciplinary groups. *Synthese, 190*(11), 1881-1898. doi:10.1007/s11229-012-0172-1

Arkin, F. (2019). Dengue vaccine fiasco leads to criminal charges for researcher in the Philippines. *Science*. Retrieved from https://www.sciencemag.org/news/2019/04/dengue-vaccine-fiasco-leads-criminal-charges-researcher-philippines

Asch, S. E. (1951). "Effects of Group Pressure upon the Modification and Distortion of Judgments." In *Groups, Leadership and Men,* edited by Harold Guetzkow, 222–236. Pittsburgh, Pennsylvania: Carnegie Press.

Benkler, Y., Faris, R. & Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. UK: Oxford University Press.

Biddle, J. B. and Kukla, R. (2017). The Geography of Epistemic Risk, in K. C. Elliott and T. Richards (eds), *Exploring Inductive Risk: Case Studies of Values in Science*, New York: Oxford University Press, pp. 215–37.

Blair, A. (2012). *Information overload's 2,300-year-old history.* Harvard business review online resources. http://blogs.hbr.org/cs/2011/03/information_overloads_2300-yea.html. Accessed 06 Feb 2020.

Bradshaw, S., Neudert, L. & Howard, P. N. (2018). *Government response to malicious use of social media.* NATO STRATCOM COE. Retrieved from https://www.stratcomcoe.org/government-responses-malicious-use-social-media

Brundage, M. & Avin, S. et al. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. Retrieved from https://maliciousaireport.com/

Bucher, T. (2018). *If... Then: Algorithmic Power and Politics.* UK: Oxford University Press.

Cárdenas, A. A., Amin, S., & Sastry, S. (2008). Research Challenges for the Security of Control Systems. In HotSec. https://static.usenix.org/events/hotsec08/tech/full_papers/cardenas/cardenas.pdf

Carlson, M. (2020). Journalistic epistemology and digital news circulation: Infrastructure, circulation practices, and epistemic contests. *New Media & Society, 22*(2), 230-246. doi:https://doi.org/10.1177/1461444819856921

Centre for Data Ethics and Innovation (2020). *Review of online targeting: Final report and recommendations.* https://www.gov.uk/government/publications/cdei-review-of-online-targeting

Charlton, E. (2019). How Finland is fighting fake news-in the classroom. In World Economic Forum. https://www.weforum.org/agenda/2019/05/how-finland-is-fighting-fake-news-in-the-classroom/

Chaslot, G. (2017) How YouTube's A.I. boosts alternative facts. Medium. https://medium.com/@guillaumechaslot/how-youtubes-a-i-boosts-alternative-facts-3cc276f47cf7

Chesney, R. & Citron, D. K. (2018). *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*. 107 California Law Review 1753 (2019), U of Texas Law, Public Law Research Paper No. 692, U of Maryland Legal Studies Research Paper No. 2018-21, Available at SSRN: https://ssrn.com/abstract=3213954 or http://dx.doi.org/10.2139/ssrn.3213954

Chessen, M. (2017). *The MADCOM future: How artificial intelligence will enhance computational propaganda, reprogram human culture, and threaten democracy... and what can be done about it.* Atlantic Council, Dinu Patriciu Eurasia Center, and Brent Scowcroft Center on International Security. Retrieved from https://www.atlanticcouncil.org/wp-content/uploads/2017/09/The_MADCOM_Future_RW_0926.pdf

Collins, A. (2019). *Forged Authenticity: Governing Deepfake Risks*. Lausanne: EPFL International Risk Governance Center.

Collins, H. M. (1992). *Changing Order: Replication and Induction in Scientific Practice*. Chicago: The University of Chicago Press.

Comprehensive report of the special adviser to the DCI on Iraq's WMD. (2004). Central Intelligence Agency. Retrieved from https://www.cia.gov/library/reports/general-reports-1/iraq_wmd_2004/

Conee, E. (2010). Rational disagreement defended. In R. Feldman & T. A. Warfield (Eds.), Disagreement (pp. 69-90). Oxford Scholarship Online: Oxford University Press.

Day, F. (2015). *You're Never Weird on the Internet (almost): A Memoir.* Simon and Schuster.

Douglas, H. (May 2017). *The materials for trust-building in expertise*. Presentation at the HPS Departmental Seminar, University of Cambridge.

Eliassi-Rad, T., Farrell, H., Garcia, D. et al. (2020) What science can do for democracy: a complexity science approach. *Humanities & Social Sciences Communications* 7, 30. doi:10.1057/s41599-020-0518-0

Epstein, S. (1995). The construction of lay expertise: AIDS activism and the forging of credibility in the reform of clinical trials. *Science, Technology, & Values, 20*(4), 408-437, http://www.jstor.org/stable/689868

European Commission (2018). *A multi-dimensional approach to disinformation: Report of the independent High level Group on fake news and online disinformation.* Luxembourg: Publications Office of the European Union. Retrieved from https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

Farell, H., & Schneier, B. (2018). Common-knowledge attacks on democracy. *Berkman Klein Center Research Publication No. 2018-7.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3273111##

Funke, D., & Famini, D. (2018). *A guide to anti-misinformation actions around the world. Poynter.* https://www. poynter.org/news/guide-anti-misinformation-actions-around-world.

Furman, K. (2020). On trusting neighbors more than experts: An ebola case study. *Frontiers in Communication, 5*(23). doi:10.3389/fcomm.2020.00023

Gersham, J. (Feb 2020). Tech Platforms aren't bound by first amendment, appeals court rules. *The Wall Street Journal.* Retreived from https://www.wsj.com/articles/tech-platforms-arent-bound-by-first-amendment-appeals-court-rules-11582748988

Giles, J. (2006) The trouble with replication. *Nature 442,* 344–347. https://doi.org/10.1038/442344a

Goldman, A. I. (2010). Epistemic relativism and reasonable disagreement. In R. Feldman & T. A. Warfield (Eds.), Disagreement (pp. 187-215). Oxford Scholarship Online: Oxford University Press.

Graves, L. (2016). *Deciding What's True: The Rise of Political Fact-Checking in American Journalism.* New York: Columbia University Press.

Graves, L., & Amazeen, M. A. (2019). Fact-Checking as Idea and Practice in Journalism. *Oxford Research Encyclopedia of Communications.* Retrieved from https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-808

Hedstrom, M. (2006). Epistemic infrastructure in the rise of the knowledge economy. In M. Hedstrom & J. L. King (Eds.), *Advancing knowledge and the knowledge economy* (pp. 113-134). Cambridge, MA: MIT Press.

Hitchins, D. K. (2008). *Systems engineering: a 21st century systems methodology.* John Wiley & Sons.

Holst, C., & Molander, A. (2017). Public deliberation and the fact of expertise: making experts accountable. Social Epistemology, 31(3), 235-250. doi:10.1080/02691728.2017.1317865

House of Commons Digital, Culture, Media and Sport Committee (2019). *Disinformation and 'fake news': Final Report Eighth Report of Session 2017–19*. Retrieved from https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf

Howard, P. N. (2020). *Lie Machines*. Yale University Press.

Howard, P. N., & Woolley, S. C. (Eds.). (2018). *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media.* UK: Oxford University Press.

Hubert, A., Bright, J., & Howard, P. N. (August 2020). *Social Media Junk News on Hydroxychloroquine and Trust in Science.* University of Oxford Retrieved from https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/07/ComProp-Coronavirus-Misinformation-Weekly-Briefing-03-08-2020.pdf

Humphreys, P. (2004). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*: Oxford University Press.

Hwang, T. (2019). *Maneuver and Manipulation.* United States Army War College Press.

Hwang, T. (2020) Deepfakes: A Grounded Threat Assessment. Center for Security and Emerging Technology, https://cset.georgetown.edu/research/deepfakes-a-grounded-threat-assessment/

Jiang, R., Chiappa, S., Lattimore, T., György, A., & Kohli, P. (2019). Degenerate feedback loops in recommender systems. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 383-390).

John, S. (2019). Science, truth and dictatorship: Wishful thinking or wishful speaking? Studies in History and Philosophy of Science. doi:https://doi.org/10.1016/j.shpsa.2018.12.003

Jourova, V. (2020). Speech of Vice President Věra Jourová on countering disinformation amid COVID-19 "From pandemic to infodemic" Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/speech_20_1000

Kahneman, D. (2011). *Thinking, fast and slow (1st ed)*. New York: Farrar, Straus and Giroux.

Kennedy J. (2017). How Drone Strikes and a Fake Vaccination Program Have Inhibited Polio Eradication in Pakistan: An Analysis of National Level Data. *International journal of health services : planning, administration, evaluation*, *47*(4), 807–825. https://doi.org/10.1177/0020731417722888

Kertysova, K. (2018). Artificial intelligence and disinformation. *Security and Human Rights* Volume 29 Issue 1-4. P.55-81. https://doi.org/10.1163/18750230-02901005

Klayman, J. (1995). Varieties of Confirmation Bias. *Psychology of Learning and Motivation, 32,* 385-418.

Kelp C., Douven I. (2012) Sustaining a Rational Disagreement. In: de Regt H., Hartmann S., Okasha S. (eds) EPSA Philosophy of Science: Amsterdam 2009. The European Philosophy of Science Association Proceedings, vol 1. Springer, Dordrecht

Kusch, M. (2002). Knowledge by agreement: The programme of communitarian epistemology. New York: Oxford Univ. Press.

Lakoff, G. (2016). *Moral politics: How liberals and conservatives think (Third edition)*. Chicago IL: The University of Chicago Press.

Landhuis, E. (2016) Scientific literature: Information overload. *Nature 535*, 457–458. https://doi.org/10.1038/nj7612-457a

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Schudson, M. (2018). The science of fake news. *Science, 359*(6380), 1094-1096.

Lin, H. (2019). The existential threat from cyber-enabled information warfare. *Bulletin of the Atomic Scientists, 75*(4), 187-196, DOI: 10.1080/00963402.2019.1629574

Longino, H. E. (2002). *The Fate of Knowledge.* New Jersey: Princeton University Press.

Malone, T. W. (2018). *Superminds: The surprising power of people and computers thinking together.* One World Publications.

Mazarr, Michael J., Bauer, R. M., Casey, A., Heintz, S., and Matthews, L. J. (2019). *The Emerging Risk of Virtual Societal Warfare: Social Manipulation in a Changing Information Environment.* Santa Monica, CA: RAND Corporation. Retrieved from https://www.rand.org/pubs/research_reports/RR2714.html.

Miller, B., & Record, I. (2013). Justified belief in a digital age: On the epistemic implication of secret internet technologies. *Episteme, 10*(2), 117-134. doi:10.1017/epi.2013.11

Navin, M. C. (2017). Prioritizing religion in vaccine exemption policies. In K. Vallier & M. Weber (Eds.), Religious Exemptions: Oxford University Press.

O'Connor, C., & Weatherall, J. O. (2019). The misinformation age: How false beliefs spread. New Haven: Yale University Press.

O'Neill, O. (2018). Linking Trust to Trustworthiness, *International Journal of Philosophical Studies, 26*(2), 293-300, DOI: 10.1080/09672559.2018.1454637

Pomerantsev, P. (2019). *This Is Not Propaganda: Adventures in the War Against Reality.* New York: Public Affairs.

Reviving the US CDC. (2020, May 16). *The Lancet, 397*(10236), 1521. https://doi.org/10.1016/S0140-6736(20)31140-5

Rini, R. (2019). *Deepfakes and the Epistemic Backstop.* Retrieved from https://philarchive.org/archive/RINDATv1

Roetzel, P.G. (2019). Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research, 12* (2), 479-522. https://doi.org/10.1007/s40685-018-0069-z

Roozenbeek, J. & van der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation, *Journal of Risk Research, 22*(5), 570-580, DOI: 10.1080/13669877.2018.1443491

Runciman, D. (2018). *How democracy ends (First US edition)*. New York: Basic Books.

Scharff, D. P., Mathews, K. J., Jackson, P., Hoffsuemmer, J., Martin, E., Edwards, D. (2010). More than Tuskegee: Understanding mistrust about research participation. *J Health Care Poor Underserved, 21*(3), 879-897. doi:10.1353/hpu.0.0323

Scott, J. C. (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press.

Simon, H. (1957). *Models of Man*. New York: John Wiley.

Simon, H. (1971). Designing Organisations for an Information Rich World. In M. Greenberger (Ed.) *Computer, communications, and public interest*. Baltimore, MD: The Johns Hopkins Press.

Shadbolt, N., O'Hara, K., De Roure, D., & Hall, D. W. (2019). *The Theory and Practice of Social Machines*: Springer.

Sunstein, C. R. (2007). *Republic.com 2.0.* Princeton: Princeton University Press.

Sunstein, C. R., and R. Hastie. (2015). Wiser: Getting beyond Groupthink to Make Groups Smarter. Boston, MA: Harvard Business Review Press

Taber, C. S., and M. Lodge. 2006. “Motivated Skepticism in the Evaluation of Political Beliefs.” American Journal of Political Science 50 (3): 755–769. doi:10.1111/j.1540-5907.2006.00214.x.

Urban, T. (2019) The Story of Us. https://waitbutwhy.com/2019/08/story-of-us.html

U.S. House of Representatives. (2019). House Intelligence Committee To Hold Open Hearing on Deepfakes and AI: The National Security Challenge of Artificial Intelligence, Manipulated

Media, and “Deepfakes” [Press release]. Retrieved from https://intelligence.house.gov/news/documentsingle.aspx?DocumentID=657

Vaidyanathan, Siva (2018). *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy.* UK: Oxford University Press.

von Solms, R. & van Niekerk, J. (2013). From information security to cyber security. *Computers & Society, 38,* 97-102. doi: https://doi.org/10.1016/j.cose.2013.04.004

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

Washington, H. (2008). *Medical apartheid : The dark history of medical experimentation on Black Americans from colonial times to the present* (1st Anchor books (Broadway Books) ed.). New York: Anchor Books.

Whitaker, R.M., Colombo, G.B. & Rand, D.G. (2018) Indirect Reciprocity and the Evolution of Prejudicial Groups. Nature, Sci Rep 8, 13247 . https://doi.org/10.1038/s41598-018-31363-z

WHO. (2020). I*nfodemic Management - Infodemiology.* Retrieved from https://www.who.int/teams/risk-communication/infodemic-management

Winsberg, E., Huebner, B., & Kukla, R. (2014). Accountability and values in radicaly collaborative research. *Studies in History and Philosophy of Science, 46*, 16-23. doi:http://dx.doi.org/10.1016/j.shpsa.2013.11.007

Wombwell, E., Fangman, M. T., Yoder, A. K., & Spero, D. L. (2015). Religious Barriers to Measles Vaccination. Journal of Community Health, 40(3), 597-604. doi:10.1007/s10900-014-9956-1

Wray, K. B. (2002). The Epistemic Significance of Collaborative Research. *Philosophy of Science, 69*(1), 150-168. doi:10.1086/338946

Zwaan, R., Etz, A., Lucas, R., & Donnellan, M. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, E120. doi:10.1017/S0140525X17001972

Appendix 1: Workshop Process & Logistics

In order to investigate challenges to epistemic security we ran a series of three one-day collaborative workshops. Our aim was to enhance our understanding of how emerging technologies both undermine and buttress collective decision making processes and to begin to identify mitigation strategies.

- **Workshop 1** focused on creating a set of six threat scenarios and developing an initial view of the interacting systems within each scenario.
- **Workshop 2** focused on identifying possible interventions for each scenario with the potential to mitigate threats to collective decision-making and enhance epistemic security.
- **Workshop 3** focused on using a variety of methods to analyse a subset of intervention options for each scenario for their potential efficacy, robustness, and potential 2nd and 3rd order effects.

The workshops were co-sponsored by the Alan Turing Institute[37], the Defence Science and Technology Laboratory (Dstl)[38] and the Centre for the Study of Existential Risk[39], and facilitated by a team provided by Dstl. The attendees included representatives from government agencies, non-governmental organisations, academia, and industry.

Workshop 1: Creating scenario based system models

The key purpose of the first workshop was to set the condition for success at the second workshop by providing a handrail to support the development of a rich set of mitigation strategies. This was achieved by creating six threat scenarios and developing a view of the interacting systems associated with each threat scenario and the wider environment within which they sit. These six scenarios were not meant to be exhaustive of all the possible ways in which epistemic security could fail due to potential future socio-technical information environments and systems. Rather, they served to illustrate a range of qualitatively different failure modes that the workshop attendees considered both plausible and impactful.

Each scenario was represented as a causal map of important and interacting elements influencing the production and exchange of information. In terms of methodology this represented taking a 'systems thinking' approach: one which focuses on the way that a

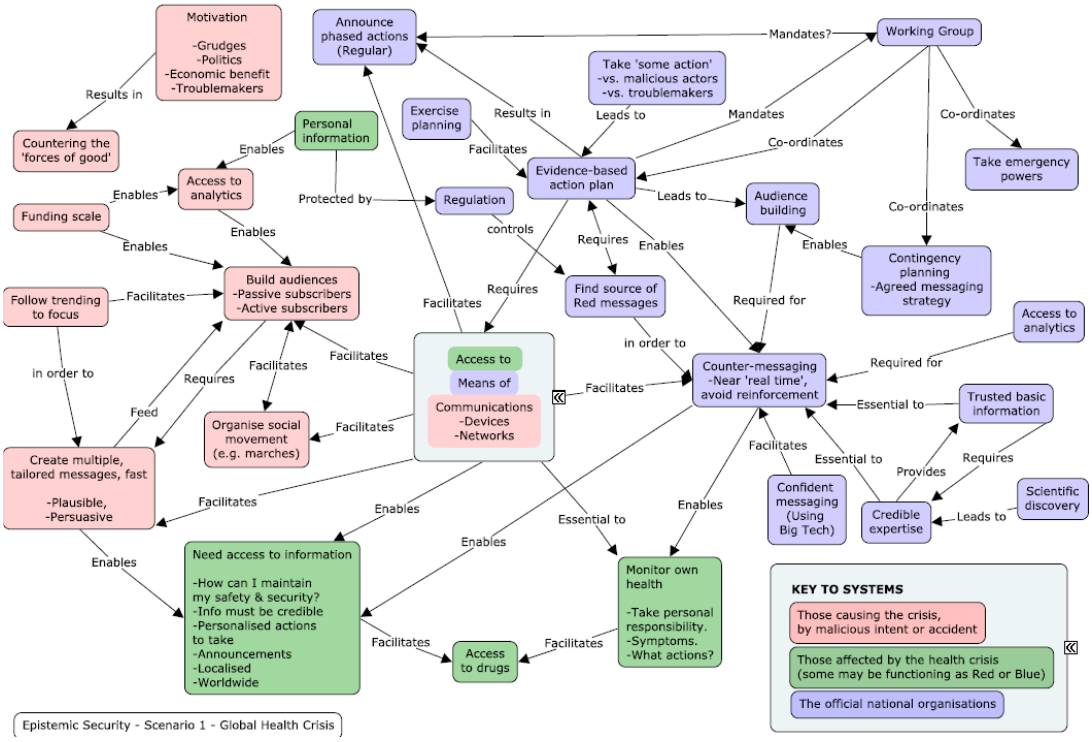
[37] https://www.turing.ac.uk
[38] https://www.gov.uk/government/organisations/defence-science-and-technology-laboratory
[39] https://www.cser.ac.uk/

system's constituent parts interrelate, as well as how systems work over time and within the context of larger systems. The approach is particularly useful for systems that have evolved in an unplanned fashion (i.e. not 'designed') such as the socio-technical systems associated with epistemic security.

A useful definition of a system from a systems thinking perspective is "an open set of complementary interacting parts, with properties, capabilities and behaviours emerging, both from the parts and their interactions, to synthesise a unified whole" (Hitchins, 2008).

Many socio-technical systems are complex, unbounded, dynamic and open. So the concept of a system is used within 'systems thinking' not simply to model situations, but to provide a method of framing and organising information about those situations; thus at the first workshop we applied the concept of a system to a situation in order to gain insight and understanding into the capabilities required by or enabling the various actors in the scenario.

We framed the discussion of the socio-technical systems associated with each crisis scenario by prompting attendees to develop a visual representation of what the Red (i.e Adversary, or Threat) system was doing (and needed in order to achieve its objectives); the Green or Target system (i.e. system whose epistemic security was under threat); and the Blue or Defending system (i.e. system attempting to mitigate the effect of disinformation). An example of the resulting scenario systems map is inserted below.



The resulting scenario systems map provides a rich picture of the interacting systems, and provides a handrail to support identification of possible interventions. In addition, the Dstl team captured a short narrative description of each scenario workshop in order to provide context to the scenario systems map.

The practicality of running a workshop with a diverse set of people, many of whom had only met each other for the first time at the workshop, meant that:

- The initial part of the first workshop was devoted to ensuring the group had a common understanding of the purpose and method.
- Attendees were divided into groups of around 10 for syndicate sessions. Each syndicate was supported by a facilitator and a scribe provided by Dstl.
- Three syndicates operated in parallel, and each attendee worked on the concept map for two of the scenarios.

Additionally we changed the membership of the syndicates between sessions to maximise the number of interactions between the attendees and to counteract 'groupthink' within syndicates.

After the workshop, the concept maps and narrative descriptions were circulated to attendees.

Workshop 2: Identifying potential interventions

The purpose of the second workshop was to build on the outputs of the first workshop by identifying intervention points within the set of interacting systems associated with a particular scenario and by identifying and grouping the interventions which might be useful at these points to mitigate the threats.

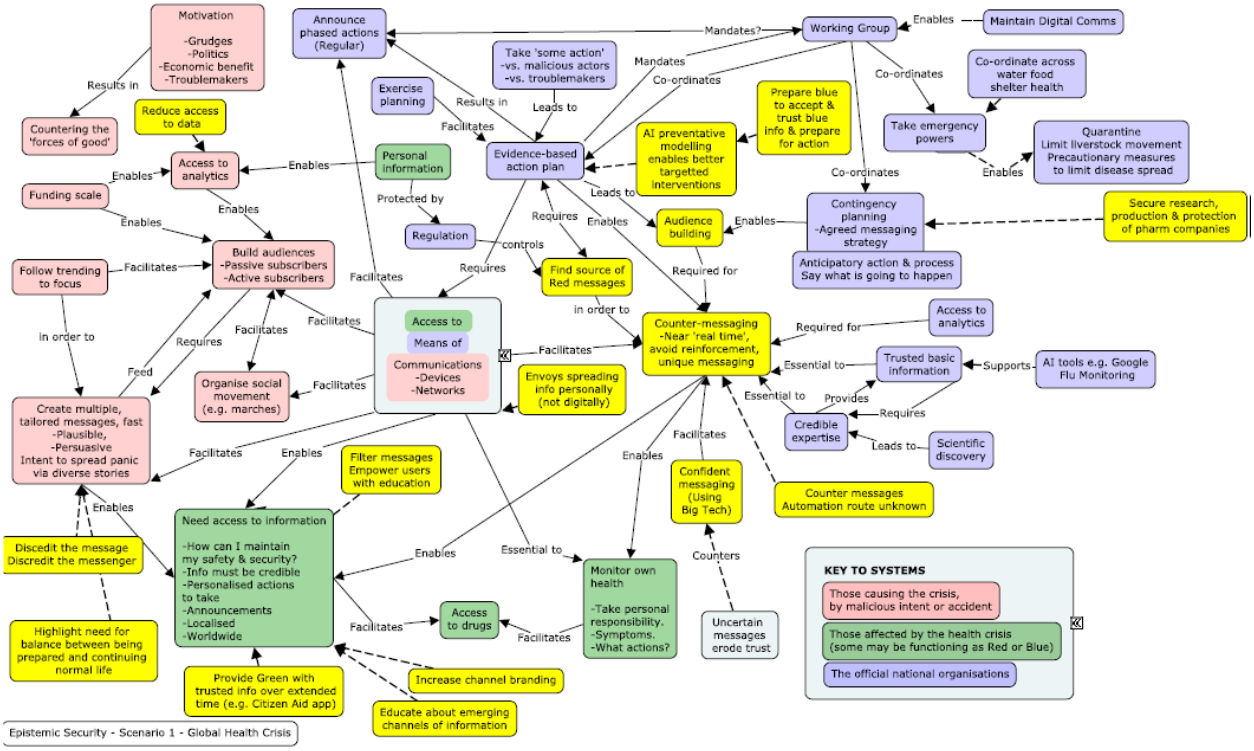
The workshop started with a reminder of the scenarios. As at the first workshop, three syndicate sessions were run in parallel with each syndicate addressing one scenario. The syndicates each had about 10 attendees and was supported by a Dstl facilitator. Again, each attendee addressed two of the scenarios, and syndicate membership was altered between sessions. Finally there was a general brief back and discussion session.

Whilst the primary purpose within the syndicate sessions was to identify and discuss potential interventions, the attendees were also encouraged to review and enhance the concept maps by considering in turn the red, blue, and green systems. The facilitators used a set of key questions to encourage and focus the discussions. For example, when considering the red system the following questions were used:

- What else do the red systems need, do, create, have, etc.?

- Are any red system elements under-represented?
- What does the red system depend on?
- Can you make things any worse?

An example of the resulting scenario systems map with suggested (but not yet assessed) interventions is presented below. The interventions are marked in yellow. When compared with the original map, a number of enhancements can be seen, for example the addition of a set of blue boxes on the top right. The main change has been the identification of a larger number of possible interventions. Again a comparison with the original scenario system map will highlight that a number of items in the original map have been re-coded as interventions in this revised version.



It is important to note that the interventions were not reviewed for their acceptability (ethical, legal or social) at this stage; clearly a formal review of the acceptability, and effectiveness, of any potential intervention would need to be undertaken implemented.

Workshop 3: ‘Red-teaming’ the interventions

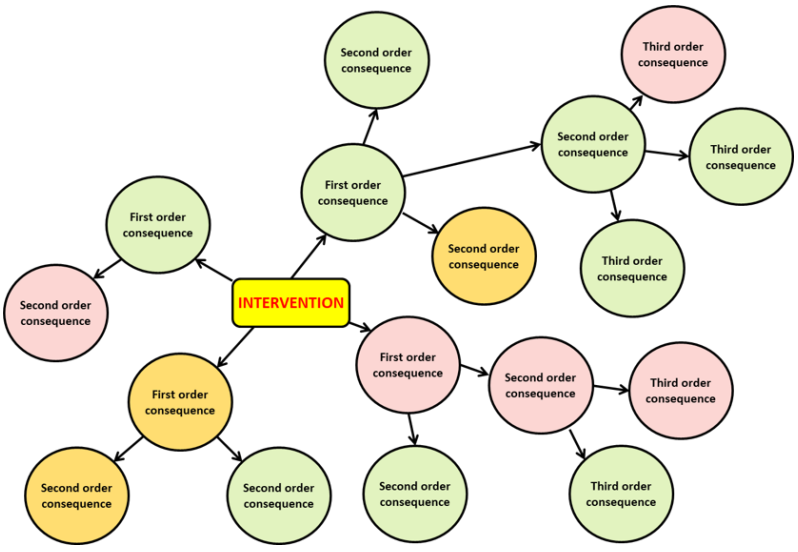
The purpose of the third workshop was to build on the outputs of the previous workshops by ‘red-teaming’ the interventions identified at the second workshop in order to examine their utility and anticipated outcomes, including the possibility of unintended consequences.

Three particular analytic techniques were used to examine the interventions:

- Pre-Mortem;
- Futures Wheel;
- SWOT (Strengths, Weaknesses, Opportunities, Threats).

The **Pre-Mortem** technique focuses attention on how plans to enact an outcome may fail, and thus overcome any optimism bias within the plan and lead to the development of a more robust plan. In outline, the participants start by describing what a fiasco (as opposed to the intended outcome) might look like, then identify what caused the plan to fail and the fiasco to occur, before developing a set of actions to address the main concerns identified.

The **Futures Wheel** technique focuses attention on potential first-, second- and third-order consequences of an action: in this case of a particular intervention. These consequences are identified as being positive, negative, neutral or uncertain.



The **SWOT** technique focuses attention again on a particular intervention, but instead looks at the intervention from the perspective of a particular actor (e.g. Blue or Red) to identify Strengths, Weaknesses, Opportunities and Threats (SWOT) associated with the intervention from the point of view of a particular actor.

STRENGTHS Which of our characteristics will help us achieve our objectives?	WEAKNESSES Which of our characteristics will work against the achievement of our objectives?
OPPORTUNITIES What events or developments might help us achieve our objectives?	THREATS What events or developments might work against the achievement of our objectives?

As in the previous workshops, the participants were divided into syndicates. The syndicates were facilitated by Dstl. Three syndicates ran in parallel, and each attendee addressed two scenarios.

The third workshop only had sufficient time to address a very small subset of the potential interventions identified in the second workshop. This illustrates that the most time-consuming part of developing an approach to mitigate risks to epistemic security is likely to be assessing potential interventions and developing a robust action plan to address the risks.

Appendix 2: Workshop Scenarios

This appendix includes the scenario narratives and system maps, including the interventions, developed during the workshops. The scenario system maps are the output from the second workshop.

The six scenarios are:

1. **Global Health Crisis**. During a major event, a pandemic in this case, a range of actors exploit the opportunity to spread misinformation and disinformation for their own reasons. This leads to dangerous practices, deaths and a significantly increased burden on health services.
2. **Character Assassination for Profit**. Identification of individuals who may be influential in the future, and the creation of long-term fake histories so there is ready-made material for blackmailing, discrediting, or otherwise manipulating these people in a way that is difficult to disprove.
3. **State Fake News**. The government of a nation state feels challenged by a set of developments and wishes to create a fake narrative to deny its responsibility or the need

to act. The narrative refers specifically to climate change, but it could apply to other issues such as historic events.

4. **Economic System Collapse**. The financial system rests upon trust. Trust is undermined by use of disinformation to manipulate trades, the use of AI-driven high frequency trading, and an increased reliance on crypto-currencies.
5. **Xenophobic Ethnic Cleansing**. A radical xenophobic group organises an attack and blames it on a minority in order to create a major societal backlash against the minority group.
6. **Epistemic Babble.** Individuals turn to information mediating technologies to help cope with an increasing deluge of claims, counter-claims, and attention-grabbing presentations. Information overload leads people to attend primarily to information that aligns with their pre-held beliefs. The ability of individuals to attend critically to complex arguments atrophies, and achieving a broad consensus for action on complicated issues becomes more difficult.

The narratives provide context to the scenario system maps. **They are not predictions of a particularly likely future**. Instead they represent important features of a potential future situation and provide a framework for the analysis of systems and interventions.

Scenario 1: Global Health Crisis

Narrative: In this scenario a health crisis sweeps across the world. Its source is unknown.

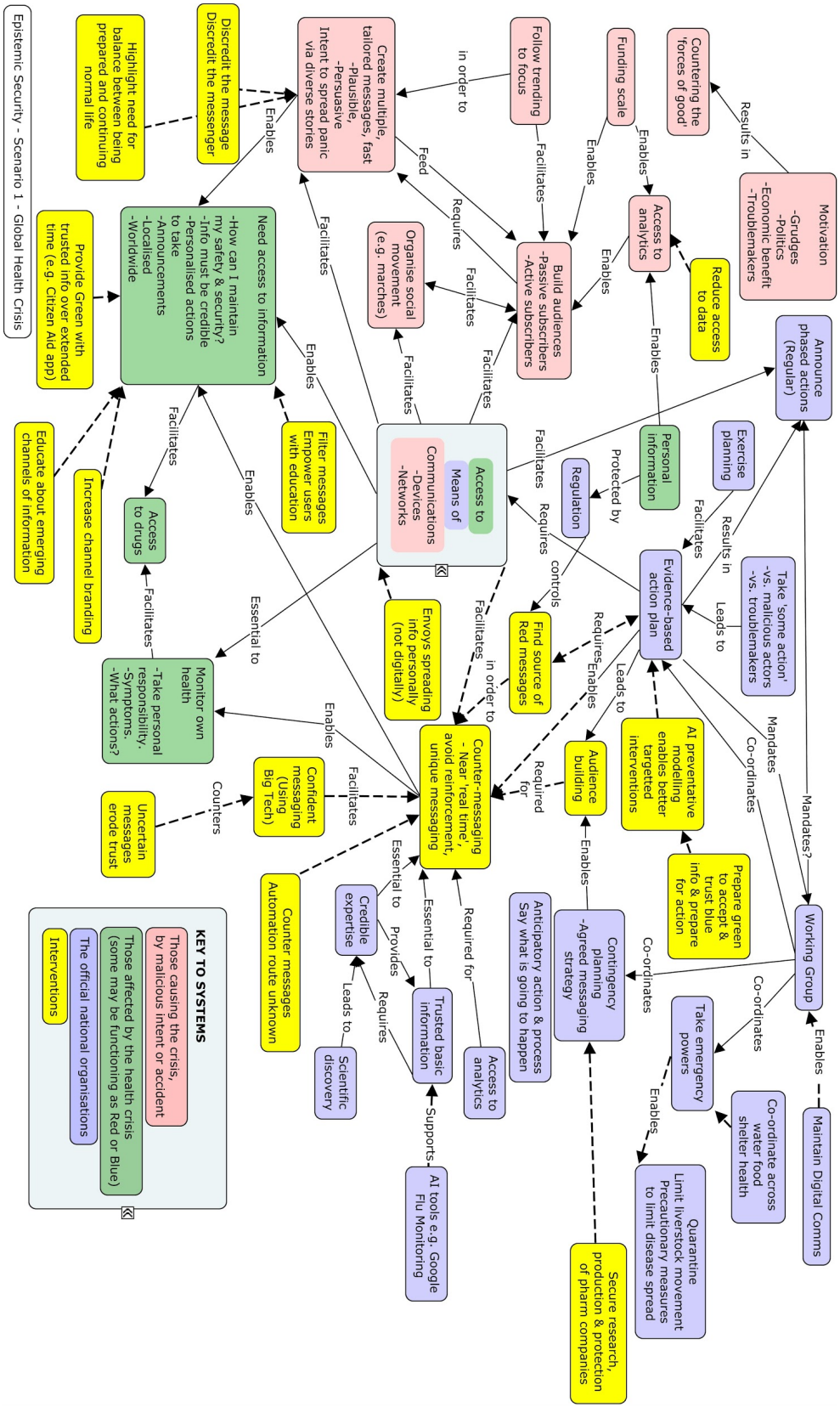
Official national organisations seek to reassure their populations, provide informed advice and guidance to keep their populations safe, and take some action to address the crisis.

The populations want to know how to stay safe, but don't know what to do. They seek guidance tailored to their situation, but are not quite sure what to believe. The slow appearance of official information drives them to seek any plausible advice. The advice is repeatedly disseminated between friends. This results in many people following dangerous practices, which increases the burden on health services. There is a rise in suicides throughout populations. There is also a rise in alternative institutions. Populations want access to drugs to protect themselves, and are willing to pay.

There are also groups operating with malicious intent, seeking to prolong and extend the crisis for their own objectives. These groups may have started the crisis, but nobody knows. Their motivations could vary from exploiting the situation for their financial benefit, to furthering grudges or political aims, to simply wanting to cause trouble.

Everyone may be affected by the health crisis: populations at large, those who may be functioning as officials, and those with malicious intent.
The result of the health crisis is the destabilisation of populations who don't know what to do, who have little faith in the ability of their official organisations to improve the situation, and who just see things getting worse. Those who think they are causing the crisis exploit the situation and enjoy the spectacle.

Global Health Crisis scenario systems map



Global Health Crisis interventions and red-teaming

The systems map, interventions and red-teaming all highlighted the importance of preparation and trust in countering malicious actors and messages during a crisis. This included actions such as:

- Development and testing of plans to provide the public with access to information taking account of the presence of malicious actors and the complexity of social epistemic systems.
- Creation and maintenance of trusted information channels before the crisis (noting that these may be both digital and non-digital). The Pre-Mortem and SWOT both highlighted the risks associated with compromise of the channels and the unintentional provision of incorrect information. The SWOT also highlighted a Blue weakness stemming from the potential complexity and uncertainty of an emergent health crisis, with corresponding opportunities for Red.

Scenario 2: Character Assassination for Profit

Narrative. In this scenario, rather than trying to implant 'fake news' into current world events, a long term view is taken in which a group of 'potential future leaders' is identified and a whole 'fake history' is created for each individual as they progress their careers.

The plan is that once these individuals become influential, some of the fake embedded facts about their early lives can be used to manipulate their actions either overtly or subliminally. As these facts will be verifiably contemporary with the events they purport to record, proving that they are false will be virtually impossible, making the damage they do much more effective.

Clearly many of the false histories created will never be of use, but the few that can be used will be sufficiently effective to justify the effort.

This influence could be achieved in a number of ways:

- Traditional blackmail either for money, as a way of financing terrorist acts, or to make them behave in a way that they would not otherwise have done (voting for example);
- Simply distracting their attention at a crisis point either political or economic so that chaos ensues;
- Releasing, what would by then be, authenticatable historic information in order to undermine their credibility and force a resignation or other desired outcome.

There are two possible motivations for this, both of which would need initial access to finance and a long term commitment:

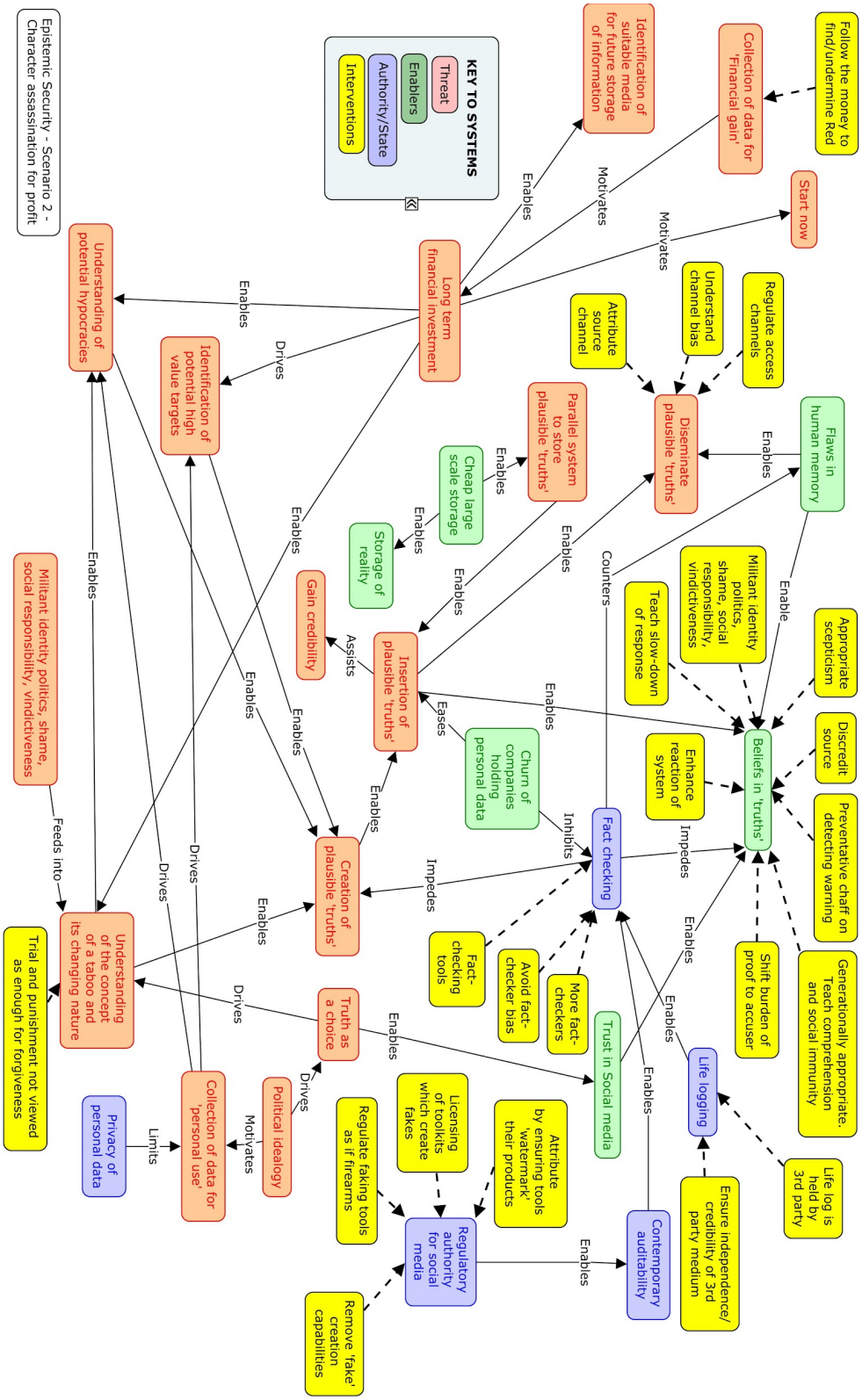
1. A politically motivated group which sees this as part of a long term strategy to undermine its enemies or to gain financing for terrorist activities.
2. A long term investment for profit in which someone with no particular political agenda simply amasses information to sell at the appropriate time and to anyone who will pay, as they would be able to sell their 'information' to a variety of different interest groups. This could increase the success rate for the 'fake histories' and therefore the potential for profit.

In either case this is enabled by the availability of cheap data storage and the technology to allow photographs and documents to be believably tampered with.

The ability and foresight to create fake documents at the appropriate time (e.g. a picture of someone dealing drugs created and stored when they were 20 and held for use in 20 or 30 years' time) adds to the authenticity of the spurious information as does the creation of a holistic story involving multiple event and associates.

This opens up a variety of opportunities which could enable the subversion of the political process or create enough confusion to reduce the ability to deal effectively with a set of simultaneous crises.

Character Assassination for Profit scenario systems map



Character Assassination for Profit interventions and red-teaming

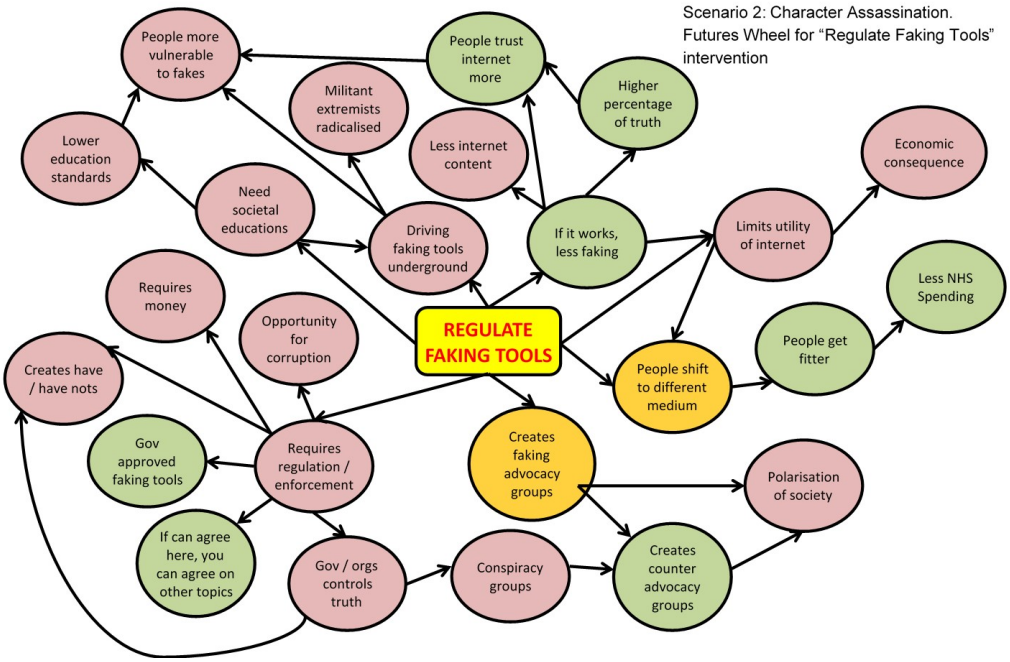
The Pre-Mortem highlighted the potential risk that we end up in a world in which trust in digital experience and testimony is effectively zero, and thus it becomes impossible to hold individuals to account and society is potentially overwhelmed by scepticism and cynicism. It is worth noting that this future world is similar to the Epistemic Babble scenario and is very similar to the "Death of Reality" future scenario in *The Emerging Risk of Virtual Societal Warfare* by M Mazarr et al (2019).

There are clusters of interventions relating to:

1. Changing societal culture associated with historic taboos, a rush-to-judgement, levels of proof and awareness of potential threat
2. Fact-checking
3. Life logs
4. Attribution of deep fakes and fake histories, including:
    a. Encouraging/regulating for digital rights management approaches which ensure tools include 'watermarks' (e.g. using steganography);
    b. Developing tools to detect 'fakes' (so raising the cost of creating fakes).

The SWOT analysis of the intervention "regulate access channels" (near top left of the map) identified that, from the Blue perspective, this would be difficult to achieve and, from the Red perspective, relatively easy to subvert (unless there was significant international agreement and cooperation).

The Futures Wheel analysis of the intervention "regulate faking tools as if firearms" created a rich picture of 2nd and 3rd order effects and potential interactions. It is included below as an example of the richness of the analysis that use of such a method can promote; but it should be noted that time precluded validation of the wheel.



Scenario 3: State Fake News

Narrative: The government of a nation state feels challenged by a set of developments and wishes to create a fake narrative to deny its responsibility or the need to act. The narrative below refers specifically to climate change; but it could apply to other issues such as historic events. (It is noticeable that the scenario systems map which was created is not specific to the example of climate change.)

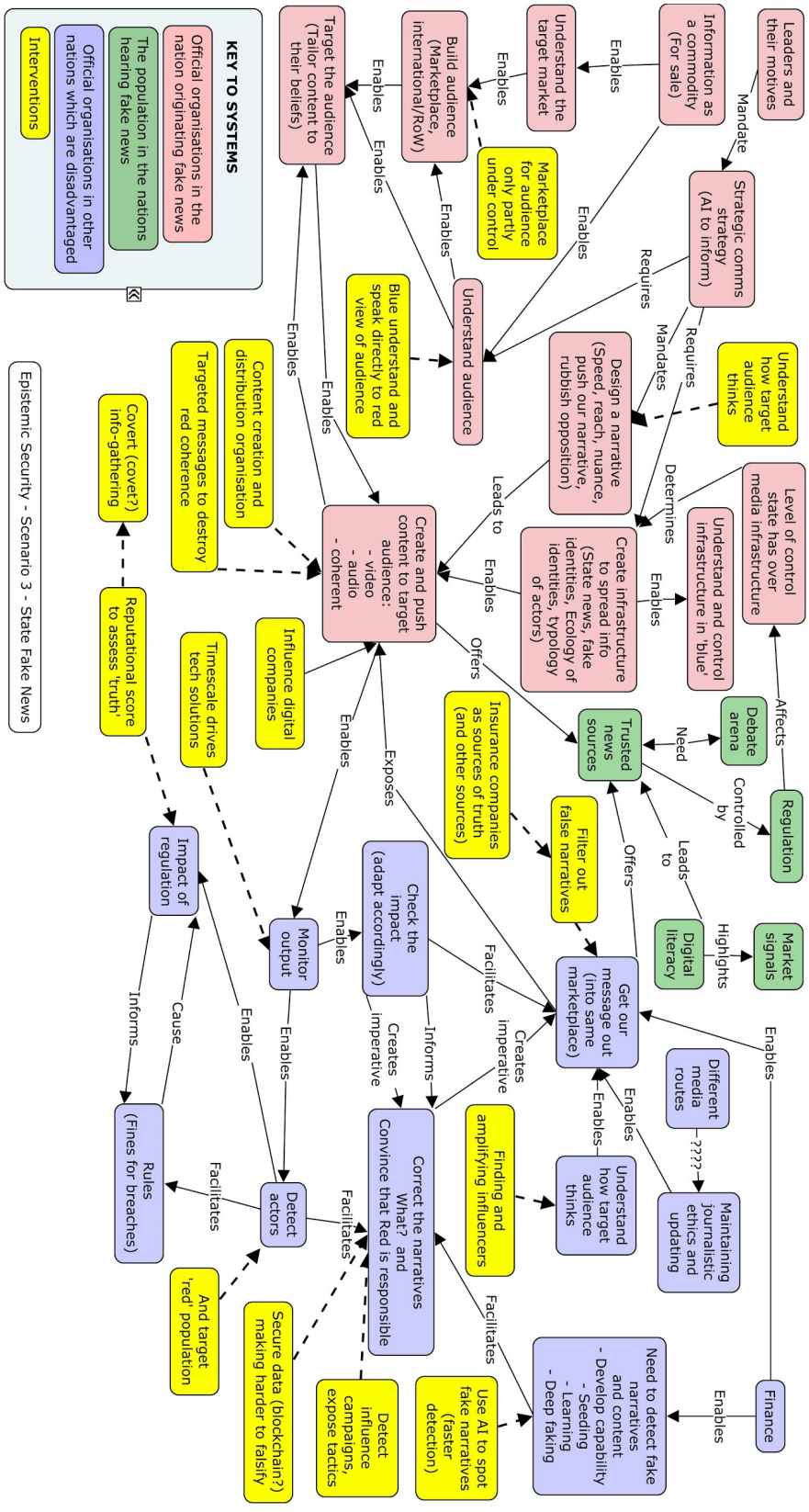
In this scenario, the leader of the government in State X is determined to maintain the high global reputation of their state with respect to its culpability for global warming. The State needs to avoid reparation costs. The State also needs to maintain global markets.

The official organisations of State X design an influence narrative. They assess the skills and technologies needed to 're-write history'. One of their first actions is to create a fake Intergovernmental Panel on Climate Change (IPCC) video showing officials denying the occurrence of global warming. They identify the appropriate channels for communicating fake news, disseminate a continuing stream and suppress any opposing views.

The populations in other nations hearing the news will access that news via many sources. They trust the news from their favourite sources, and accept the narrative initiated by State X. However, populations still see that sea levels are rising and choose to move away from littoral zones. This results in the mass migration of populations inland and to other states.

Official organisations in other nations recognise they are being disadvantaged by the emerging narrative. The nations experiencing mass immigration struggle to adapt at the pace needed, and experience a degree of unrest. Official organisations seek to counter the false narratives by detecting them and disseminating corrected information through the same news channels. Through regulation, these official organisations seek to highlight and limit the activities of fake news perpetrators.

State Fake News scenario systems map



State Fake News interventions and red-teaming

The systems map highlights the central role played, in this scenario, by (1) trusted sources of testimonial knowledge, (2) the ability to understand, build and micro-target audiences, (3) the ability to understand the objectives, methods and effect of Red narratives, and (4) access to channels used by audiences (& audiences' access to channels).

The red-teaming highlighted the following issues and challenges:

- There is the potential for Blue to win the 'battle' but lose the 'war' if not clear on long term objectives (aka moral reflection on Blue position with respect to 'truth');
- Accountability and reach of broadcast channels, and role of dialogue channels versus broadcast channels;
- Blue and audience access to channels.

Scenario 4: Economic System Collapse

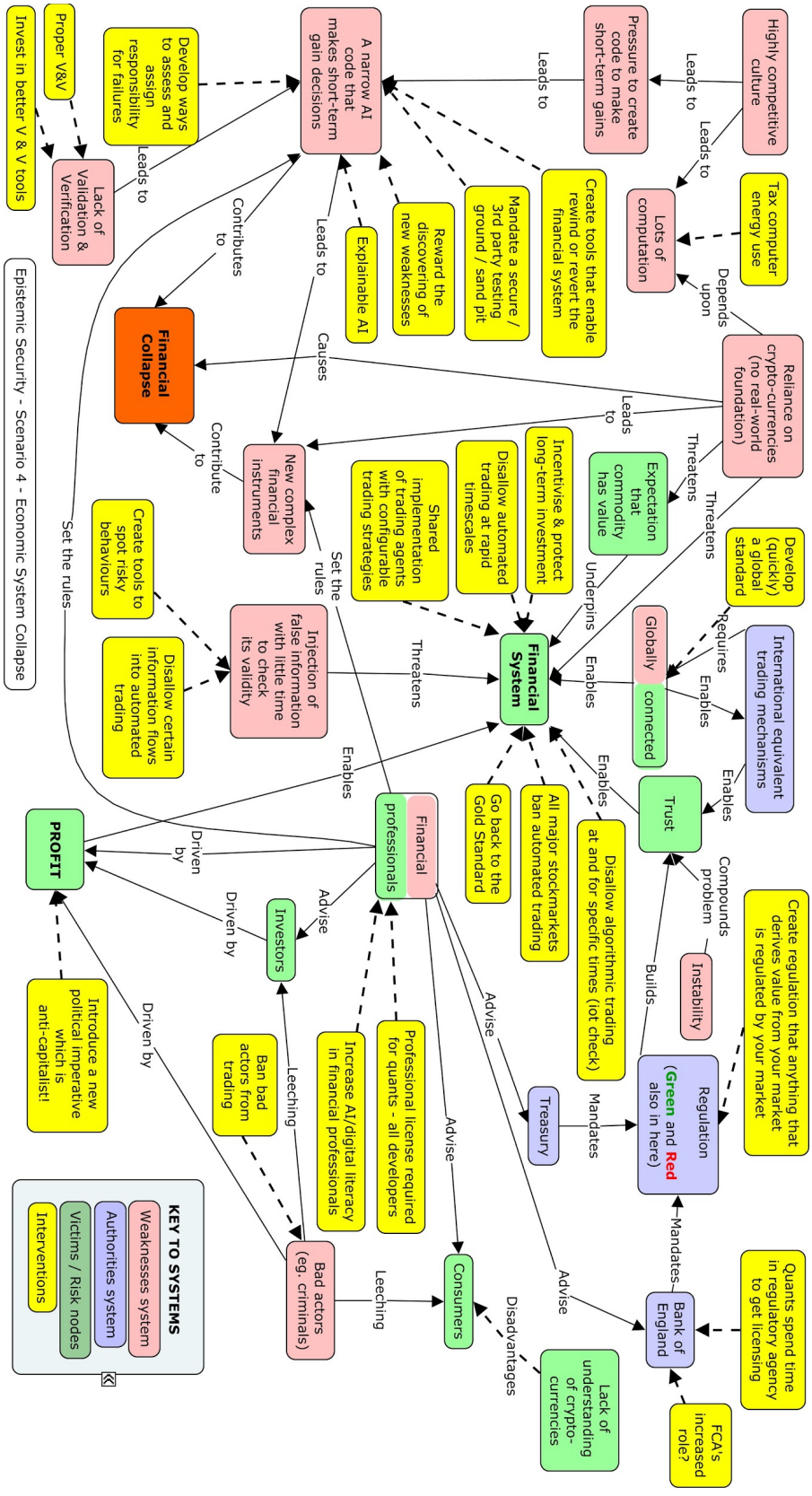
Narrative: In this scenario, the creation of narrow AI code by the financial industry for short-term gain, coupled with a reliance on crypto-currencies, leads to economic collapse.

The highly competitive culture within the financial industry and its focus on short-term gain leads the industry to create narrow AI code, lacking proper validation and verification, which makes short-term gain decisions. Financial professionals, driven by profit, encourage the short-term gain culture. They also begin to advise investors to rely on crypto-currencies. Criminals and other bad actors devise new ways to leech profit for themselves from this development.

The expectation that commodity has value underpins the international financial system, so dependence on crypto-currencies, which are not linked to any tangible asset nor supported by any national government, threatens that system.

International regulation by the financial authorities, designed to keep a financial system based on commodities with value stable, fails to adapt sufficiently fast to the new models of financial dealing driven by the use of crypto-currency. Before the authorities can enact international controls on these new models, investors suddenly lose faith in a specific crypto-currency when its technical underpinning is compromised. They attempt to realise the capital it represented in large numbers. The crypto-currency cannot deliver and this leads to the disintegration of other financial models linked to it. As investors have traded real commodities using crypto-currency, this undermines the currency of countries which invested real money in the production of those commodities. Traditional financial models therefore also begin to break down, leading to a failure of most parts of the financial system and economic collapse.

Economic System Collapse scenario system maps



Economic System Collapse interventions and red-teaming

The potential interventions identified were very wide ranging - though the Pre-Mortem suggested at least one was probably effectively impossible (returning to the Gold Standard). They included:

- Changing the objectives and behaviours of the financial system;
- Enhanced training, inc. ethics, for quants and developers;
- Enhanced testing of algorithmic tools used within the financial system
- Methods to mitigate the impact of algorithmic based trading (including both pre-event, during an event and post-event);
- Methods and tools to understand the behaviour of the system-of-systems (including in response to invalid information).

The Pre-Mortem highlighted the problem that at least some actors within the system will push the boundaries and/or cheat due to a perception that this will allow them to 'win'.

The SWOT focused on the testing of algorithms in a sand-pit environment. It highlighted challenges associated with implementing a testing regime that is both effective and efficient, and which can not be 'gamed'. However, it also identified that it offered potential opportunities for development and testing of new theories and products, and for leadership in development of norms.

Scenario 5: Xenophobic Ethnic Cleansing

Narrative: In this scenario a far right xenophobic faction forces the departure of a specific ethnic community.

The faction's controlling group decides on a strategy of implicating the ethnic community in a chemical or biological attack, which will turn public opinion against that community so much that extreme violence against them will be considered justifiable by elements of the population. Thus, these elements are the 'target population' of the far right faction.

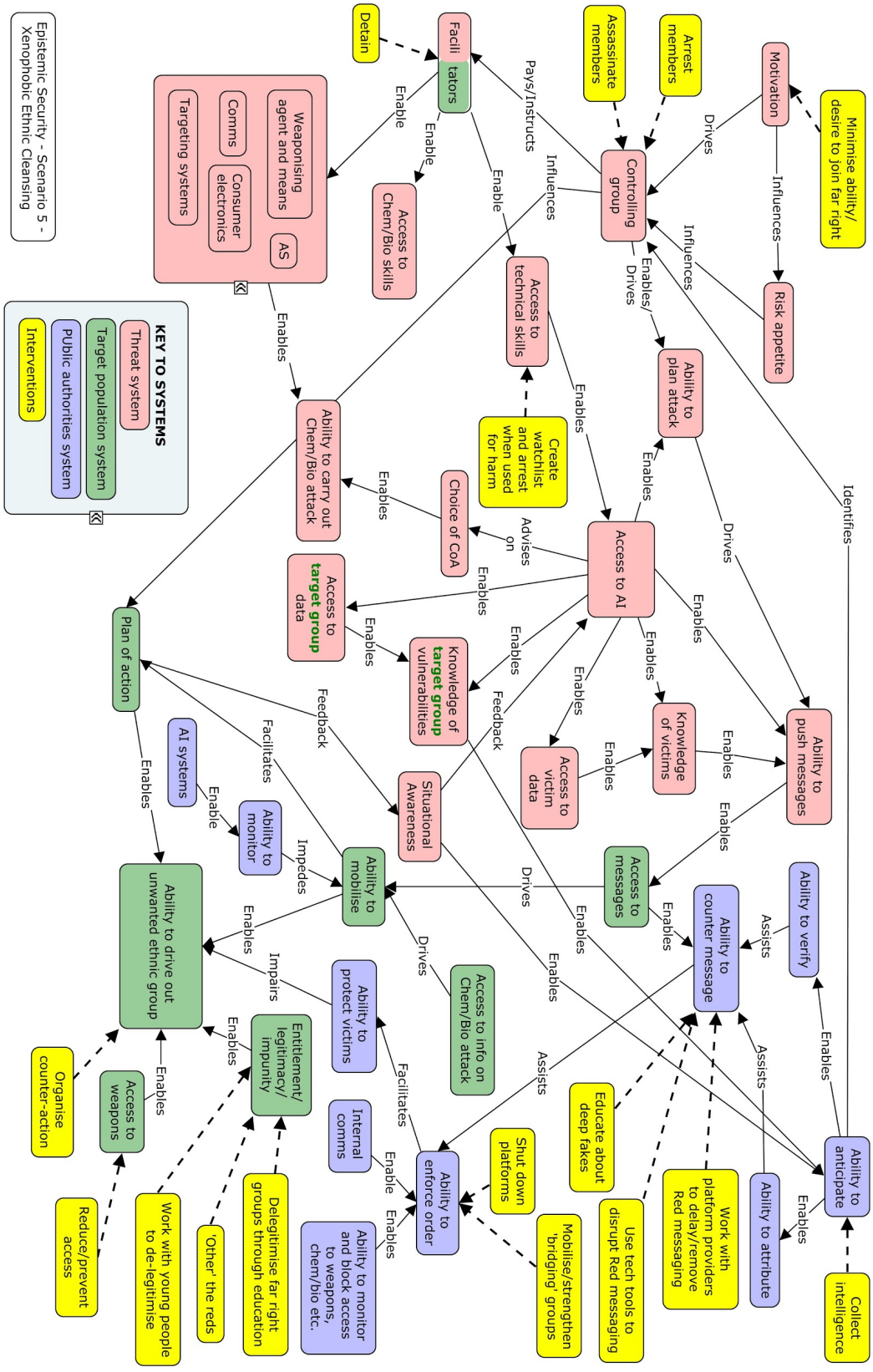
The controlling group uses facilitators to access chemical/biological weapon skills, the actual materials required for an attack, and technical skills in AI. Before they stage the chemical/biological attack, the group uses AI to research the vulnerabilities of their target group, gain as much knowledge of its intended victim ethnic community as possible, and researches and then selects the most effective and damaging course of action (CoA) possible for the chemical/biological attack.

The far right faction perpetrates the chemical/biological attack. Its controlling group then uses AI to push messages at speed and scale to the target population (through the channels that population uses to receive news and communicate) claiming that the specific ethnic community is responsible for this attack.

Elements of the target population then mobilise, arm themselves and use violence to drive out the ethnic community. The far right faction has achieved its objective.

To combat this situation, the blue team authorities have the ability to detect, verify, attribute, and stop the activities perpetrated by the far right faction. They must be able to create and push a credible counter-message at speed and scale to the target population and ultimately, they must be able to halt the violence against the ethnic community.

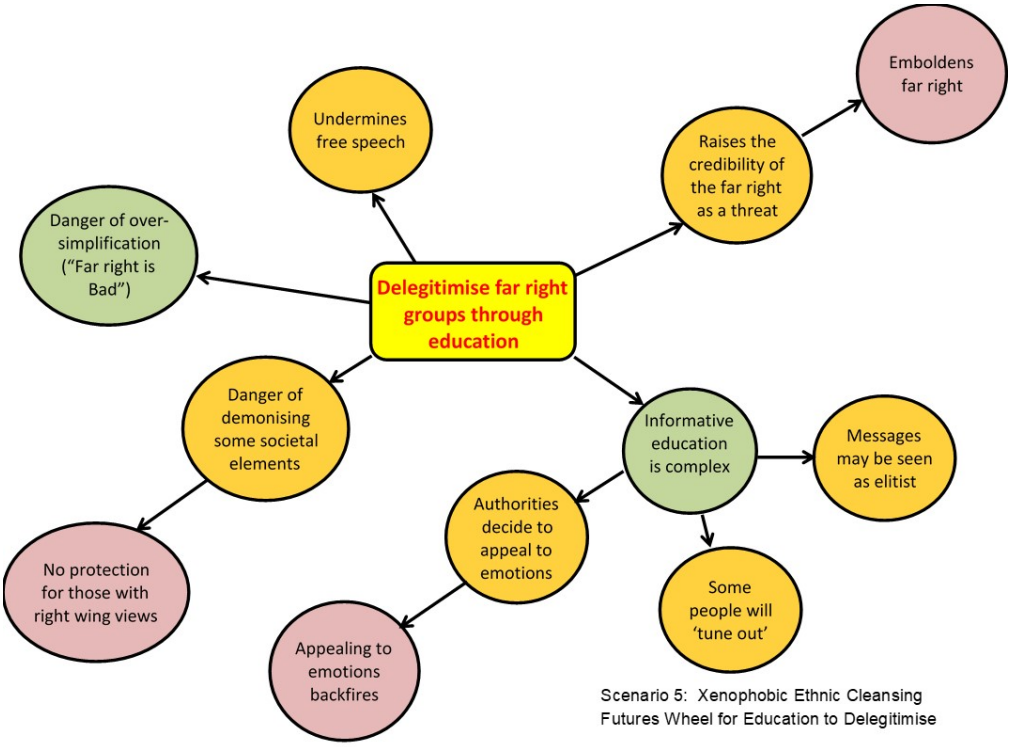
Xenophobic Ethnic Cleansing scenario systems map



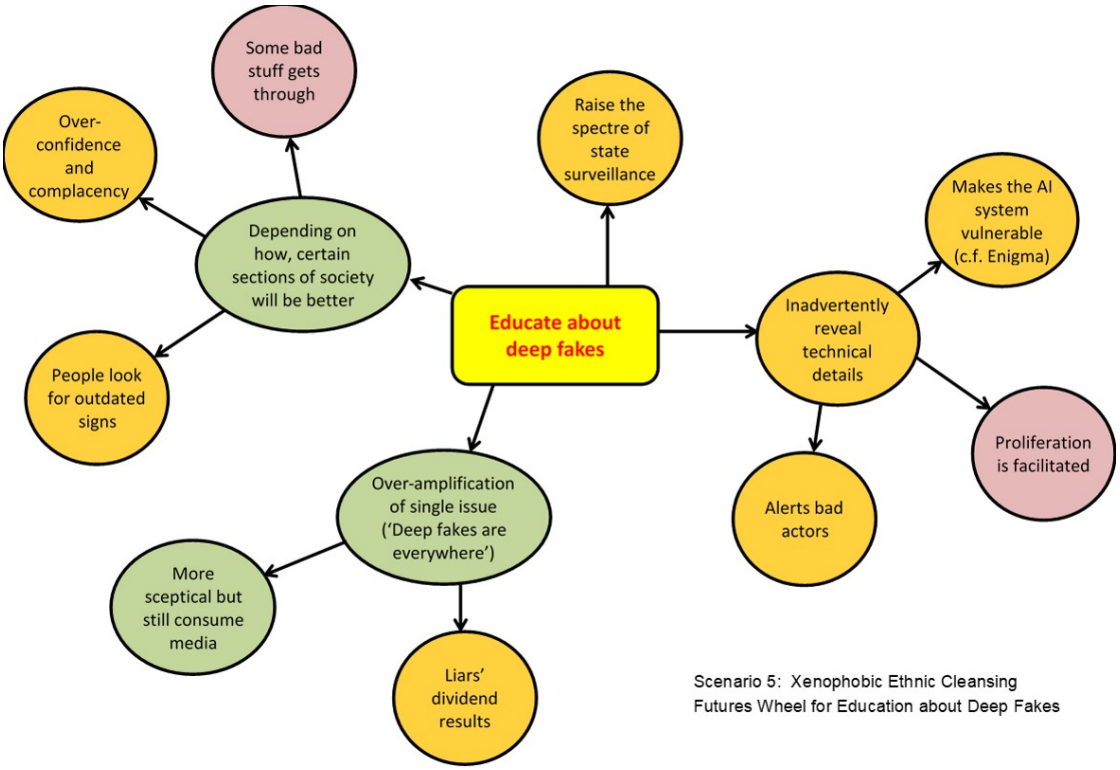
Xenophobic Ethnic Cleansing interventions and red-teaming

Many of the interventions identified are traditional counter terrorism and de-radicalisation techniques. The interventions which are less traditional, as they relate to this specific scenario, are those related to (a) countering deep fakes and (b) adding delay into the communication channels (in order to slow mobilisation and "mob" behaviour).

The Futures Wheels provided a rapid way of beginning to analyse the complexities and potential down-sides associated with interventions. The Futures Wheels for the two interventions which use education to enhance target resilience to (a) far right messages and (b) deep fakes are included below.



Scenario 5: Xenophobic Ethnic Cleansing
Futures Wheel for Education to Delegitimise



Scenario 5: Xenophobic Ethnic Cleansing
Futures Wheel for Education about Deep Fakes

Scenario 6: Epistemic Babble

Narrative: In this scenario the ability for the general population to tell the difference between truth and fiction (presented as truth) is lost.

Social media has allowed people to put forward spurious views and for them to be accepted along with the views of true experts as being of equal value. This in turn has led to expertise being devalued and people no longer respecting or accepting the views of educated and knowledgeable people, or accepting authority in any way.

Although information is easily available, people routinely purport to be other than themselves on electronic media and this goes undetected, so people cannot tell whether the information they are receiving is reliable or not.

People connect with others but have little idea of who that person really is or where they come from. 'Social' interaction is less 'social' and more computer-based. These developments have resulted in societies both mixing more and fracturing, with contradictory views being strongly held by different factions within a society.

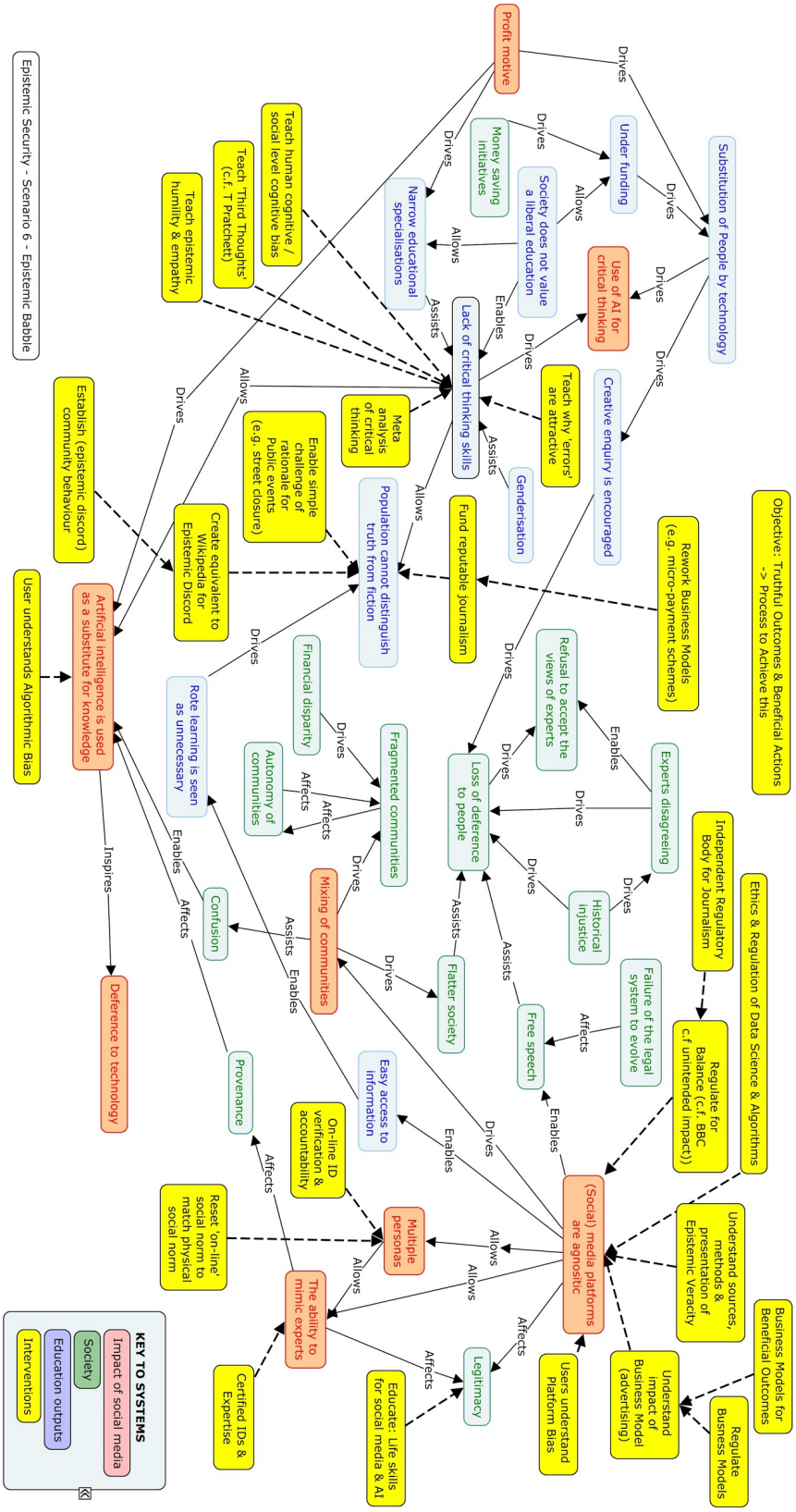
Additionally the education system relies on digital technologies to radically reduce the number of real teachers (without adequate testing of the change). This results in pupils not developing

the ability to apply critical thinking to the information that is presented to them, as so much is, and has been presented, without the guidance of an adult.

The result of this ‘Epistemic Babble’ is that there is an environment of ‘knowledge’ and belief that could be easily manipulated, and hence the views and actions of society are ‘up for grabs’. The question is who will ‘grab’ them?

Note: This scenario lacks the malicious actors present in the other scenarios. There are strong resonances between this scenario and the “Death of Reality” future scenario in *The Emerging Risk of Virtual Societal Warfare* by Mazarr et. al. (2019)

Epistemic Babble scenario system maps



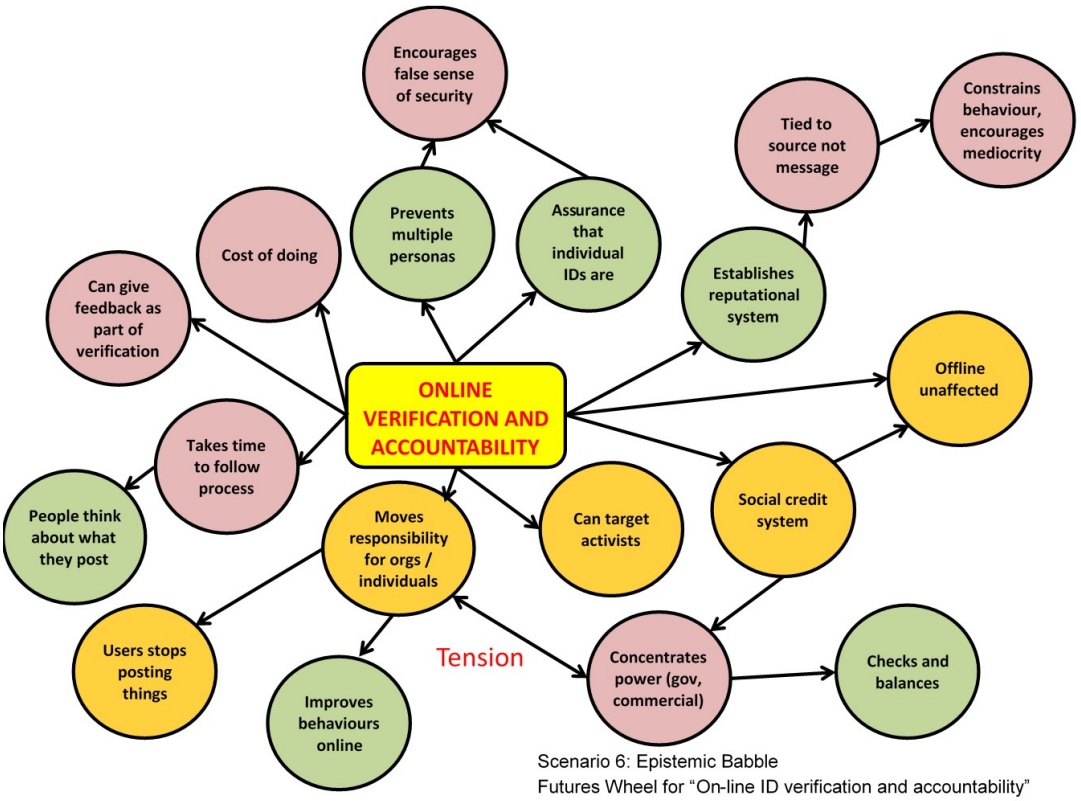
Epistemic Babble interventions and red-teaming

The systems map has several major intervention clusters, including interventions aimed at:

- Enhancing the capability and resilience of the population
- Enhancing participation and development of best practice in areas of epistemic discord
- Enhanced methods of verification of epistemic veracity
- Changing the media ecosystem
- Changing the behaviour of developers of future data systems (inc. algorithms)
- Changing on-line behaviour

In the State Fake News scenario the analysis of interventions identified the potential for Blue to win the 'battle' but lose the 'war' if not clear on long term objectives (aka moral reflection on Blue position with respect to 'truth'). In this scenario an intervention was identified which was focused on clarifying the long term objective, and thus the pre-mortem identified a set of potential issues associated with that objective.

The Futures Wheel again proved to be useful in providing a rapid identification of potential issues related to an intervention. In this case the intervention was to create a method of on-line identification verification and accountability.



Scenario 6: Epistemic Babble
Futures Wheel for "On-line ID verification and accountability"

Appendix 3: Technological threats and fixes

Here we collect a non-comprehensive list of known or expected threats from emerging technologies to epistemic security and of some of the proposed fixes to these threats. We are not necessarily advocating for the adoption of any of these proposals; as highlighted in the main text, we strongly urge any policy to take a systemic, holistic approach and pay careful attention to ways specific policies or solutions may backfire, using for example a "red team" methodology.

At the end of appendix 3, table 3.1 maps the epistemic threats described in this appendix to the different basic knowledge sources (experience, memory, reason, and testimony). Table 3.2 similarly maps the proposed “fixes” presented in this appendix to the epistemic vulnerabilities described in the main text (attention, community, trust, and adversaries)

Viral spread of misinformation in social media

False, fake, or otherwise misleading information is easily created and easily propagated on social media (Lazer et al. 2018; Vosoughi, Roy & Aral 2018). Some proposed "fixes" include:

- **Novel regulation:** across the world, governments are responding to misinformation through legislation, regulations and task forces (Funke & Famini 2018). For example, Singapore has introduced a fine for the spreading of misinformation, and also established by law a right for government to publish corrections alongside misinformation on various platforms (e.g. Facebook) and to take action to stop the spread of misinformation on messaging platforms.
- **Platform intelligence / automated tripwires:** communication and media platforms such as Facebook and Twitter explore a combination of manual and automated processes to detect rapidly spreading content on their platform, which can then be flagged for fact-checking or other screening and handling.
- **Platform automated warnings / pre-scripted messaging:** for certain topics (e.g. vaccines) platforms like Facebook and Youtube show warnings for content that is suspected to be misinformation, and show pre-scripted messages linking to authoritative sources on the matter.
- **Information literacy / misinformation education:** governments, academics, NGOs and corporations are investing resources in educating platform users and the general public on how to detect misinformation and in general information literacy. For example, Finland has rolled out such education, including a specific focus on state-run propaganda, in its school system (Charlton, 2019); a Google-supported partnership of

NGOs and academics have created an online course on navigating digital information[40]; and Cambridge academics have created an educational game showing how disinformation is generated and spread (Roozenbeek & van der Linden, 2019).

- **Platform-wide reputation / "karma" / editorial system:** platforms for user-generated content or contributions, like Wikipedia, Stack Overflow, or Hacker News operate a platform-wide system to control content additions, modifications, deletions and prevalence. For example, Wikipedia operates a distributed editorial system, whereas Stack Overflow and Hacker News operate "karma" systems where positive feedback from other platform users leads to greater editorial control.
- **Per-community tools for reputation / "karma" / editorial management:** some platforms, like Facebook or Reddit, delegate editorial control to administrators, or "admins", per community (e.g. Facebook/Whatsapp group, subreddit) on the platform, with no user having editorial control across the platform (except employees of the company working under the company's guidelines and local laws and regulations).
- **Law enforcement work with platforms to enforce misinformation-related laws:** government law enforcement works with platforms under the remit of available laws (either novel misinformation-related laws as in Singapore, or existing laws that cover topics like libel, defamation, hate speech or intellectual property rights) to address misinformation and prosecute offenders on the platforms.
- **Financial support for real-time fact checking:** governments, corporations, philanthropists and crowds provide financial support to real-time fact checking organisations, such as Politifact and Snopes, that can help mitigate the effect of misinformation.
- **Long-term support for trustworthy information sources:** In addition to fact-checking sources, which respond to content appearing on other channels, governments and other organisations also work to establish and maintain the trustworthiness of information sources such as the British Broadcast Corporation. This is done by distancing them from the need to fight for advertising revenue and by subjecting them to regulation.

Machine generated fake evidence / malicious synthetic media

Digital tools, like Photoshop, and especially AI-based tools (like Generative Adversarial Networks) can be used to generate fake media (e.g. images, audio) that appears real. If used at scale by malicious actors, this technical capability could significantly increase the cost of finding and establishing the truth about a particular matter, up to the point of potentially losing trust in digital evidential sources. Some proposed "fixes" to this threat include:

[40] https://www.youtube.com/playlist?list=PL8dPuuaLjXtN07XYqqWSKpPrtNDiCHTzU

- **Automatic detection:** depending on the tool used to generate the synthetic / manipulated media, it may be possible to detect artefacts and identify the fake purely using data contained in the e.g. image/sounds sample. This is more easily achieved when the detecting party has access to the tool or machine learning model that was used to generate the fake. When it is possible to detect fakes in an automated manner in real time, platforms can use such detectors to prevent the upload or spread of such content.
- **Retrospective forensics:** if real-time detection is not possible or available, retrospective forensics can be used to assess the media as well as the relevant context, to judge whether a piece of evidence has been faked. This could be especially relevant in legal contexts, or in other contexts where individual pieces of evidence can carry significant weight.
- **Prospective cryptographic provenance assurance:** extremely-difficult-to-forge digital signatures exist and are used to verify signatures on documents and to authenticate identities online. There are some proposals to add similar digital signatures to media (e.g. at the camera or microphone) and then check for them downstream. However, to deploy this across all devices (both capture and display/playback) would be a major undertaking, likely requiring international standards or other major coordination efforts, and could also make everyday use of image and video editing problematic.
- **Proliferate "radioactive" data:** research has shown that image datasets can be manipulated to make them "radioactive", such that outputs from synthetic generators trained on these datasets can be more easily detected. If a majority of public datasets currently used to train synthetic generators can be treated in this way, it will lower the cost of detection, or increase the cost for adversaries who wish to hide the fact that such outputs are synthetic, without harming legitimate uses of synthetic media generators (Hwang, 2020).
- **Limit access to technology:** given the potential risks from maliciously generated synthetic media, some have called for developers of the enabling technologies to conduct pre-publication risk assessment, and in some cases choose to abandon certain developments or otherwise adopt selective sharing of their findings, e.g. first with developers of detection tools and the platforms who can deploy them.

Automated testimonial sources (bots, language models)

Much like fake images and audio, progress in natural language understanding and generation and in statistical language modelling has enabled the creation of automated text and conversation generators (generative language models and chat bots), which could be maliciously used to engage a large number of online users (Chessen 2017). Some proposed "fixes" include:

- **Automatic detection:** as in the case of machine generated synthetic media, artefacts of the generation process may be detected and used to flag automatically generated text, especially if the developers of the detector have access to the code or model used to generate the text.
- **Detection through behaviour and social patterns:** similar to detection of fake images and audio, behaviour and social patterns could help to provide a contextual assessment of textual content being machine-generated.
- **Identity verification:** some have suggested requiring identity verification for users on a platform, to make sure they are not bots (e.g. similar to the identity required now to engage with government or financial institutions and online services). This may be particularly critical in response to concerns of orchestrated and massive-scale disinformation campaigns that could be run through bots.
- **Limit access to technology:** this involves considerations and approaches similar to the case of machine generated synthetic media (above).

Online targeting and customisation

With many aspects of life now taking place online via interaction with digital platforms, it becomes increasingly possible to for these platforms, or certain users of these platforms (such as content contributors or advertisers), to match certain content items to certain users, or to otherwise modify the user's experience of the platform based on information provided by the user and/or based on the behaviour of the user on the platform. For example, a user's engagement with certain videos on a platform like YouTube, or with certain pages and posts on a platform like Facebook, influences the content that appears in the user's recommendations or news feed. It also provides criteria for advertisers to select which users see which ads. Concerns have been raised that this technology undermines users' privacy, and allows malicious actors to target vulnerable individuals and communities (Centre for Data Ethics and Innovation 2020). Some proposed "fixes" include:

- **Advertising restrictions and transparency:** the advertising targeting systems integrated into e.g. Facebook and Google platforms provide one of the easiest methods for online targeting, both for legitimate users and malicious ones. Regulations around advertising (e.g. age restrictions for addictive or harmful substances, spending caps and transparency of political ads), and their enforcement on these platforms, is one suggested method for limiting the harms.
- **Privacy and accountability regulations:** legal frameworks like the European Union's General Data Protection Regulation (GDPR) require platforms to create governance structures that empower users and limit their exposure to unwanted targeting.
- **Alternative business models:** some have suggested that the current advertising-based business model of online platforms necessarily requires greater surveillance of users

and exposure to targeting harms (under the moniker of "Surveillance Capitalism"), and suggest a shift to other business models (public support, crowd support, subscription, or other) would be required to avert the harms.

Amplification of falsehoods by recommendation algorithms

In addition to spread of misinformation through social networks, and through targeting, it has also been observed that media platforms with user-generated content and algorithmic recommendations, such as YouTube, can significantly amplify the attention received by misinforming content available on their platforms due to the way users interact with such content (Chaslot 2017). Some proposed "fixes" include:

- **Manual tweaks:** for certain topics or content types, platforms can create a stock of trusted channels or sources, or promote and demote content relative to its authoritative reference.
- **Specific topic warnings:** for certain controversial or potentially harmful topics and keywords, platforms can display warning messages that link to authoritative sources. For example, Twitter chose this approach in response to misleading information relating to COVID-19.[41]
- **Algorithmic changes:** some platforms have undertaken deeper investigations into the causes of algorithmic amplification of misinformation, and have experimented with changes to the algorithm to reduce the occurrence of such amplification.

Algorithmically-reinforced echo chambers and filter bubbles

People have a tendency to seek information that fits their pre-held beliefs ("confirmation bias"), and users of media platforms are no different. However, in addition to the user-driven dynamic of "echo chamber" formation discussed in the main text, there is a further concern for systems with algorithmic recommendation of content. Here, the system may learn the bias of the user and thus create a reinforcing feedback loop whereby users become increasingly exposed to information that coheres with their biases (Jiang et al. 2019). Some proposed "fixes" include:

- **Education to encourage users to diversify media sources, "media diet":** the easy availability of rich and diverse media sources means that users can easily be informed by more than one source (as long as they are educated and incentivised to do so), which can counteract the "echo chamber" effect.
- **Algorithm changes:** technical proposals have been made that suggest algorithms with a higher emphasis on exploration and diversity of information sources can slow down the formation of a feedback loop.

[41] https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html

- **Create and maintain alternative fora for information sharing (online and offline):** in addition to increasing the richness of media sources, organisations interested in promoting a healthy information diet can create new and more diverse fora where different groups can meet and exchange ideas.
- **Create and maintain "argument fora":** going beyond a diversity of sources, some call for the creation of dedicated spaces (online and offline) where the best arguments from all sides of a debate are collected and curated, in order to improve the overall quality of discussion and disagreement and to help both sides of an argument understand the other.

Sensor spoofing

The rapid diffusion of cyber-physical systems and the "Internet of Things" mean that more information than ever is now collected and processed by digital sensors and computers before passing to a human to make a decision (if at all). While this has many benefits, it also creates an opportunity for adversaries to interfere with information input channels, either by providing a sensor with a malicious input (where a human might not have been fooled) or by hacking the communications or upstream processing from the sensor (Cárdenas, Amin & Sastry 2008). Some proposed "fixes" include:

- **Cross-checking across numerous modalities and independent sensors:** adding more sensors that are independent from one another, and that capture different aspects of the phenomena, can help detect spoofing and even correct for it.
- **Tamper-proof and secure sensors:** security techniques are available to increase the cost of physically tampering with, or computationally compromising, sensors and their readouts.

Digital memory tampering or manipulation

With increasing amounts of "memories" (retrievable past experiences) now stored on networked digital devices, there is a growing threat of those memories being tampered with by adversaries or through accidents. Memories could be changed by a malicious actor who gains virtual access (local or remote) to the system on which memories are stored, or through direct physical access to the hardware on which the memories are stored, or, for distributed systems, if a malicious actor is a contributing participant in the process that generates and/or records the memories. Various "fixes" have been proposed to address these different threats:

- **Software and hardware based access control:** contemporary operating systems and some hardware designs maintain records of which users and processes are allowed to access which memories, and will prevent unauthorised access.

- **Digital signatures:** a cryptographic method to generate a short text that is derived from the memory contents and a secret key, which allows detection of tampering.
- **Audit logs:** some systems maintain a record of all attempts to read, add, modify or delete content in memory, allowing the detection of unauthorised access.
- **Redundancy and error correction:** some systems retain numerous copies of the content in memory, or store the content in a format that includes in-built redundancy, which allows the detection of errors and incongruities and, if enough copies are available, allows the reconstruction of the original and the correction of errors.
- **Tamper-proof algorithms:** some cryptographic systems have been devised to retain certain security requirements even under conditions when an adversary has access to parts of the memory used by the algorithm.
- **Read-only, distributed ledgers:** some cryptographic systems have been devised to allow a large number of users who do not necessarily trust each other to nonetheless agree on a single shared collection of memories and their order.

Table 3.1: Epistemic threats from technology mapped to the basic knowledge sources

Knowledge Source	**Epistemic Threat**
Experience	• Maliciously generated synthetic media (e.g. deepfakes) • Sensor spoofing
Memory	• Digital memory tampering
Reason	• Exploitation of recommendation algorithms • Algorithmically reinforced echo chambers
Testimony	• Viral spread of misinformation in social media • Online targeting • Automated testimonial sources (bots, language models)

Table 3.2: Tech-based solutions to epistemic threats mapped to the epistemic vulnerabilities

Epistemic Vulnerability	Solutions	Potential Unintended Consequences
Attention	• Showing responses to misinformation directly next to the original content • Detecting and visually flagging problematic content • Advertising restriction • Privacy regulations • Changes to recommendation algorithms away from engagement	• Same technology can be misused for censorship • Algorithmic bias can lead to information bias • Economic damage • Innovation slowdown • Unclear what metric to use - and there is no metric for 'truth'.
Community	• Community tools for managing epistemic reputation • Education about the value of diversity • Support for a range of alternative foras for different communities to come together	• Incentivise behaviour to 'game' the reputation system • Partisanship in the education system • Competition between fora, placing further pressure on the 'attention economy'
Trust	• Education and information literacy • Karma systems • Real-time fact checking • Public support for trustworthy reporting	• Influence or capture by vested interests • Gamed behaviour • Polarization pro/con fact

	• Identity verification • Support for fora and platforms that maintain high epistemic norms even for active debates and contested topics • Secure-by-design sensors, channels and protocols	checkers • Risk of pro-government bias • Silence dissent • Such fora seen as elitist and non-inclusive. Increased competition for attention. • Increased costs excluding developing markets.
Adversarial Influence	• Communication platform monitoring • Law enforcement on communication platforms • Automated detection of fakes (bots, machine generated media) • Restricted access to dual-use technology • Cross-validation across modalities to detect spoofing, anomalies	• Potential for misuse as censorship • Reduced public scrutiny of law enforcement • Algorithmic bias leading to information bias • Innovation slowdown • Reduce visibility of single-source truth (e.g. whistleblowing)

Appendix 4: A model for understanding the costs of maintaining epistemic security and the impacts of emerging technologies thereon.

Here we build a preliminary model to describe the costs of maintaining or undermining the epistemic security of a society given the development of new information and communication technologies. The purpose of the model is to demonstrate how different factors influence the capacity of an actor (e.g. a government institution or adversary) for influencing a society's ability to organize well-informed collective action by disseminating information.

In the model, we keep track of the following costs:

Symbol	Epistemic-related cost
I	Total cost of having informed decision making, i.e. of being epistemically secure.
G	Cost of gathering information directly from the world
E	Cost of establishing and maintain a new information channel
R	Cost of retrieving information from an existing channel
A	Cost of deciding whether to attend to an information channel
M	Cost of merging information from multiple channels
D	Cost of detecting information channels controlled by adversaries
C	Cost of counteracting information channels controlled by adversaries, either through counter-messaging or coordination amongst decision makers to ignore adversarial channels
P	Cost of preventing adversaries from establishing information channels

And the following quantities:

#DM	Number of decision makers
#IU	Number of information units
#IC	Number of existing information channels
#A	Number of adversaries
#AC	Number of adversary-controlled channels

We build the model up in stages. We start from a simple, unrealistic, non-representative state, one of information scarcity, a single decision maker, and no adversaries, and slowly replace each component with ones that add complexity but better represent our current challenges. We found the process of conceptualizing the model to be just as informative as the end product.

Stage 1: Information scarcity, single decision maker

In a context of information scarcity, a single decision maker will take actions to gain information about the world. These actions can be one-off, where time and resources (costs) are expended to gain decision-relevant information, or longer-term, where costs are spent establishing an information channel, which reduces the cost for later information gathering actions.
In the first alternative, the decision-maker gathers the information directly from the world

$I = G*\#IU$

In the second alternative, the decision-maker establishes an information source and a channel to that source, and then retrieves information from that channel.

$I = E + R*\#IU$

Stage 2: Information scarcity, multiple decision makers

When there are many decision makers, and they stand to gain from collective, coordinated action, they may choose to share information sources, and choose to jointly invest in establishing information sources.

$I = E + R*\#DM*\#IU$

Because coordination requires a shared picture of the world, a channel needs to be established that connects all decision-makers, which will be more expensive than in the previous stage (*E* is higher). Note, however, that the cost to establish that channel and connect it to all decision-makers may be far lower than the cost to establish one channel per decision-maker (which would be *E*#DM*).

Stage 3: Information abundance, single decision maker

When many sources of information have been developed, or where the cost of generating information drops, the amount of available information increases. At some point it becomes more advantageous to get information from an already established information source (even if it is not controlled by the decision maker) than it is to develop new information sources. The task then shifts from creating information sources to evaluating information sources, allocating attention efficiently, and combining information from different sources into a coherent picture of the world.

I = A#IC + R*#IU + M*#IU*

Note that the cost of information retrieval (*R*) in this situation is near-zero.

Stage 4: Information abundance, single decision maker, with adversaries

In the presence of adversaries, who may control some of the numerous information channels, and who are feeding deliberately misleading information through these channels, the decision maker needs not only to allocate attention, they must also learn to identify and discard misinformation channels, and if possible take actions to restrict the adversaries' ability to create or use information channels. When the cost for the adversary to establish (or co-opt) information channels decreases, or when the cost for the adversary to masquerade as a high-value information channel decreases, the cost to the decision maker increases.

In the first alternative, the decision-maker detects and counteracts the adversarial channels, and then proceeds as before to attend, retrieve and merge information from the remaining "safe" channels.

I = (D+C)#AC + A*(#IC-#AC) + R*#IU + M*#IU*

Alternatively, the decision-maker can choose to prevent adversaries from creating information channels, and thus proceed in the knowledge that all channels are "safe".

I = P#A + A*#IC + R*#IU + M*#IU*

As the cost for adversaries to establish new channels decreases, the cost to the decision-maker increases, by either increasing the detection and counteraction costs or by increasing the prevention costs. The choice between detection and prevention may depend on the context of the decision-maker, and the various consequences of adopting one strategy or the other (for example, prevention may be unacceptably costly in a society with strong "free speech" protections).

Stage 5: Information abundance, multiple decision makers, with adversaries

In this setting it is not enough for the decision maker to identify and discard information sources controlled by adversaries, they also need to prevent adversaries from affecting other decision makers - either by preventing their access to information creation and distribution, or by exposing and communicating their control by an adversary, or by creating controlled communication channels that would gain more attention than the adversaries' channels. As the cost for the adversaries to establish channels to numerous decision makers decreases, and as the adversaries' ability to capture attention and masquerade as a high reliability information source increases, the cost for the key decision maker to maintain coordination of the numerous decision makers increases.

Again, the decision-makers have two alternatives, one to detect and counter adversaries and the other is to prevent adversaries' access to channel creation. However, this scenario now has the added cost of coordinating the response to adversaries across all decision makers.

Detection and countering:

I = D#AC + C*#AC*#DM + A*#IC*#DM + R*#IU + M*#IU*

Note the need to coordinate the response to adversarial channels, which increases the cost as the number of decision-makers increases: while detection of adversarial channels can be performed by a small subset of decision-makers, they need to be countered effectively for every decision-maker, or decision-making group, who might otherwise be influenced by them. With increased targeting and customisation, the cost of countering increases.

Prevention:

I = P#A + C*#IC*#DM + R*#IU + M*#IU*

Given the above, here are the components of the challenge we are facing:

- Democracies are defined by having numerous decision makers for key collective decisions (#*DM* is high).
- In many democracies there are also limits to restrictions that can be placed on actors' ability to access or create information channels, due to protections of "free speech" (*P* is high).
- Our current technology already makes the cost of establishing new information channels (both broadcast and targeted) near-zero (making *#AC* large).
- Current and near-future technologies make the cost of grabbing attention and faking high reliability near-zero, or at least orders of magnitude cheaper than before (making *A* and *D* high).

This means that, even while innovations in information technology have significantly reduced the cost of retrieving and merging information (making *R* and *M* low), the overall cost for those interested in maintaining informed, coordinated decision-making is rapidly increasing.

turing.ac.uk
@turinginst