DEMOS

EPISTEMIC SECURITY FOR CRISIS RESILIENCE

ELIZABETH SEGER
SAM STOCKWELL
TYREESE CALNAN
HENRY AJDER
JAMIE HANCOCK
HANNAH PERRY

JANUARY 2026

CE TaS Centre for Emerging Technology and Security

EPISTEMIC SECURITY NETWORK
DEMOS

Epistemic Security for Crisis Resilience:

An analysis of information threats, vulnerabilities, and priority interventions for the maintenance of effective crisis response capacity in democratic societies

AUTHORS:

- Elizabeth Seger (Demos)
- Sam Stockwell (CETaS / ATI)
- Tyreese Calnan (Demos)
- Henry Ajder (University of Cambridge)
- Jamie Hancock (Demos)
- Hannah Perry (Demos)

CONTRIBUTORS:

(in alphabetical order, surname)

- Phoebe Arnold (Full Fact)
- Sacha Babuta (CETaS / ATI)
- Damian Collins (Geradin Partners)
- James Eaton-Lee (Human Rights Watch)
- Dan Hawkins (Cardinal Insights)
- Pia Huesch (RUSI)
- Julian Huppert (Jesus College, Cambridge)
- François Nel (University of Lancashire)
- Sameer Padania (Macroscope)
- Nicola Stokes (Demos)
- Tim Squirrell (Foxglove)
- Gia Thom (Impress)
- Matt Warman (Independent)

ACKNOWLEDGEMENTS:

(in alphabetical order, surname)

- Shahar Avin (University of Cambridge)
- Aaron Bailey Athias (CETaS / ATI)
- Chloe Burke (Demos)
- Polly Curtis (Demos)
- Francesca Federico (University of California, Berkeley)
- Megan Hughes (CETaS / ATI)
- Chris Raggett (CETaS / ATI)
- Kevin Luca Zandermann (Tony Blair Institute for Global Change)

We would like to thank the Quadrature Foundation and Will Perrin for financially supporting this work.

Open Access. Some rights reserved.

Published by Demos January 2026
© Demos. Some rights reserved.
15 Whitehall, London, SW1A 2DD
T: 020 3878 3955
hello@demos.co.uk
www.demos.co.uk

CONTENTS

CONTENTS CONT.

ABOUT THIS PROJECT

This project, conducted in partnership between the **Demos** and **Centre for Emerging Technology and Security (CETaS)** at the Alan Turing Institute, is a core contribution to Demos' Epistemic Security program.

Epistemic security is about building and preserving healthy information pipelines that are robust to adverse influence. At Demos we see epistemic security as a critical component to the well-functioning of democratic societies. However in a modern digital era riddled with hyperrealistic fake content, epistemic security is becoming ever more difficult to defend.

This particular project highlights the role of epistemic security in underpinning a democratic society's capacity to navigate crises and complex challenges ranging from climate change and pandemics to economic collapse and election interference. In times of crisis robust information supply chains are needed to keep citizens informed with high-quality, decision relevant information. Adverse influences on information supply chains can exacerbate crises by adding fuel to ongoing crises, slowing response to crises in action, and by nurturing seeds of public dissatisfaction, distrust, and unrest to kindle crises from minor incidents.

This project gathered a leading group of experts on subjects related to epistemic security to analyse a diversity of hypothetical crisis scenarios in order to better understand epistemic drivers of crisis and to identify the most promising interventions thereon for mitigating crisis likelihood and severity.

This work sits at the intersection of Demos's strategic pillars on ***Healthy Information Ecosystems*** and ***Digital Policy*** recognising the critical and ever evolving influence new and emergency digital technologies have on shaping how information is produced, accessed, consumed, and evaluated.

ABOUT EPISTEMIC SECURITY AT DEMOS

BUILDING HEALTHY INFORMATION ECOSYSTEMS FOR DEMOCRATIC RENEWAL

At Demos we see epistemic security as critical not only to building crisis resilience in democratic societies, but also to preserving and strengthening democracy itself.

The UK is facing a democratic emergency. Declining trust in government institutions, increasing polarisation, a lack of trusted news, and the weaponisation of misinformation are driving discord and undermining healthy democratic discourse. In our agenda setting paper, Epistemic Security 2029, we argue that the UK must urgently attend to our nation's epistemic security – to securing healthy and robust information supply chains within the UK and building resilience to adverse influences thereupon.

Like in the US, democracy in the UK is already under considerable pressure after a turbulent decade. Inflation, economic inequality, high cost of living, buckling public services, and a rapid turnover of prime ministers have fostered a growing disillusionment with the potential of the UK's democratic system to serve citizens' needs, leaving it vulnerable to manipulation and antidemocratic forces.

And at the same time threats to the UK's information supply chain are worsening, driving the wedge of dissatisfaction and discord deeper and making it all the more difficult to facilitate the kind of democratically enriching discourse needed to underpin well functioning democracy.

Local news infrastructure in the UK has been decimated leaving a void in trusted information about the issues most locally relevant to citizens. Private social media companies own the primary mode of communication between citizens and between citizens and government with a profound impact on the shape and flavour of public political discourse. And powerful voices both domestically and internationally target UK citizens and elected representatives with inflammatory smears and harmful, ideological rhetoric that risks public safety, sows discord and fuels political disengagement.

Securing the UK's information supply chain and building resilience to adverse influence on our democratic processes is paramount. We believe there is no hope for desperately needed national democratic renewal without it.

JOIN THE EPISTEMIC SECURITY NETWORK

Demos launched the **Epistemic Security Network (ESN)** in June 2025 to provide a home for collective efforts to protect democracy by cleaning up and fortifying our information supply chains. Together we mobilise evidence and design strategies to strengthen the UK's information environments, and we examine media policy, regulation, public service models, local news structures, and citizens education to tackle this vital element of the democratic emergency.

Join the network to receive information about our events and to receive our ESN newsletter.

https://demos.co.uk/epistemic-security-network/

EPISTEMIC
SECURITY
NETWORK
DEMOS

EXECUTIVE SUMMARY

CONTEXT

Robust and healthy information ecosystems are critical to a democratic society's ability to navigate complex challenges and crises. From climate change to pandemics, effective and timely response to the most pressing threats against nations and citizens depends on a societies' ability to coordinate public action and response. This in turn requires robust information supply chains that keep citizens informed with high-quality, decision relevant information.

INFORMATION SUPPLY CHAIN[1]

The flow of information from its initial point of production through its dissemination, consumption, appraisal, and use to inform beliefs and decisions. Today, most stages of information supply chains are facilitated by digital mediators (e.g. social media platforms, AI assistants, news broadcast, search algorithms), and each step in the supply chain is a potential point of vulnerability – a spot where an adversarial actor, foreign influencer or unwitting blunderer could interfere, sowing discord, driving polarisation, or eroding trust.

Information Production

People produce information drawing on e.g. scientific research, their observations, or through their won reasoning or imagination.

Information Distribution

That information is distributed to wider communities via various platforms e.g. through textbooks, by a teacher, news broadcast, over email, messaging apps, or social media.

Information Acquisition

Information recipients retrieve information (e.g. tweets, blogs, news articles, video clips) via some platform. Communities often form around specific issues on particular platforms (e.g. Anti-vaccination communities on Facebook).

[1] Adapted from Seger et al.'s (2020) depiction of an 'epistemic process'.

Information Evaluation

Information recipients must decide whether they should believe the information they receive and to what extend they should use it to inform their beliefs and decisions. People will communicate with each other in trying to navigate this challenge.

Decision-making

People decide how they act in response to the information they receive. Democratic processes like elections are a form of collective decision-making.

Coordination & Action

Where challenges require collective response (like responding to a pandemic or tackling climate change) decisions and rationale must be conveyed between individuals to hopefully instigate an effective coordinated action.

Where information supply chains falter, democratic societies are made more vulnerable to a variety of potential crisis scenarios. Perturbations and threats to information supply chains can exacerbate crisis by adding fuel to ongoing crises (as during the UK's Southport riots), slowing response to crises in action (as with sluggish climate change response and poor health advice adherence during Covid), and by nurturing seeds of public dissatisfaction, distrust, and unrest to kindle crises from minor incidents (again as with Southport, and as might lead to bank runs or economic crash).

In this report we present results from a series of workshops in which Demos and CETaS convened leading experts on topics relating to information ecosystem dynamics and threats to analyse these mechanisms in action across a diverse set of hypothetical crisis scenarios.

Over two days, experts worked to develop and press a realistic set of hypothetical scenarios, to holistically map those scenarios with their constituent actors and influencing factors, to identify critical points of epistemic vulnerability and threat, and identify and test the most promising angles of intervention.

A PREDICTIVE METHODOLOGY

Such scenario planning exercises are routine practice in other areas of national security, and we have high confidence in their application to crises perpetuated by epistemic threat as well.[2] In 2018 researchers at Dstl, The Alan Turing Institute, and the Centre for the Study of Existential Risk (CSER) at the University of Cambridge undertook a first iteration of this work with notable predictive success.[3] Two of the scenarios detailed during the 2018 workshops included a 'global health' scenario that foresaw the Covid 19 pandemic and a 'xenophobic violence' scenario that closely mirrored events as they unfolded leading to the 2025 Southport riots:

[2] UK Ministry of Defence (2021). Red Teaming Handbook. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1027158/20210625-Red_Teaming_Handbook.pdf

[3] Seger, E., et al., (2020). Tackling threats to informed decisionmaking in democratic societies. https://www.turing.ac.uk/sites/default/files/2020-10/epistemic-security-report_final.pdf

- **Global Heath:** In this hypothetical scenario articulated pre-Covid, a pandemic rapidly spreads around the globe with rampant speculation about intentional origin. Populations want to know how to stay safe, but they don't know where to turn. Slow appearance of official information drives them to seek any plausible advice from friends and online communities resulting in many people following dangerous practices, which increases the burden on health services.
- **Xenophobic Violence:** In this hypothetical scenario, a xenophobic radical far-right group stages a chemical attack near a school and circulates online misinformation blaming members of a minority refugee community. In the Southport case, six years later, a knife attack on a dance class in the town of Southport was not staged with intention to frame, but the effect was nonetheless very similar with rampant disinformation online alleging the perpetrator to be a Muslim immigrant fueling a summer of violent protests and attacks.

For this project and report we repeat the 2018 mapping exercise within our new geopolitical and socio technical context in 2025. What threats to information supply chains or other epistemic failures in the UK have the potential to catalyse crises for our democracy and society? What actors are relevant, what factors would facilitate adverse events, and what are the most promising points for intervention?

METHODS

The workshops brought together a group of 25 leading experts across academia, civil society, industry and government who focus on various aspects of epistemic security including information ecosystem dynamics, online safety, election security, foreign influence, news media, cybersecurity, and AI impacts.

The workshop goals, articulated as follows, were pursued over the course three working sessions including one preparatory session and two full-day workshops:

Workshop preparation:

- **Ideate and develop a diverse set of hypothetical crisis scenarios** exacerbated or instigated by epistemic threats or vulnerabilities that could realistically occur over the next 5 years, taking into account current and projected technological, geopolitical, economic, social, legal, and environmental context. (Section 2.1)

Workshop 1:

- **Refine and prioritise scenarios** to select four scenarios featuring diverse crisis types and mechanisms to be taken forward for further analysis. (Section 2.2)

The expert working group selected the following scenarios:

Xenophobic violence leads to breakdown in trust:

Summary: A far right xenophobic faction organises a violent chemical attack on a school. They blame it on the minority community, and spreading lies with deepfakes and coordinated social media messaging in order to instigate public ill will and violence against the minority community.

AI driven breakdown of legal system:

Summary: Ubiquitous generative AI use pollutes all areas of criminal justice, enabling individuals to exonerate themselves or rewrite their own records, leading to a complete breakdown of the legal system and wider information environment. Alongside this, digital vulnerabilities enable attacks on information held by public bodies. With no non-digital backups, huge amounts of information are lost forever and public trust in the legal system is destroyed, which leads to a further breakdown of trust in public institutions. Opportunists take advantage to rewrite history, burying and confusing evidence. As proof becomes unverifiable, accountability collapses, emboldening malicious actors to commit crimes with impunity.

Bank run and economic collapse:

Summary: Hostile state actors exploit UK financial vulnerabilities through disinformation, deepfakes, and coordinated online campaigns. Exaggerated reports, troll-driven narratives, and small-scale bank thefts erode public trust. When the thefts are exposed, reassurances appear deceitful. Panic spreads via social media, fueling bank runs. Within 72 hours, misinformation and fear trigger systemic financial collapse.

Foreign tech superpower undermines UK Sovereignty:

Summary: The U.S. requisitions major tech firms and the global digital infrastructure they control, cutting access for non-U.S. entities. Business grinds to a halt, small businesses are crippled, scientific collaborations stall, essential services halt. Partial restoration comes with strict oversight, turning essential services into powerful leverage in trade, diplomacy, and geopolitical conflicts.

The 'scenario close-up's' in Appendix A provide a 'real world grounding' analysis describing how each hypothetical scenario is rooted in real world events and influencing factors.

- **Systems map the crisis scenarios** to build a holistic-as-possible view of how numerous potential factors and actors interact. (Section 2.3)

Workshop 2:

- **Identify key intervention points** in each scenario map for mitigating epistemic threats and vulnerabilities (Section 3.1)
- **Red-team the interventions** to test for unintended downstream consequences and to improve intervention robustness. (Section 3.2)
- **Identify critical cross-cutting intervention areas** that are common across a diverse set of crisis scenarios. These are the areas which, if adequately addressed, are likely to have the widest ranging impact on mitigating epistemic drivers of crisis. (Section 4)

FINDINGS

We identify seven **cross-cutting intervention areas** that featured across the range scenarios analysed. These are the most efficient intervention areas in the sense that, if adequately addressed, they are likely to have the widest ranging impact on mitigating epistemic drivers of crisis. We offer targeted recommendations for pursuing each area of intervention in the UK (Section 4).

1. **Content and data provenance, digital signatures, and watermarking:** developing and requiring technical interventions to make the source of digital content traceable.
2. **Media and information literacy:** helping people to be more discerning regarding the information they consume and more responsible in how they share content.
3. **Digital infrastructure & cybersecurity:** Making sure the basic infrastructure we rely on to process, store, transmit, and access information is secure, reliably, and dependably accessible.
4. **Crisis protocols for platforms:** Requiring social media platforms to implement protocols for mitigation of inflammatory and harmful content during time-limited, and well-defined crisis periods.
5. **Government and regulator crisis preparedness:** Updating the existing civil contingency preparedness arrangements to address information crises as a cross-cutting risk. Publishing an information crisis response protocol, similar to Canada's Elections Crisis Protocol and to provisions in the EU DSA.
6. **AI risk management:** Managing the epistemic impacts of AI systems, particularly with respect to frontier AI developments and projected developments in full multimodality, agency, deceptive capacity, and cyber applications.
7. **Rebuilding local and regional news ecosystems:** Helping people access reliable and trustworthy news content that is relevant to their local context.

FINAL TAKEAWAYS

The report concludes with final take-aways from the project reflecting key insights on the nature of epistemic security and our efforts to preserve it (Section 5).

1. **There are no silver bullet interventions for epistemic security and crisis resilience.** A diversity of interdependent technical, social, and policy interventions is needed for meaningful progress.
2. **Government must adopt a whole-of-society approach to epistemic security for crisis resilience, and cannot rely on its own levers of power alone.** Government interventions in isolation may struggle to gain traction or backfire as overreach. Forming partnerships with non-governmental voices to help e.g. amplify credible/verified information or deliver media literacy training will help build trust and a sense of community ownership of the epistemic environments.
3. **Handing over significant power to an external party that offers solutions may help solve one crisis but can lead to new, bigger ones.** We must ensure that in our present epistemic crisis mindset, we are not "handing over the keys to the city" by making decisions, passing laws, and setting precedents today that will pave the way for overreach and oppression tomorrow.

4. **Beware leaving the door open for a future authoritarian government by implementing measures or failing to close loopholes they could exploit.** It is an extension on the previous point. Policymakers must also beware of enabling authoritarian drift through a slow drip drip of measures that limit civil liberties. Each might be minor (too much slack in the definition of crisis here, an excessive intervention duration there), but the cumulative effect can be substantial.

5. **There is a significant epistemic threat plaguing UK democracy and crisis resilience rooted in social and economic instability, not just information manipulation.** After years of rising cost of living, impossible housing costs, and buckling public services, there is a pervasive sense our present democratic system is not serving people well. The underlying current of dissatisfaction is a tinderbox for conflict and polarisation.

6. **We have grounds for optimism.** A report on crises and epistemic collapse is a lot of doom and gloom. But remember that while human beings are really good at finding trouble, history is also littered with moments when people have pulled off some incredible coordinated efforts to save the day.

HOW TO USE THIS REPORT

This report lays out the methodologies employed throughout the expert workshops and presents the core recommendations emerging from the workshops for minimising epistemic threats and vulnerabilities that catalyse crises.

For academics, civil society, or government researchers interested in repeating or iterating upon the methodologies:

- **Sections 2 and 3** detail the methodologies for workshop preparation and each workshop.
- **Appendix A** may offer useful reflections on the moderators' experiences leading the crisis working groups. See below for details.
- **Appendix C** articulates some initial learnings from the first workshop on scenario ideation and mapping which follow-up projects may wish to take into consideration.

For researchers and policymakers interested in crisis scenario narratives our expert groups developed, refined, and analysed:

- **The end of section 2.2** presents the full text for the 4 scenarios chosen and refined by the expert working group during the first workshop.
- **Appendix A** provides a close up on each of the four scenarios mapped by the expert working group. Each crisis close-up includes:
    - a summary of the crisis
    - a "real world grounding" analysis detailing how the building blocks of each hypothetical scenario are rooted in real world events and technological developments.
    - full systems maps pre- and post- intervention identification
    - a "moderator insight" note reflecting on interesting features or experiences from each groups' crisis development and analysis processes that may not be evident in the output material.
    - list of "preliminary interventions" specific to each scenario as identified and prioritised by the expert working groups.
- **Appendix B** contains the full text for the 10 starting crisis scenarios developed during workshop preparation.

For those primarily interested in workshop findings and policy implications:

- **Executive summary and Introduction** provide a summary overview of the project which will allow findings to be understood in context
- **Section 4** discusses the seven cross-cutting intervention areas that featured across scenarios. For each cross-cutting intervention area we:
    - (i) Evaluate why is a critical epistemic intervention area for mitigating the likelihood and severity of crisis

  - (ii) Recommend concrete interventions for UK implementation grounded in further research building on the expert workshops
  - (iii) Identify similar interventions being employed in other jurisdictions that might be emulated or expanded upon in the UK
- **Section 5** presents a final set of final take-aways from the project reflecting key insights on the nature of epistemic security and our efforts to preserve it.

KEY TERMS

Information supply chain: The flow of information from its initial point of production through its dissemination, consumption, appraisal, and use to inform beliefs and decisions. Today, most stages of information supply chains are facilitated by digital mediators (e.g. social media platforms, AI assistants, news broadcast, search algorithms), and each step in the supply chain is a potential point of vulnerability – a spot where an adversarial actor, foreign influencer or unwitting blunderer could interfere, sowing discord, driving polarisation, or eroding trust.

Epistemic security: The field of study and work relating to the strengthening and protection of information supply chains. Epistemic security is often used as an umbrella term to unify a wide array of often disparate efforts to build healthy and democratically enriching information ecosystems.

Epistemic threat: Factors that undermine epistemic security by adversely influencing on information supply chains

Epistemic vulnerability: weak points information supply chains and wider social epistemic systems that are in danger of succumbing to epistemic threats.

Epistemic Crisis: A crisis scenario instigated or exacerbated by epistemic threats and vulnerabilities. Mechanisms of crisis exacerbation via information supply chain perturbations include: (i) fueling ongoing crises; (ii) slowing response to crises in action; (iii) nurturing seeds of public dissatisfaction, distrust, division, uncertainty, and/or unrest to kindle crises from minor incidents

Red-teaming: Red-teaming is the process by which experts search for failure modes in a plan or system to challenge assumptions and make solutions more robust. It is a practice routinely employed in national security and military settings to help agencies test defenses, improve decision-making, and anticipate threats. In the context of this report, red-teaming was used to investigate potential adverse second- and higher-order consequences of interventions for improving epistemic security.

Cross-cutting intervention areas: Common kinds or themes of intervention that have been independently recommended across a diverse set of crisis scenarios. These are the 'most efficient' intervention areas in the sense that, if pursued and effectively implemented, they are most likely to have the furthest reaching impact in mitigating the likelihood and severity of a wide variety of potential crisis scenarios. are likely to have the widest ranging impact on mitigating epistemic drivers of crisis.

INTRODUCTION

Robust and healthy information ecosystems are critical to a democratic society's ability to navigate complex challenges and crises. From climate change to pandemics, effective and timely response to the most pressing threats against nations and citizens depends on a societies' ability to coordinate public action and response. This in turn requires robust information supply chains that keep citizens informed with high-quality, decision relevant information.

When information supply chains falter, messages scramble, trust crumbles, and divisions widen as people struggle to differentiate fact from fiction, and the consequences can be terrible. The 2024 **Southport riots** were case and point; delayed official reporting regarding an isolated act of violence left a news void to be filled with speculation and falsehoods. Fabricated stories about perpetrator's racial identity and immigrant status spread rapidly on social media platforms driving a violent wave of xenophobia yielding incitements to violent retaliation against Muslim and migrant communities and a summer of riots across the UK.[4]

The fight against **climate change** has similarly suffered from prolonged campaigns to manipulate information supply chains, involving industry and government efforts to tamper with the foundational stages of scientific inquiry and to manipulate public narratives to sow doubt and slow policy intervention.[5,6] Now consequences of climate change from rising sea levels to drought are threatening large-scale population displacement and famine, with knock-on implications for economic and political stability.[7]

Our **elections** are vulnerable too. Propaganda and mis/disinformation campaigns targeted at national and local elections are nothing new, but the scale and precision of attack has expanded. We've seen a rapid proliferation of deepfakes used to impersonate and intimidate candidates, and influence campaigns have transformed from slow, expensive and data poor endeavours, to fast, AI-enabled, and data rich targeted attacks on public sentiment and belief.[8] The UK is also now suffering influence from classic foreign aggressors as well as traditional allies and we are not sure how to respond.[9] If substantial election interference is ever suspected we have no warning protocol in place, nor any plan to respond, and consequently any attempt to delay, cancel, or recount an election would likely be catastrophic to public trust in democratic process and government.[10]

[4] Perry et al. (2025). Community Disorder: How do we prevent an information emergency? https://demos.co.uk/research/community-disorder-how-do-we-prevent-an-information-emergency/

[5] Oreskes, N. (2011). Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Climate Change. Bloomsbury. Print.

[6] Milman, O. (2024). 'Smoking gun proof': fossil fuel industry knew of climate danger as early as 1954, documents show. The Guardian. https://www.theguardian.com/us-news/2024/jan/30/fossil-fuel-industry-air-pollution-fund-research-caltech-climate-change-denial

[7] Amnesty International (2025). When people are displaced by climate change, what rights do they have? https://www.amnesty.org/en/latest/campaigns/2025/10/when-people-are-displaced-by-climate-change-what-rights-do-they-have/

[8] Dobber, T. et al. (2020). Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? The International Journal of Press/Politics. https://doi.org/10.1177/1940161220944364

[9] Stockwell, S., et al. (2025). "AI-Enabled Influence Operations: The Threat to the UK General Election," CETaS Briefing Paper https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-uk-general-election

[10] Demos (2026) Epistemic Security briefing: Elections Bill. https://demos.co.uk/research/epistemic-security-briefing-the-elections-bill/

At Demos we have developed a program of work around the preservation of democratically enriching information ecosystems using the term '**epistemic security**' as an umbrella to unify the various areas of study pertaining to information ecosystem health and impacts thereon.[11] We think in terms of '**information supply chains**' - the processes by which knowledge is produced, disseminated, consumed, appraised, and used to inform decisions. Like water, oil, and electricity, information is a critical national resource. It is essential to a democratic society's capacity for well-informed decision making and collective action, and it should be protected as such.[12] Epistemic security is about keeping information supply chains safe. It involves reinforcing supply chain vulnerabilities (epistemic vulnerabilities) and guarding against adverse influences (epistemic threats).

Perturbations and threats to information supply chains exacerbate crises in three ways:

1. **Adding fuel to ongoing crises** - in the way inflammatory and extremist content spread widely on social media fuelled the January 6th insurrection on the US capital.[13]
2. **Slowing response to crises in action** - as with sluggish climate change response thwarted by prolonged attacks on climate science,[14] and poor health advice adherence during Covid due to rampant disinformation.[15]
3. **Nurturing seeds of public dissatisfaction, distrust, and unrest to kindle crises from minor incidents** - as with the Southport riots as described above.

In this report we present results from a series of workshops in which Demos and CETaS convened leading experts on topics relating to information ecosystem dynamics and threats to analyse these mechanisms in action across a diverse set of hypothetical crisis scenarios.

Over two days, experts worked to develop and stress-test a realistic set of hypothetical scenarios, to holistically map those scenarios with their constituent actors and influencing factors, to identify critical points of epistemic vulnerability and threat, and identify and test the most promising angels of intervention.

WHY NOW?

In 2020 a number of this report's authors published the first epistemic security report which coined the term and presented a similar series of expert workshops analysing a set of hypothetical crisis scenarios instigated or exacerbated by epistemic threats and vulnerabilities.[16] The scenarios considered a variety of epistemic threats and vulnerabilities - ranging from coordinated foreign influence campaigns to nascent deep fake technology in the hands of unwitting blunders - and crisis types ranging from economic collapse to political character assassination. It was a highly predictive endeavor. Two of the scenarios detailed during the 2018 workshops included a 'global health' scenario that foresaw the Covid-19 pandemic and a 'xenophobic violence' scenario that closely mirrored events as they unfolded leading to the 2025 Southport riots.

[11] Demos (2024). Epistemic Security 2029: Protecting the UK's information supply chain and strengthening democratic discourse for the next political era. https://demos.co.uk/blogs/epistemic-security-2029-protecting-the-uks-information-supply-chain-and-strengthening-democratic-discourse-for-the-next-political-era/

[12] A similar argument is presented in Kallioniemi, P. (2025). Beyond Defence: A Proactive Strategy for the West in the Information Domain. International Centre for Defence and Security. https://icds.ee/en/beyond-defence-a-proactive-strategy-for-the-west-in-the-information-domain/

[13] Eisenstat, Y., Hendrix, J. & Kreiss, D. (2024). Tech Platforms Must Do More to Avoid Contributing to Potential Political Violence. Tech Policy Press. https://www.techpolicy.press/tech-platforms-must-do-more-to-avoid-contributing-to-potential-political-violence/

[14] Oreskes, N. (2011). Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Climate Change. Bloomsbury. Print.

[15] Artificial Intelligence and Digital Public Health Special Interest Group. 'Written evidence submitted by the Faculty of Public Health (SMH0011)'. https://committees.parliament.uk/writtenevidence/132776/html/ (Accessed Nov 20, 2025).

[16] Seger, E. et al. (2020). Tackling threats to informed decision- making in democratic societies: Promoting epistemic security in a technologically-advanced world. The Alan Turing Institute. https://www.turing.ac.uk/news/publications/tackling-threats-informed-decision-making-democratic-societies

Unfortunately the 2018 epistemic crisis mapping exercises struggled to gain policy traction published amidst the Covid turmoil, but the report has received renewed attention in the past year. Following the Southport riots and amid mounting concerns about foreign election influence, concentrated corporate control digital platforms, and the impact of rapidly evolving AI technologies on information consumption, calls for greater crisis preparedness and mitigations to growing information threats are intensifying.

This report offers a critically-timed follow-up to the first. We have reconducted the crisis scenario ideation and mapping exercises of the first report accounting for current geopolitical, technological, economic, social, and legal context with a 5 year forecast of predicted developments (Section 2). As in the first workshop series, experts also identified and red-teamed potential interventions for each scenario (Section 3).

But in this second iteration we went a step further in working solutions for bolstering systems-wide epistemic security which we offer as the project's core contribution. Building on the expert intervention analysis, we reviewed all workshop outputs to identify the critical '**cross-cutting intervention areas**' that featured prominently across the diverse range of workshopped scenarios (section 4). These are the most efficient intervention areas in the sense that, if adequately addressed, they are likely to have the widest ranging impact on mitigating epistemic drivers of crisis.

The report concludes with final take-aways from the project reflecting key insights on the nature of epistemic security and our efforts to preserve it (Section 5).

ABOUT THE WORKSHOPS

Demos and the Center for Emerging Technology and Security (CETas) jointly designed and delivered a series of workshops to forecast and analyse potential crisis scenarios instigated or exacerbated by current and emerging epistemic threats and vulnerabilities.

Expert working groups

The workshops brought together a group of 25 leading experts across academia, civil society, industry and government who focus on various aspects of epistemic security. Some experts were 'generalists' who brought a holistic view of interconnected information ecosystem dynamics. Others were 'subject area' experts with deep insight to specific epistemic vulnerabilities, threats, threat actors, or mechanisms - for example, news media publishers and regulators, AI safety researchers, national security practitioners, and experts in cybersecurity, psychology, and electoral procedure. The workshop delivery team also held significant experience in designing and moderating systems mapping, red-teaming, and war-gaming exercises.

Format & Goals

The workshop goals, articulated as follows, were pursued over the course three workings sessions including one preparatory session and two full-day workshops:

Workshop preparation:

- **Ideate and develop a diverse set of hypothetical crisis scenarios** exacerbated or instigated by epistemic threats or vulnerabilities that could realistically occur over the next 5 years, taking into account current and projected technological, geopolitical, economic, social, legal, and environmental context. (Section 2.1)

Workshop 1:

- **Refine and prioritise scenarios** to select four rich and feasible scenarios featuring diverse crisis types and mechanisms to be taken forward for further analysis. (Section 2.2)
- **Systems map the crisis scenarios** to build a holistic-as-possible view of how numerous potential factors and actors interact to form a dynamic web of influences impacting the materialisation of each crisis scenario. (Section 2.3)

Workshop 2:

- **Identify key intervention points** in each scenario map for mitigating epistemic threats and vulnerabilities to reduce the likelihood and/or severity of the crisis scenario. (Section 3.1)
- **Red-team the interventions** to test for unintended downstream consequences and to improve intervention robustness. Red-teaming is a practice routinely employed in national settings to test failure modes in a plan or system to improve decision-making and build more robust solutions.[17] (Section 3.2)
- **Identify critical cross-cutting intervention areas** that are common across a diverse set of crisis scenarios and present the most efficient epistemic security interventions; these are the areas which, if adequately addressed, are likely to have the widest ranging impact on mitigating epistemic drivers of crisis. (Section 4)

[17] UK Ministry of Defence. (2021). Red Teaming Handbook.https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1027158/20210625-Red_Teaming_Handbook.pdf

2. CRISIS DEVELOPMENT & SYSTEMS MAPPING

METHODOLOGY: PREPARATION & WORKSHOP 1

The first stage of our project involved developing and mapping out a set of hypothetical crisis scenarios. The aims of this first stage were threefold:

- Ideate a variety of crisis scenarios grounded in real world events and influenced by real and emerging technological, social, political, environmental, and economic factors. (Section 2.1)
- Refine and prioritise four scenarios for further analysis. (Section 2.2)
- Systems map the selected scenarios to identify the breadth of factors and actors influencing the process by which the crisis scenario rolls out. (Section 2.3)

2.1 SCENARIO IDEATION (WORKSHOP PREPARATION)

Prior to the first workshop, the Demos/CETaS team brought together eight leading experts on information ecosystem dynamics and emerging tech to ideate an initial set of crisis scenarios. Several of these experts had previous experience in defence and security wargaming and helped to collaboratively guide the exercise.

We started by brainstorming crisis scenario building blocks. These were components, which when brought together in different combinations, could yield a variety of hypothetical crises in which some epistemic vulnerability or threat would catalyse or exacerbate the realisation of a crisis event. Examples of building blocks drawn from the initial brainstorm are in Table 1.

TABLE 1
CRISIS SCENARIO BUILDING BLOCKS

SCENARIO BUILDING BLOCKS	EXAMPLES
Kinds of crisis	• Acute environmental catastrophe (e.g. flood, hurricane.) • Climate change effects (e.g. famine, displacement, etc.) • Health (e.g. pandemic) • War / civil-war • Election failure • Legal system failure • Xenophobic mass-violence • Non-ideological mass violence • Resource distribution failure • Authoritarian takeover / democracy collapse • Critical infrastructure failure (e.g. mass extended internet or electricity outage) • Loss of human knowledge / skill
Actor dynamics	• Targeted coordinated disinfo campaign / epistemic attack. (e.g. info campaign to sway an election or a cyberattack). Foreign or domestic. • Untargeted coordinated disinfo campaign (e.g. to sow general confusion, discord, or distrust). Foreign or domestic. • Uncoordinated bad actors (e.g. individuals producing a sharing harmful xenophobic AI slop / uncritically reposing knowingly harmful and divisive content online) • Unintentional Blundering: e.g. people sharing content online with good intention but unwittingly sowing harmful disinformation (e.g. vaccine disinfo) that gets picked up and amplified • Mass confusion / epistemic babble. No clear actors work intentionally or unintentionally to perpetuate a crisis situation. A confluence of technological, political, and social factors leads to gradual degradation to 'epistemic babble' - the complete inability for citizens to distinguish truth from fiction and trustworthy from untrustworthy sources. Undercuts societies ability to respond to any complex challenges or crises

SCENARIO BUILDING BLOCKS	EXAMPLES
Influencing technological factors With 3-4 year projection on technological development	• Core data set corruption / theft or leakage of core data sets • Cyberattacks (on critical infrastructure, financial institutions, government departments, etc.) • Bias in / control of divisive social media ranking algorithms • Online echo chambers • Fragmentation of online information space • Increasingly deceptive AI (deceives users and safety risk assessments) • Convincing multimodal deepfakes • AI slop degrades data commons • AI Agent polluting information ecosystems (e.g. updating wikipedia hourly) • AI Agents with extensive and permissive access to personal data and accounts (e.g. contact info, browser history, financial accounts, email, calendars, etc.) • AI bot relationship (messing with trust heuristics / proving a route for user manipulation) • 'Grooming' bots deployed by malicious actor groups • Automated fact checking / content moderation (benefits and potential harms from shortcoming) • Content provenance / watermarking mechanisms (benefits and current shortcomings) • Hyperpersonalised mis/disinformation • Vulnerable infrastructure (e.g. undersea cables, data centres) • Monopolies / centralised ownership of critical digital infrastructure (largely in the US) yields concentrated power and political influence. Makes others vulnerable to digital infrastructure access failing for being cut off. • Intense geopolitical competition on tech (e.g AI race, chip export bans) • Digital exclusion

SCENARIO BUILDING BLOCKS	EXAMPLES
Influencing social, political, economic, and legal factors	• Current or upcoming legislation • UK Online Safety Act • UK Election Bill • UK Curriculum review • UK Digital Inclusion Action Plan • UK AI Bill(?) / EU AI Act / various US state AI legislation • Rising cost of living / housing costs • Political instability (e.g. high PM turnover riddled with scandals) • Low public trust in government • Buckling public services (contributes to citizen disenchantment with gov.) • Overburdened legal system (can it deal with an onslaught of fabricated evidence) • Rising polarisation (including over key issues like immigration) / echo chambers (perpetuated online) • Distrust of media • Decimated local news ecosystems leaves information voids open for speculation and conspiracism

After the initial building block brainstorm, we then started combining factors (influencing factors, technological exacerbations, kinds of crises) to generate a large pool of approximately 40 short (2-3 sentence) crisis vignettes.[18] The expert group then worked to sort, combine, and refine the pool of short crisis vignettes into ten scenario narratives that would be brought forward into the two subsequent full-day workshops for further refinement and analysis by a larger group of experts. These ten starting scenarios were composed to span a diverse set of kinds of crises (e.g. climate, health, election), actor dynamics (e.g. targeted attacks v. uncoordinated blundering), and influencing factors that pose or heighten epistemic threats and vulnerabilities throughout the crisis narrative. The ten starting scenarios and their building blocks are summarised in Table 2. The full narratives for the ten starting scenarios can be found in Appendix B.

[18] Short crisis narratives or partial crisis narratives in the following style, for example: Archivists struggle to build an official account of events following a short but violent conflict due to the volumes of AI generated content shared during the fighting. In turn, this threatens legal proceedings in international courts and threatens to warp how future generations remember the conflict. It emerges that one of the nations involved is actively trying to sabotage records such as wikipedia and other repos to create false memories and records of events to paint them in a better light and dodge justice.

TABLE 2
SUMMARY OF 10 STARTING SCENARIOS DEVELOPED FOR INPUT TO FULL-DAY EXPERT WORKSHOPS

1. XENOPHOBIC VIOLENCE

Crisis type	Actor Dynamic	Influencing tech factors
Ethnic violence & collapse of trust	Targeted coordinated (domestic)	Deepfakes, social media

Summary: An extremist xenophobic faction organises a violent chemical attack on a school. They blame it on the minority community, and spreading lies with deepfakes and coordinated social media messaging in order to instigate public ill will and violence against the community.

2. POPULATION CONTROL WITH AGENTIC AI

Crisis type	Actor Dynamic	Influencing tech factors
Loss of human agency + Authoritarian takeover	Targeted coordinated + uncoordinated blundering	Agentic AI

Summary: Popular agentic AI assistants heavily influence people's lives - not just what information they consume, but directly dictating where they go, when they travel, what they buy, who they see, and what conversations they have. The companies that provide these tools (and the government to which they pander) have a perfect surveillance tool and ability to manipulate public activity and opinion. They use it to orchestrate unrest and coordinate support for an authoritarian takeover domestically and internationally.

3. CLIMATE CRISIS SPIRALS TO COLLAPSE OF CIVIC DISCOURSE

Crisis type	Actor Dynamic	Influencing tech factors
Climate	Coordinated & targeted → uncoordinated blundering → epistemic babble	Streaming platforms, generative AI, social media, AI bots

Summary: A government-backed documentary downplaying climate urgency dominates streaming platforms, amplified by AI-generated social media content posing as expert opinion. Public pressure rolls back net-zero commitments and reactivates fossil fuel infrastructure. Critics face deepfakes and threats, while radicalised climate vigilantes emerge. Society polarises, truth fractures, and collective response to the climate crisis collapses.

4. LOSS OF HUMAN KNOWLEDGE AND SKILL

Crisis type	Actor Dynamic	Influencing tech factors
Epistemic	Blundering + Epistemic babble	AI, Agentic AI, Data security

Summary: AI systems become central to global knowledge infrastructure, replacing human expertise. Over time, foundational datasets degrade, setting research back decades. Humanity enters an epistemic dark age.

5. MANIPULATING HISTORY TO ESCAPE ACCOUNTABILITY FOR WAR ATROCITIES

Crisis type	Actor Dynamic	Influencing tech factors
War atrocities + epistemic	Targeted coordinated	Cyberattack, AI Agents, Generative AI

Summary: A cyberattack cripples archives and libraries, erasing access to vital records. Opportunists take advantage to rewrite history burying and confusing evidence of war crimes. As proof becomes unverifiable, accountability collapses, emboldening regimes and normalising mass violence against civilians on a global scale

6. BANK RUN AND ECONOMIC CRASH

Crisis type	Actor Dynamic	Influencing tech factors
Economic	Targeted coordinated (foreign influence) + uncoordinated blundering	Social media, generative AI, encrypted messaging

Summary: A hostile state uses LLMs to persistently erode economic confidence. After viral footage of the Chancellor crying, disinformation floods encrypted channels: fake ATM closures, spoofed alerts, deepfakes. Panic spreads, officials lose control, and a real bank run unfolds. Disinformation becomes reality within 72 hours.

7. YOUNG VOTER ISOLATION AND ELECTION INFLUENCE

Crisis type	Actor Dynamic	Influencing tech factors
Election / Democratic	Targeted uncoordinated (opportunistic influence)	Virtual reality, AI agents/ avatars, AI relationships

Summary: A populist party rapidly gains support by targeting young voters isolated in virtual spaces and trusting AI avatars. Influencers and foreign actors use memes, virtual campaigns, and shaming tactics to silence opposition. The movement grows largely unchecked, sweeps the election, and establishes authoritarian control.

8. FOOD SUPPLY CHAIN COLLAPSE

Crisis type	Actor Dynamic	Influencing tech factors
Climate & Resource distribution	Mainly uncoordinated + blundering w/ some opportunistic targeting	Social media

Summary: A crop virus devastates food supplies already strained by climate change. Disinformation fuels panic and unrest, while collapsing trust leaves governments paralysed amid worsening malnutrition and deaths

9. CYBERATTACK ON EDI INITIATIVES

Crisis type	Actor Dynamic	Influencing tech factors
Misogynistic surge	Targeted coordinated	Cyberattack, Data corruption

Summary: A coordinated men's rights network launches cyberattacks against major UK employers' HR systems, corrupting recruitment data to fabricate evidence that women are being hired despite inferior qualifications. Fake statistics flood social media, triggering widespread discrimination and lawsuits.

10. FOREIGN CONTROL OF CRITICAL INFORMATION INFRASTRUCTURE		
Crisis type	**Actor Dynamic**	**Influencing tech factors**
Economic, social, political, epistemic	Targeted coordinated	Digital infrastructure ownership concentration
Summary: The U.S. requisitions major tech firms and the global digital infrastructure they control, cutting access for non-U.S. entities. Business grinds to a halt, small businesses are crippled, scientific collaborations stall, essential services halt. Partial restoration comes with strict oversight, turning essential services into powerful leverage in trade, diplomacy, and geopolitical conflicts.		

2.2 SCENARIO PRIORITISATION AND REFINEMENT (WORKSHOP 1A)

The first full day workshop brought together a group of 25 leading experts across civil society, academia, industry and government to analyse the 10 starting scenarios. The experts included epistemic security generalists who are leaders in research regarding information ecosystem dynamics in the context of human psychology and emerging technology. We also brought together subject area experts in frontier AI capabilities and safety, foreign influence, cybersecurity, news media, content moderation and online regulation, economics, and national security.

Our first step as a group was to narrow down the list of 10 starting scenarios to 4 refined scenarios that we would take forward for further analysis. This prioritisation exercise was an opportunity for the full expert group to contribute to the scenario ideation process, bringing their more diverse expertise to bear in helping to make scenarios more pressing and realistic. The 10 starting scenarios (Table 2 / Appendix B) served as building blocks for focused, productive discussion.

The prioritisation and refinement exercise took place in two phases:

Phase 1: Prioritisation

Expert participants were randomly broken into five working groups. Each working group was tasked with identifying and ranking their top 4 scenarios from the starting set, keeping in mind that there would be opportunity to modify the scenarios later (phase 2) in case there were any key features or omissions from the original that would prevent an otherwise important scenario from making the cut for further analysis. In other words, participants were asked to rank scenarios based on the potential they saw for a refined scenario, not strictly based on its original articulation.

The groups were asked to consider the following questions in their scenario ranking process:

- How realistic is the scenario? (e.g. are there any core narrative components that are unlikely to happen due e.g. to the current regulation or the state of expected technological capability over the next 5 years?)
- How well trodden are mechanisms leading to crisis? (we might want to focus on ones that are less well understood)
- How catastrophic are the potential consequences for society? (we might want to focus on analysing mitigations for the most harmful scenarios)
- Does the set of 4 scenarios have a good variety of mechanisms (e.g. actors and factors driving the crisis realisation process) and kinds of crises?

We then came together as a full group to compare rankings and collectively decide which four scenarios we would carry forward. There was immediate full group consensus on moving forward Scenarios 6 (Bank Run) and Scenario 10 (Foreign Control of critical information infrastructure) with only minor modifications expected. There was a wider spread of opinion regarding the other scenarios, but through further discussion we came a consensus about moving forward Scenario 1 (Xenophobic Violence) due to a clear close link with current events making it a high likelihood scenario, and Scenario 5 (Manipulating History) combined with components of Scenario 4 (Loss of human knowledge).

Scenario 8 (Food supply chain collapse) was deprioritised because it featured similar mechanisms to Scenario 6 (Bank run) in its spiralling-collapse-of-trust dynamic. While both scenario 6 and 8 were considered pressing, realistic, and important targets of further analysis, it was felt that scenario 6 was more immediately pressing and likely to galvanise policy action.

Scenario 3 (Climate crisis spirals to collapse of discourse), also featured the spiralling-collapse-of-trust dynamic but it is more difficult to identify the specific crisis which would make it more difficult to analyse. Some experts argued that disintegration of our ability as a society to distinguish truth from fiction and trustworthy sources from untrustworthy sources is itself a crisis, as it undermines a society's capacity to engage in productive democratic discourse and make timely decisions to respond to complex external challenges like climate change or pandemic. Indeed the 2020 Epistemic Security project did analyse a scenario of this sort - dispersed societal epistemic breakdown - which was titled "epistemic babble".[19] We ultimately agreed to deprioritise the scenario for this workshop because dispersed breakdown in trust and communication ultimately evolves as a theme influencing all the crisis scenarios we were discussing. As such, the dynamic would not be overlooked in our analyses.

Scenario 2 (Population control with agentic AI) was generally recognised as illustrating a critical epistemic threat from emerging technology, but that the crisis in the scenario was difficult to pinpoint. Instead the expert working group concluded that the likely influence of AI agents on epistemic security in coming years should be integrated into the prioritised scenarios where appropriate.

Scenario 9 (Cyberattack on EDI Initiative) was set aside as being unnecessarily polarising and also because the cyberattack threat also features in other scenarios.

Finally, Scenario 7 (Young voter isolation and election influence) was deprioritised, not because it was less important or unrealistic, but because election influence is a very well trodden topic with substantial policy attention. One of our goals with these workshops is to help illuminate the wider breadth of potential crisis scenarios influenced by epistemic threats and vulnerabilities.

Overall, the following scenarios were prioritised to undergo further refinement:

- Scenario 1 - **Xenophobic violence**
- Scenario 5 (with aspects of 4) - **AI driven breakdown of legal system**
- Scenario 6 - **Bank run and economic crash**
- Scenario 10 - **Foreign control of critical tech infrastructure**

[19] Seger, E. et al. (2020). Tackling threats to informed decision- making in democratic societies: Promoting epistemic security in a technologically-advanced world. The Alan Turing Institute. https://www.turing.ac.uk/news/publications/tackling-threats-informed-decision-making-democratic-societies

Phase 2: Refinement (making the scenarios "even worse")

The next activity was to update and refine the selected scenarios to make them "even worse". This meant adding or modifying detail to the scenarios (e.g. influencing factors, actors, background conditions, narrative) that would (a) make negative impact more severe, and (b) make the scenario more realistic and therefore more likely to happen.

For this activity, experts broke out into four subgroups to refine each of the scenarios. Subgroups were self-selected as we could not pre-assign members based on their expertise without knowing which scenarios we would ultimately be considering. Moderators ensured an even balance of generalists between groups and allowed subject area experts to self-sort into the most relevant scenarios according to their knowledge and experience. A moderator accompanied each group to guide discussion.

The refinement process began with each group undertaking a PESTLE analysis of their respective scenarios. A PESTLE involves identifying the various Political, Economic, Social, Technological, Legal, and Environmental factors that might be influencing a situation.

FIGURE 1
PESTLE CHART GUIDELINES

P	**Political Factors**	Includes government policy, political stability, taxation, trade regulations and lobbying dynamics.
E	**Economic Factors**	Covers inflation, interest rates, economic growth, unemployment, exchange rates, and consumer spending.
S	**Social Factors**	Encompasses cultural trends, demographic changes, lifestyle preferences, education, and population shifts.
T	**Technological Factors**	Involves innovation, R&D, automation, digital infrastructure, and technological adoption, technological capabilities
L	**Legal Factors**	Focuses on laws and regulations such as employment, tech regulation, intellectual property, and safety standards.
E	**Environmental Factors**	Highlights sustainability, climate change, ecological concerns.

Using a PESTLE chart as a guide, the expert groups were first prompted to identify factors likely to be influencing their scenarios as currently articulated. They were then prompted to build on this initial list by adding or trading in factors that might make the scenario consequences more severe or increase the likelihood of the crisis coming to fruition. For example, in the Bank Run scenario details about digital infrastructure vulnerability were added that were not present in the original. The scenario's heavy reliance on deepfake videos of cash point mobs and financial leader statements as a primary mechanism for driving panic was also swapped for malicious actors manipulating and misconstruing narratives around real events to slowly eat away at public confidence. The latter was felt to be a more realistic process.

The expert groups then worked with the moderators and notetakers to update their respective scenario narratives to reflect the changes introduced during the PESTLE analysis. Not all of the details articulated during PESTLE analyses were included in the updated scenarios. The scenario narratives only convey the core mechanisms by which the scenario rolls out over time (e.g. event

A leads to event B, event B leads to event C, and so on). The additional details on the numerous influencing factors identified during the PESTLE would underpin the subsequent systems mapping exercise (section 2.3).

The full updated narrative texts for each scenario are presented in the boxes that follow.

SCENARIO 1
XENOPHOBIC VIOLENCE TO COLLAPSE IN TRUST

Summary: A far right xenophobic faction organises a violent chemical attack on a school. They blame it on the minority community, and spreading lies with deepfakes and coordinated social media messaging in order to instigate public ill will and violence against the minority community.

- It is shortly before Purdah in 2029. There has been exacerbating pressures on the economy and the refugee crisis has worsened. The current Government has found no solution to where asylum seekers are held (hotels). The BBC is under duress and likely diminished funding and slumping trust.
- Foreign powers continue to provide legitimacy and endorse attacks on migrants and asylum seekers.
- Migrants have been framed as political actors in support of a foreign power.
- A radical xenophobic group decides to turn the population of their country against a minority community of recently-arrived refugees.
- They mount a low-grade chemical attack near a school in a poverty-stricken suburb of a major city, taking several videos of the operation.
    - This group has significant financial support and connections with those with mass platform influence among English-language nations.
    - This group also has support from other far-right groups in other European nations.
- The videos are edited to make it appear as if a recognised figure from the refugee community was the perpetrator and released as “breaking news” on right-leaning social media groups and through various messaging apps.
- The extremist group also releases messages saying “the government is going to cover this up and prevent us from speaking freely about the truth of what’s happening to our country”.
- Because of the upcoming election timing, a social media campaign continues amidst rising violence that drip-feeds additional videos indicating specific refugees participating in the election illegitimately (voting). This calls into question the legitimacy of the election.
- 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.

- Shocked, afraid and angry mobs rally and carry out vandalism and, on some occasions, assault.
- A deluge of both real and deepfake videos seemingly being recorded on phones in real time are spread on social media and messaging apps confusing the narrative to further sow discord and fear.
- 'Trusted' mainstream media platforms share inaccurate information via AI overview alert at a critically vulnerable time in violence.
- The government acts to counter the spread of disinformation, pushing for new policies like requiring encrypted messaging apps to install backdoors to enable surveillance during crises and to broaden the definition of "illegal content" as articulated in the Online Safety Act.
- The general public shifts communication channels to encrypted platforms, making visibility of communications harder for the authorities.
- Journalists and civil society leaders begin to sound the alarm. Is this overreach?
- Public trust collapses rendering government incapable of doing anything to mitigate the epistemic risks it sought to address. Any action at all will be met with extreme public scepticism and distrust, taken as proof of state overreach and attempts at population control.
- Policing is rejected and undermined - officers are attacked or ignored. Rise is vigilante attacks across multiple communities.
- To convince the public of its continued commitment to preserving freedom of expression, the Government is forced to abandon the Online Safety Act as a whole.

SCENARIO 5

AI DRIVEN BREAKDOWN OF LEGAL SYSTEM

Summary: Ubiquitous generative AI use pollutes all areas of criminal justice, enabling individuals to exonerate themselves or rewrite their own records, leading to a complete breakdown of the legal system and wider information environment. Alongside this, digital vulnerabilities enable attacks on information held by public bodies. With no non-digital backups, huge amounts of information are lost forever and public trust in the legal system is destroyed, which leads to a further breakdown of trust in public institutions. Opportunists take advantage to rewrite history, burying and confusing evidence. As proof becomes unverifiable, accountability collapses, emboldening malicious actors to commit crimes with impunity.

- A new AI product that can create fake evidence and manipulate existing evidence is created to help exonerate criminals and marketed to organised crime groups, and later is made freely available online and accessible to the general public
- Various criminal justice and police data systems are compromised. Some evidence is destroyed completely, while new evidence is planted and existing records are manipulated, rendering it impossible to discern true information from what has been fabricated.
- With no non-digital backups, information relating to court cases, prison records, and police records is lost forever. This further emboldens serious criminals to conduct more crime at a larger scale with impunity.
- As organised crime groups' power and influence grows, they are able to further corrupt or blackmail officials within the legal system, e.g. through sophisticated AI-enabled sextortion.
- Over time, trust in the legal system completely breaks down as victims and their families do not receive justice.
- This then leads to widespread public anger against the system, including physical protests and disruption.

SCENARIO 6

BANK RUN AND ECONOMIC COLLAPSE

Summary: Hostile state actors exploit UK financial vulnerabilities through disinformation, deepfakes, and coordinated online campaigns. Exaggerated reports, troll-driven narratives, and small-scale bank thefts erode public trust. When the thefts are exposed, reassurances appear deceitful. Panic spreads via social media, fueling bank runs. Within 72 hours, misinformation and fear trigger systemic financial collapse.

- Over an extended period of time hostile state actors cater a continuous stream of reports elucidating instabilities and corruption in UK financial institutions. The reports are often exaggerated and misleadingly framed, but are seeded in a grain of truth making them difficult to refute.
- For example, the hostile actors identify politicians with shareholdings in important banks or financial institutions, or with compromising relationships to key financial sector leaders. The reports imply high likelihood of government coverups of financial instability and willingness to bail banks out regardless of risky behavior. These narratives eat away at public confidence.

- Meanwhile, troll factories (people and bots) begin producing thousands of plausible-sounding posts, articles, and personal testimonies, all hinting at behind-the-scenes financial collapse. This sows doubt about the UK's economic stability.
- Government and financial leaders offer public reassurances about the state of the economy, the stability of the pound, and the importance of saving and investing with banks to plan for your financial future.
- Against this backdrop the hostile actors have been carrying out a gradual series of real thefts, hacking systems to steal small amounts (e.g. £1) out of tens of thousands of personal accounts. It initially goes unnoticed, but then the hammer falls!
- The hostile actor uncovers the thefts, reporting that the banks went out of their way to cover it up. They replay the real clips of public figures giving re-assurances during the same period that the accounts were being compromised. They say, either the public leaders were unaware of the thefts or were deliberately covering them up and are now caught in the lie. Both are bad!.
- Meanwhile banks are slow to respond, plagued with legacy IT and unprepared for this kind of information warfare.
- Within hours, simmering public uncertainty about the country's financial stability quickly spirals to outright panic. Well-meaning citizens begin sharing amateur financial advice—urging followers to pull cash out of banks and move savings into gold and high risk crypto investments. Martin Lewis deepfakes proliferate urging the same.
- Cascading financial collapse builds its own momentum. The hostile state actors need only lightly nudge it along. They deploy real people and AI agents to hold one-to-one and one-to-many conversations on encrypted messaging platforms claiming that major banks are on the verge of collapse.
- The first few instances of people being unable to execute based financial transactions occur (e.g. unable to withdraw cash or transfer between accounts). These stories are picked up and amplified. Panic heightens building into a genuine bank run.
- Real and fake footage of mobs at cashpoints circulate.
- But it's not just cash withdrawals. Because it's the digital age, everyone has access to banking apps on their phone with the ability to instigate immediate fund transfers. The tsunami of customer activity overwhelms the system. More and more financial transactions fail, and some banking apps crash, cutting off citizen account access.
- Officials try to respond, urging calm, but their messages are dismissed. No one trusts government or financial officials now! Most people just want to know what Martin Lewis is saying.
- Then a major retail bank requires emergency overnight funding from the Treasury. Reports of its insolvency are the final nail in the coffin. Other banks are forced to follow as citizens rush to withdraw.
- From first reports of the thefts to widespread insolvency, catastrophic financial collapse is realised in 72 hours.

SCENARIO 10
FOREIGN TECH SUPERPOWER LEVERAGES UK OVERRELIANCE ON ITS TECHNOLOGY (PRIMARILY AI) TO UNDERMINE SOVEREIGNTY

Summary: Over the next few years, sectors across UK society come to rely on American AI services (or services with a significant AI component) and underpinning infrastructure. The US leverages the UK's reliance on its technology to force the UK to adopt US-dictated policies and laws, threatening to cut off access with catastrophic effect for business and economy. Meanwhile US control of popular platforms and messaging apps provides an avenue to tracking and nudging UK citizen behavior. The UK becomes a de facto US vassal state. In an alternative version of this scenario, the UK comes to rely on AI from China rather than the US; but the outcome is similar.

- In the next few years, the UK comes to rely on American AI-based or AI-related technologies throughout its government, economy, and society:
- US-controlled AI is embedded into critical national infrastructure, such as the water and energy systems.
- US-controlled data centres form the vast majority of the UK's data centre capacity.
- US-controlled cloud services are relied on throughout the UK government.
- US-controlled AI services are relied on throughout the UK government for information, productivity tasks and decision-making.
- US-controlled AI corporate services are relied on throughout the UK economy for information, productivity tasks, and decision-making.
- US-controlled AI consumer services are depended on by the UK public for information, socialising, and emotional support.
- The US government decides to exercise influence on the UK to assert its economic dominance and to force the UK to adopt policies that align with its ideological goals, such as restricting the rights of racial minorities or banning 'woke' ideas.
- To achieve this goal, the US either directly requisitions US companies as state assets or controls them covertly through backdoor channels. In either case, the US uses this control to gather sensitive intelligence on the UK government's decisions, control the UK's public's information access, shape public opinion, and/or manipulate the government's AI-assisted decision-making in the US' favour. The US also takes advantage of its information access to gather compromising information on senior UK politicians and blackmail them.
- The US then threatens to cut off the UK's access to the data and AI that it has come to depend on.
- If the UK refuses to acquiesce to American demands, the US retaliates by cutting off its access to all US-controlled AI and data throughout the UK's tech stack. This leads to a failure of communication systems; rolling energy and internet blackouts;

the loss of government data and records; and the sudden withdrawal of services that the public had become dependent on for emotional support and advice. The UK economy collapses, government decision-making is paralysed, and public disorder breaks out.

- If the UK accepts the US' demands, the UK is forced to adopt US-dicted policies and laws – essentially making it into a vassal state.

2.3 SYSTEMS MAPPING (WORKSHOP 1B)

In this final stage of the workshop, each or the scenario groups engaged in '**Systems Mapping**' exercise to visually map the interactions and interdependencies within and between the social epistemic systems influencing a scenario. '**Social Epistemic Systems**' refers to the collection of processes, actors, and factors (e.g. technological, social, political) that influence how information is produced, distributed, modified, consumed, and evaluated within a society. Each scenario involves multiple overlapping social epistemic systems. These may include, for example, a 'malicious actor' system involving bad actors and the various technological or social capacities that facilitate their actions; an "official authority" system (e.g. government) involving the various factors and capacities that enable or inhibit the authorities' activities in countering the malicious actors; and a 'public societal system' involving all the factors influencing how the public is able to access, consume, and process information to inform belief formation, decision-making, and action.

Systems mapping is a useful tool for stepping back and holistically visualising social epistemic systems, which in turn helps developers to identify points of epistemic vulnerability (some of which only emerge where actor systems intersect) and potential leverage points for intervention to strengthen epistemic security. The holistic view is important because complex epistemic systems most often suffer from multiple overlapping epistemic threats and vulnerabilities such that a solution to one might exacerbate another. For example, 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 - potentially lending the group greater credibility - or backfire if it is seen as government overreach infringing on freedom of expression.

Of course given the complexity of real world social epistemic systems it is almost always impossible to comprehensively map any epistemic system with all relevant internal and external influences thereon. Our goal in taking a systems-based appraisal is therefore not to produce an exhaustive list of vulnerabilities and threats, but to visualise the complex interactions at play and to identify most likely threat mechanisms and prioritise areas of intervention accordingly.

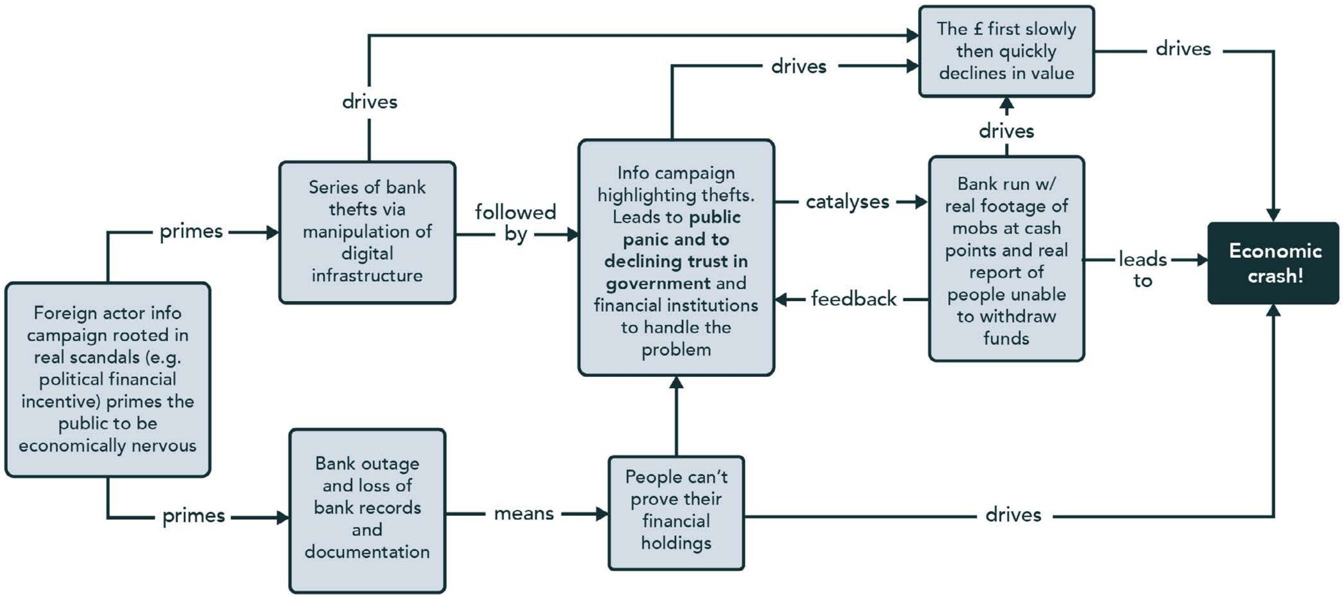
We attend interventions in Section 3. The first workshop produced the initial scenario system maps that would be the scaffold for intervention analysis in the second workshop.

Initial Scenario Mapping:

At the start of this systems mapping exercise, participants were encouraged to shuffle to a new scenario group if they felt their specific expertise would apply. This was to prevent group think by bringing in a fresh set of perspectives for building on and translating the initial PESTLE exercise into the system's map. At least two members of each original group were retained for continuity.

Once settled, the group moderator sketched out a flow chart of the base scenario mechanism as articulated in updated scenario narrative (see figure 2 - an example from scenario 6). This provided a scaffold for the group to build on.

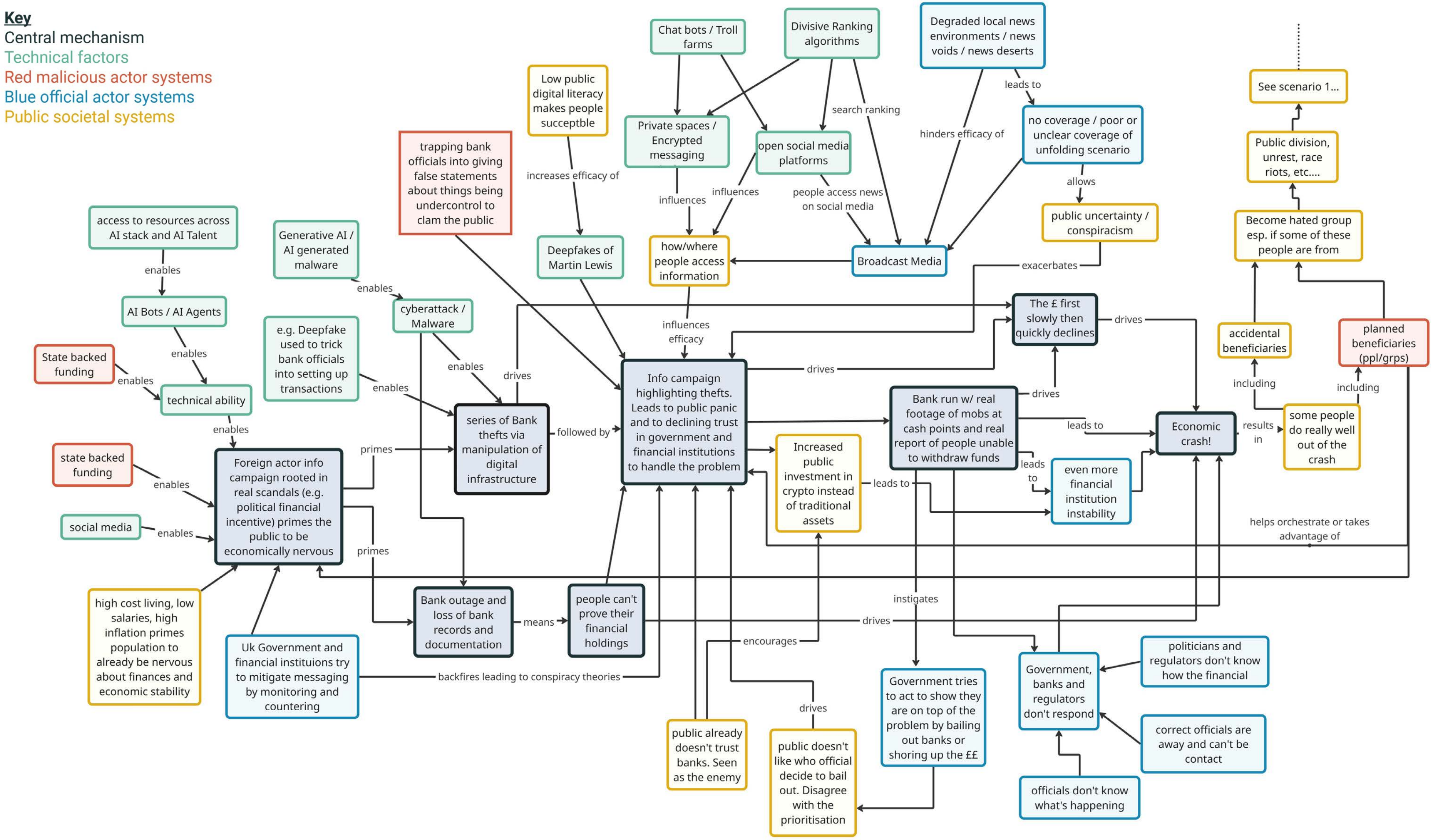
FIGURE 2
BASE NARRATIVE FLOW CHART FOR SCENARIO 6 "BANK RUN"



Foreign actor info campaign rooted in real scandals (e.g. political financial incentive) primes the public to be economically nervous
primes
Series of bank thefts via manipulation of digital infrastructure
drives
followed by
Info campaign highlighting thefts. Leads to **public panic and to declining trust in government** and financial institutions to handle the problem
drives
The £ first slowly then quickly declines in value
drives
catalyses
feedback
Bank run w/ real footage of mobs at cash points and real report of people unable to withdraw funds
drives
leads to
Economic crash!
primes
Bank outage and loss of bank records and documentation
means
People can't prove their financial holdings
drives

Via moderated discussion, the groups then began mapping on influencing factors, starting with those already articulated in the updated scenario narrative to help establish a sense for the causal mapping process and to build momentum in the discussion. The groups continued by mapping additional influencing factors identified in the PESTLE as well generating completely new factors, actors, influence pathways, and narrative dynamics that emerged as the maps fleshed out. (See figure 3 building on figure 2). The complete maps for all four scenarios are provided in Appendix A.

FIGURE 3
SYSTEM MAP OUTPUT FROM WORKSHOP 1 FOR SCENARIO 6 “BANK RUN”


Key
Central mechanism
Technical factors
Red malicious actor systems
Blue official actor systems
Public societal systems
Chat bots / Troll farms
Divisive Ranking algorithms
Degraded local news environments / news voids / news deserts
leads to
See scenario 1...
Low public digital literacy makes people succeptble
Private spaces / Encrypted messaging
open social media platforms
search ranking
hinders efficacy of
no coverage / poor or unclear coverage of unfolding scenario
Public division, unrest, race riots, etc....
trapping bank officials into giving false statements about things being undercontrol to clam the public
increases efficacy of
influences
influences
people access news on social media
allows
public uncertainty / conspiracism
Become hated group esp. if some of these people are from
access to resources across AI stack and AI Talent
Generative AI / AI generated malware
Deepfakes of Martin Lewis
how/where people access information
Broadcast Media
exacerbates
enables
enables
cyberattack / Malware
influences efficacy
The £ first slowly then quickly declines
drives
AI Bots / AI Agents
e.g. Deepfake used to trick bank officials into setting up transactions
accidental beneficiaries
planned beneficiaries (ppl/grps)
State backed funding
enables
enables
enables
drives
Info campaign highlighting thefts. Leads to public panic and to declining trust in government and financial institutions to handle the problem
drives
Bank run w/ real footage of mobs at cash points and real report of people unable to withdraw funds
drives
including
including
enables
technical ability
enables
series of Bank thefts via manipulation of digital infrastructure
followed by
leads to
Economic crash!
results in
some people do really well out of the crash
enables
state backed funding
enables
Foreign actor info campaign rooted in real scandals (e.g. political financial incentive) primes the public to be economically nervous
primes
Increased public investment in crypto instead of traditional assets
leads to
leads to
even more financial institution instability
social media
enables
primes
helps orchestrate or takes advantage of
high cost living, low salaries, high inflation primes population to already be nervous about finances and economic stability
Bank outage and loss of bank records and documentation
means
people can't prove their financial holdings
instigates
drives
encourages
Uk Government and financial instituions try to mitigate messaging by monitoring and countering
backfires leading to conspiracy theories
Government, banks and regulators don't respond
politicians and regulators don't know how the financial
drives
Government tries to act to show they are on top of the problem by bailing out banks or shoring up the ££
correct officials are away and can't be contact
public already doesn't trust banks. Seen as the enemy
public doesn't like who official decide to bail out. Disagree with the prioritisation
officials don't know what's happening

Here ends workshop 1. For readers interested in conducting follow-up projects that iterate on this methodology, we articulate further learnings from the process in Appendix C which may be helpful.

3. INTERVENTION IDENTIFICATION & ANALYSIS

METHODOLOGY: WORKSHOP 2

In the first workshop expert participants worked to develop a holistic, systems-level understanding of four distinct crisis scenarios exacerbated by epistemic threats. In the second workshop experts turned their attention to interventions.

The aims of the second workshop were twofold:

- Use the system maps developed in the first workshop to identify key leverage points for intervention to reduce the risk of such a crisis unfolding (Section 3.1)
- Red-team the most promising interventions to identify potentially harmful downstream consequences and to mitigate those potentialities where possible (Section 3.2)

3.1 INTERVENTION IDENTIFICATION (WORKSHOP 2A)

For the second full-day workshop the original group of experts reconvened along with a few new faces brought in for additional subject area expertise.

The second workshop commenced with four moderators heading up each of the crisis scenario subgroups. Expert participants were again asked to self-sort with the exception of new subject area experts that were brought in to help attend to specific scenarios (e.g. a financial sector professional sent to the 'Bank Run' scenario).

The intervention identification processes took place in three phases:

Phase 1: Systems map review

Each group was provided an A0 printout of the scenario systems map produced in the first workshop. Experts spent the first 15 minutes reviewing the scenario. This was an opportunity for experts to refamiliarise with the map and for new subject experts to help identify any influencing factors or system dynamics that may have been missed in the previous workshop.

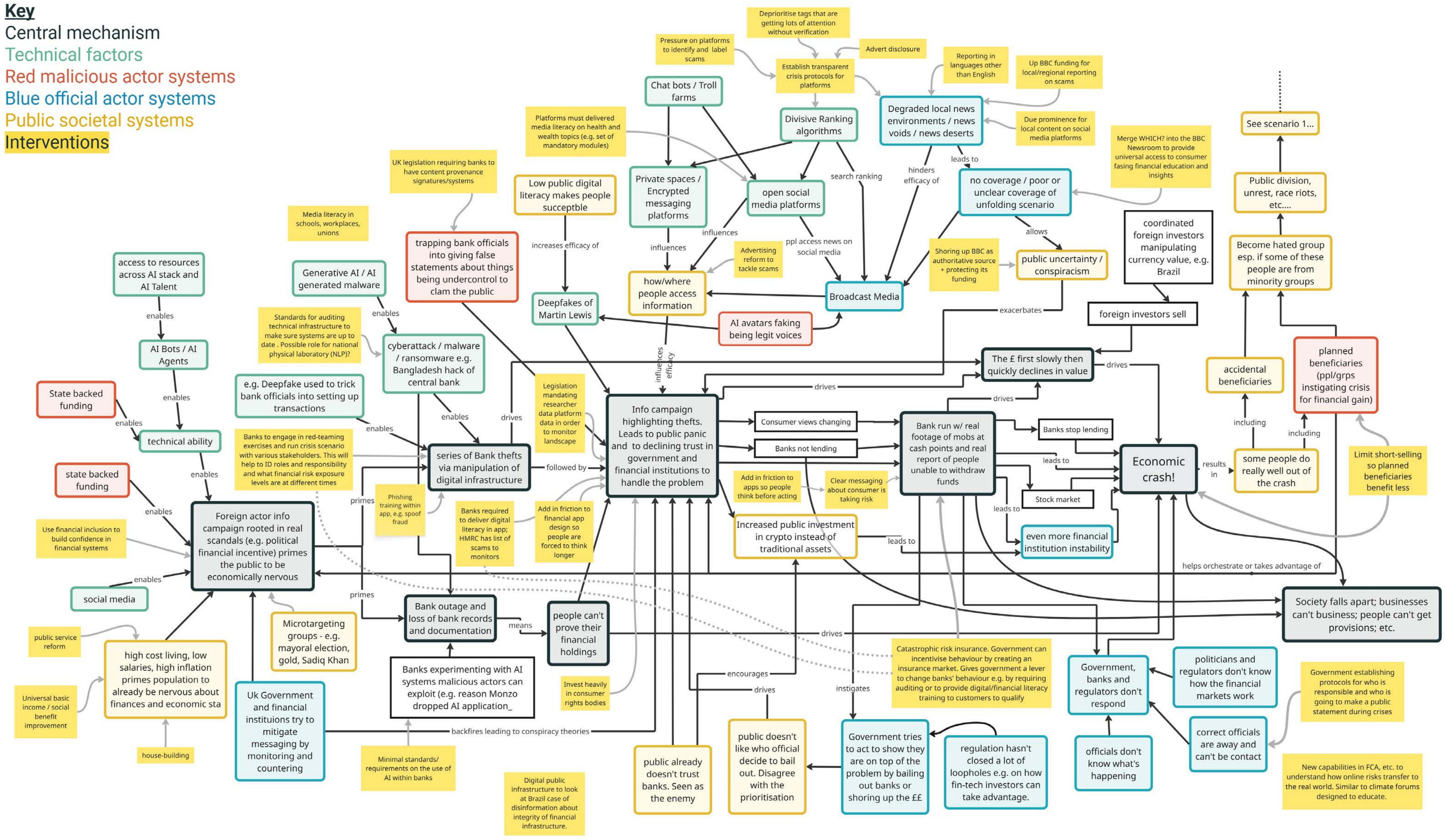
While it is impossible to create a perfectly complete systems map (a map representing all possible factors, actors, and interactions influencing a scenario), the process we employed yields a holistically considered approximation with a high likelihood of depicting the most critical drivers.

Phase 2: Intervention ideation and mapping

The mapped pathways provide the scaffold for identifying intervention points. In this second phase participants worked systematically through the causal pathways, considering what kinds of interventions might be implemented on each to address epistemic vulnerabilities or threat vectors exacerbating the scenario. These intervention potentials were added to the systems map. (See figure 4, building on figure 3). The layered intervention maps for all four scenarios are provided in Appendix A.

Halfway through the intervention mapping process, experts were given the opportunity to switch groups to bring fresh perspective.

FIGURE 4
SCENARIO 6 SYSTEM'S MAP LAYERED WITH INTERVENTIONS



Key
Central mechanism
Technical factors
Red malicious actor systems
Blue official actor systems
Public societal systems
Interventions

access to resources across AI stack and AI Talent
enables
AI Bots / AI Agents
enables
State backed funding
enables
technical ability
enables
state backed funding
enables
Use financial inclusion to build confidence in financial systems
social media
enables
Foreign actor info campaign rooted in real scandals (e.g. political financial incentive) primes the public to be economically nervous
public service reform
high cost living, low salaries, high inflation primes population to already be nervous about finances and economic sta
Universal basic income / social benefit improvement
house-building
Uk Government and financial instituions try to mitigate messaging by monitoring and countering
Microtargeting groups - e.g. mayoral election, gold, Sadiq Khan
Media literacy in schools, workplaces, unions
Generative AI / AI generated malware
enables
Standards for auditing technical infrastructure to make sure systems are up to date . Possible role for national physical laboratory (NLP)?
cyberattack / malware / ransomware e.g. Bangladesh hack of central bank
e.g. Deepfake used to trick bank officials into setting up transactions
enables
Banks to engage in red-teaming exercises and run crisis scenario with various stakeholders. This will help to ID roles and responsibility and what financial risk exposure levels are at different times
primes
Phishing training within app, e.g. spoof fraud
enables
drives
series of Bank thefts via manipulation of digital infrastructure
followed by
primes
Bank outage and loss of bank records and documentation
means
people can't prove their financial holdings
Banks experimenting with AI systems malicious actors can exploit (e.g. reason Monzo dropped AI application_
Minimal standards/ requirements on the use of AI within banks
backfires leading to conspiracy theories
Digital public infrastructure to look at Brazil case of disinformation about integrity of financial infrastructure.
UK legislation requiring banks to have content provenance signatures/systems
trapping bank officials into giving false statements about things being undercontrol to clam the public
Legislation mandating researcher data platform data in order to monitor landscape
Banks required to deliver digital literacy in app; HMRC has list of scams to monitors
Add in friction to financial app design so people are forced to think longer
Invest heavily in consumer rights bodies
Platforms must delivered media literacy on health and wealth topics (e.g. set of mandatory modules)
Low public digital literacy makes people susceptble
increases efficacy of
Deepfakes of Martin Lewis
Pressure on platforms to identify and label scams
Chat bots / Troll farms
Private spaces / Encrypted messaging platforms
influences
influences
how/where people access information
influences efficacy
Advertising reform to tackle scams
Deprioritise tags that are getting lots of attention without verification
Advert disclosure
Establish transparent crisis protocols for platforms
Divisive Ranking algorithms
open social media platforms
ppl access news on social media
search ranking
Broadcast Media
AI avatars faking being legit voices
Reporting in languages other than English
Up BBC funding for local/regional reporting on scams
Degraded local news environments / news voids / news deserts
Due prominence for local content on social media platforms
hinders efficacy of
leads to
no coverage / poor or unclear coverage of unfolding scenario
Shoring up BBC as authoritative source + protecting its funding
allows
public uncertainty / conspiracism
exacerbates
Merge WHICH? into the BBC Newsroom to provide universal access to consumer fasing financial education and insights
coordinated foreign investors manipulating currency value, e.g. Brazil
foreign investors sell
The £ first slowly then quickly declines in value
drives
drives
drives
Info campaign highlighting thefts. Leads to public panic and to declining trust in government and financial institutions to handle the problem
Consumer views changing
Banks not lending
Add in friction to apps so people think before acting
Clear messaging about consumer is taking risk
Increased public investment in crypto instead of traditional assets
leads to
Bank run w/ real footage of mobs at cash points and real report of people unable to withdraw funds
Banks stop lending
leads to
Stock market
leads to
even more financial institution instability
Economic crash!
results in
public already doesn't trust banks. Seen as the enemy
encourages
drives
public doesn't like who official decide to bail out. Disagree with the prioritisation
instigates
Catastrophic risk insurance. Government can incentivise behaviour by creating an insurance market. Gives government a lever to change banks' behaviour e.g. by requiring auditing or to provide digital/financial literacy training to customers to qualify
Government tries to act to show they are on top of the problem by bailing out banks or shoring up the ££
regulation hasn't closed a lot of loopholes e.g. on how fin-tech investors can take advantage.
Government, banks and regulators don't respond
officials don't know what's happening
politicians and regulators don't know how the financial markets work
correct officials are away and can't be contact
Government establishing protocols for who is responsible and who is going to make a public statement during crises
New capabilities in FCA, etc. to understand how online risks transfer to the real world. Similar to climate forums designed to educate.
helps orchestrate or takes advantage of
Society falls apart; businesses can't business; people can't get provisions; etc.
See scenario 1...
Public division, unrest, race riots, etc....
Become hated group esp. if some of these people are from minority groups
accidental beneficiaries
planned beneficiaries (ppl/grps instigating crisis for financial gain)
including
including
some people do really well out of the crash
Limit short-selling so planned beneficiaries benefit less

Phase 3: Intervention prioritisation and refinement

Following the initial intervention brainstorm, the expert working groups were then asked to narrow down their list of possible interventions to a list of 5-10 priority interventions they felt were most important - those that should receive further analysis with an eye towards implementation to mitigate the risk of such a crisis scenario coming to fruition.

The purpose of prioritising interventions was (a) to narrow down a tractable number of interventions to be brought forward to the subsequent red-teaming exercise, and (b) to quickly identify which interventions were expected to be most plausible and impactful. These interventions could later be compared with those identified in the other crisis scenario subgroups. Common prioritised interventions across subgroups would indicate the most likely targets for furthest reaching impact.

The expert working groups were provided loose guidelines for intervention prioritisation as articulated in Table 4.

TABLE 4
INTERVENTION PRIORITISATION GUIDELINES

Instructions: Select up to 10 interventions. Similar interventions, or interventions that are dependent on another for effective implementation can be bundled into one. Some factors you may wish to consider in your prioritisation include:

Actors
Can we identify specific actors who clearly ought to be responsible for implementing the interventions? How likely would the actor be to intervene? Are there any large barriers to doing so? (if so, this does not necessarily mean the intervention should be deprioritised, but it may need to be bundled with an upstream intervention to help increase the likelihood of the responsible actor intervening)
Expected impact
If successful, how large do you expect the impact will be on mitigating the scenario rollout or the negative impacts of the scenario? Is the mechanism that the intervention intervenes upon a core/powerful driver of the scenario unfolding?
Projected risk
Is it an intervention that seems potentially no/low risks for a notable impact? Is the intervention a potentially high impact but high risk intervention that needs to be carefully scrutinised with risk benefit analysis before implementation? What are some potential unintended consequences?

After intervention prioritisation, experts were tasked with refining the selected interventions where needed. The aim in this step was to clarify the intervention mechanisms where necessary by articulating the target actors (who is meant to enact the intervention?), as well as the specific activity (e.g. policy implementation, risk assessment, technological development) the intervention is meant to involve.

3.2 RED-TEAMING (WORKSHOP 2B)

In the second half of the workshop, the expert working groups then turned to red-team the prioritised interventions. '**Red- teaming**' broadly describes the process by which experts search for failure modes in a plan or system in order to challenge assumptions and make solutions more robust. Together, scenario mapping and red-teaming are routinely employed in national security settings to help agencies test defenses, improve decision-making, and anticipate threats by taking an opposing stance to rigorously challenge it.[20] Systems oriented red-teaming is particularly important in the context of ecosystem security for detecting potential unforeseen consequences of interventions arising from the complex interactions between overlapping epistemic systems. If an intervention is considered in isolation, potential negative consequences or shortcomings in the intervention may not be immediately clear, but when considered against the richer systems backdrop the issue is more easily illuminated.

A robust red-teaming exercise for our workshop would ideally involve splitting the expert scenario groups into 'red-teams' tasked with bypassing or otherwise thwarting interventions and 'blue-teams' tasked with ensuring the interventions' effective implementation. The idea is that as the game plays out between red-team and blue-team, red-team activity helps illustrate the scope of human ingenuity that would come into play in the real world thwarting interventions (highlighting likely mechanisms of intervention failure or backfire). Meanwhile responding blue-team activity workshops how interventions can be strengthened against these pitfalls as the scenario unfolds.

Limitations from time constraint:

However, coming to the end of the second workshop, we lacked the time necessary to enact the kind of deep red-teaming exercise described above. We decided instead to run a quick "light touch red-teaming" session using a set of common red-teaming tools to help to scan for potential pitfalls in each group's prioritised scenarios.

Cutting down on the red-teaming segment of the workshops has meant that we are unable to offer as robust a set of targeted recommendations as we would like for mitigating the likelihood and severity of four specific crisis scenarios we workshopped. In the scenario 'close-ups' provided in Appendix A you will see the scenario specific recommendations are labelled "Preliminary priority interventions". While our expert working groups expect these are key intervention pathways that ought to be enacted, each intervention should receive deeper analysis for unintended risks and ideal implementation mechanisms.

Nonetheless, a core goal of this project has, from the start, been to identify "cross-cutting areas of intervention" - closely related intervention recommendations (e.g. recommendations relating to content provenance protocols and requirements) that are common across a variety of crisis scenarios. These are the intervention areas which are likely to have the furthest reaching impact for mitigating the likelihood and severity of a diverse variety of potential crises scenarios. For this purpose of identifying these cross-cutting intervention areas, the systemic intervention mapping, prioritisation, and light touch red-teaming measures implemented are more than adequate. Results of the cross-cutting intervention analysis are presented in section 4.

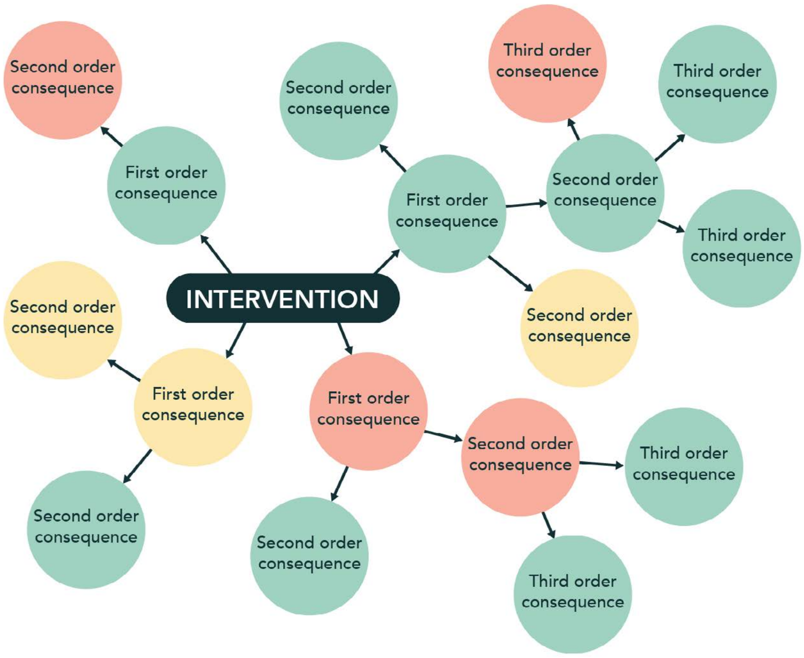
Red-teaming tools employed:

For the light-touch red-teaming exercise, the scenario groups were encouraged to apply either '**Futures Wheels**' or '**Pre-mortems**' to scan for pitfalls or unintended second-, third-, or higher-order consequences of intervention.

[20] UK Ministry of Defence (2021). Red Teaming Handbook.https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1027158/20210625-Red_Teaming_Handbook.pdf

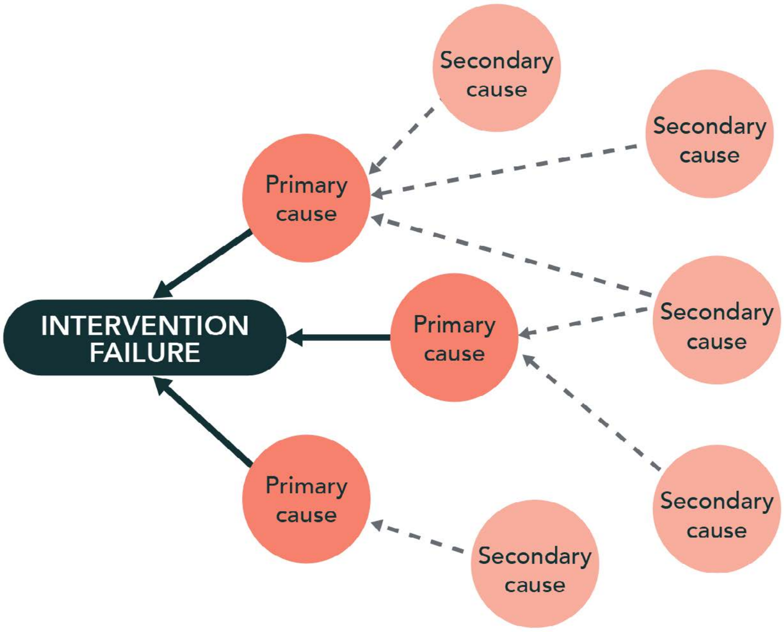
In scanning for these potential risk pathways the expert groups also had an opportunity to identify if any additional 'sub-recommendations' could be bundled in with the title recommendation to mitigate against the undesirable consequence. For example, watermarking requirements for artificially generated media might be implemented to help citizens distinguish between real and artificially generated content as artificially generated content becomes increasingly visually indistinguishable from real photographs and videos. However watermarks are currently trivial to remove. Consequently, if citizens imply from the new requirement that watermarks will indicate artificial content, they may become more vulnerable to misplacing trust in fake content from which watermarks have been removed thus increasing the potential influence of that content. To mitigate this potentiality, watermarking requirements might be bundled with improvements to media literacy modules offered in schools and by employers or by requiring online platforms to present users with media literacy notices before use.

FIGURE 5
RED-TEAMING TOOLS



A futures wheel is a red-teaming technique in which one starts with an intervention and works outwards to identify possible first-, second- and higher-order consequences of the intervention. First the initial consequences of an intervention (first-order impacts) are traced out, and knock-on impacts (higher-order intervention) are built out from there. Both high probability consequences and low-probability but high-impact consequences are included and identified as positive (green), negative (red), or neutral or uncertain (yellow) outcomes.

A pre-mortem starts by assuming that an intervention has failed. The practitioner then works backwards to construct causal pathways that could lead to such a failure. With the possible pathways 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.



4. FINDINGS
CROSS-CUTTING INTERVENTION AREAS

For each scenario expert working groups mapped, prioritised, and lightly red-teamed a set of interventions they expect would have the greatest impact minimising the likelihood and/or severity of one such crisis scenario occurring. These preliminary scenario-specific recommendations are listed under their respective scenario briefs in Appendix A. We do not expound on them here for two reasons.

First, as noted at the end of section 2, our expert working group was composed heavily of leading expertise on topics pertaining to information ecosystems dynamics – including expertise on news environments, online safety, epistemic threats from emerging tech capabilities and governance, and foreign influence. This was in order to consider an initially unknown variety of potential crisis scenarios, but it meant the team was not tailored with context-specific stakeholders and subject area expertise to robustly analyse the more specialised intervention. For example, to mount the most thorough and robust mapping and interventions red-teaming exercise for Scenario 6 "Bank Run and Economic Crash", the expert group would also need to have included representatives from HMT, PRA, and the Bank of England.

Second, time restrictions on the second day limited us to a light-touch analysis of unintended consequences. We are confident that the scenario-specific interventions identified in Appendix A illustrate important areas or angles for intervention for each scenario, but would encourage further analysis of each scenario and its interventions by a tailored group of experts and stakeholders.

Despite these limitations, the expert groups were well equipped to deliver on the project's ultimate objective which was to analyse a diversity of potential crisis scenarios and generate a breadth of intervention proposals in order to later step back and observe the key '**cross-cutting intervention areas**' that are common across the board. These are the most efficient intervention areas in the sense that, if pursued and effectively implemented, they are most likely to have the furthest reaching impact in mitigating the likelihood and severity of a wide variety of potential crisis scenarios.

We identified seven such **cross-cutting intervention areas** in the workshop outputs. In no particular order they are:

1. Content and data provenance, digital signatures, and watermarking (section 4.1)
2. Media and information literacy (section 4.2)
3. Digital infrastructure & cybersecurity (section 4.3)
4. Crisis protocols for platforms (section 4.4)
5. Government and regulator crisis preparedness (section 4.5)
6. AI risk management (section 4.6)
7. Rebuilding local and regional news ecosystems (section 4.7)

In what follows we provide a brief description of each cross-cutting intervention area, propose specific, tractable interventions that might be pursued under each heading, and provide examples of current initiatives in the UK or internationally where similar such interventions have been implemented or are under consideration.

4.1 CONTENT AND DATA PROVENANCE, DIGITAL SIGNATURES, AND WATERMARKING

The increasing realism and sophistication of AI-generated content is making it more and more difficult for everyday users and even the experts to identify it on their devices.[21] During a crisis, the ability of citizens to identify official, decision relevant content is essential for a coordinated response. However, hyperrealistic fake content spread by malicious actors or unwitting blunders of mistakenly identifying AI-generated material as authentic muddies the waters, sowing confusion and discord. Well-placed fake content could also trigger crises, inviting violence by framing minority groups for crimes or catching out political leaders in fake lies.[22]

Alongside adults, youth are also struggling in this context – a third of UK students in a recent survey found it difficult to spot AI-generated people in videos.[23] As the lines between fact and fiction become ever blurrier, and users start to doubt everything they see online, threats to epistemic security only grow worse.[24] Political candidates can dismiss potentially credible allegations of corruption as 'deepfake smears' by opponents through the phenomenon known as the 'Liar's Dividend', while justice may be undermined when legal evidence is thrown out due to vague possibilities that it could be manipulated by AI tools.[25] These types of threats centred around the reliability of content were particularly prominent in our "AI-driven breakdown of legal system", "bank run and economic crash", and "xenophobic violence" scenarios – where threat actors were able to compromise evidential records, sow doubt over the truth, and amplify hatred by flooding the information zone with a deluge of indistinguishable real and fake content. Yet even in this respect, we are already seeing record levels of AI-generated disinformation polluting our online ecosystems, reflecting that this is very much a 'here and now' problem that cannot be ignored.[26]

[21] Horvitz, E. (2024). Better Together: Joining Forces on Digital Media Provenance. https://erichorvitz.com/better_together_digital_media_provenance.htm

[22] Spring, M. (2024). Did social media fan the flames of riot in Southport?. BBC News. https://www.bbc.co.uk/news/articles/cd1e8d7llg9o; Spring, M. (2024). Sadiq Khan says fake AI audio of him nearly led to serious disorder. BBC News. https://www.bbc.co.uk/news/uk-68146053.

[23] White, M. (2025). Pupils struggle to spot untrue AI content - report. https://www.bbc.co.uk/news/articles/c891lwye5d8o

[24] New Statesman (2025). In the age of AI, how do we trust what we see online? https://www.newstatesman.com/spotlight/2025/02/in-the-age-of-ai-how-do-we-trust-what-we-see-online

[25] Christopher, N. (2023). An Indian politician says scandalous audio clips are AI deepfakes. We had them tested. Rest of World. https://restofworld.org/2023/indian-politician-leaked-audio-ai-deepfake/; Murphy, H. (2024). The trouble with deepfakes — liar's dividend. Financial Times. https://www.ft.com/content/7f22ce59-1c6c-4d84-bca8-dc539992e286; Runyon, N. (2025). Deepfakes on trial: How judges are navigating AI evidence authentication. Thomson Reuters. https://www.thomsonreuters.com/en-us/posts/ai-in-courts/deepfakes-evidence-authentication/

[26] EDMO (2025). AI-generated disinformation hits a new record in October as information integrity crumbles. https://edmo.eu/wp-content/uploads/2025/11/EDMO-Horizontal-53-1.pdf.

In our current digital age, a capacity to verify content source and authenticity is essential to crisis resilience in democratic societies. Aside from avoiding the dangers of relying on users going with their 'gut instinct' in trying to navigate what they see on online platforms, it is also unfair and unreasonable to place the primary burden on them to do so.[27] Content provenance tools which cryptographically secure metadata (e.g. the who, what, where and how) to digital assets, provide users with the ability to know how a piece of content was created.[28] Consistent employment of these tools would have helped significantly to counteract many of the AI-enabled threats which emerged across our scenarios.

Crucially however, content provenance tools are not silver bullets by themselves;[29] they also require widespread cooperation between different sectors and an awareness among users as to what this information reveals or how any secured metadata included as part of these features should be interpreted.[30] As such, it is vital that they are paired with pragmatic media and information literacy initiatives (4.2), as well as measures by media platforms to help moderate device and fake content (4.4), efforts to revitalise trustworthy news ecosystems (4.7)

Recommendations

There are already some provenance tools which have seen adoption among multiple different sectors in society. For example, the Coalition for Content Provenance and Authenticity's (C2PA) open standard is utilised by over 5,000 members of the C2PA community, ranging from camera and phone providers to AI developers and social media platforms. [31]

However, there are still obstacles which prevent provenance tools like C2PA from being more effective in combatting threats to epistemic security. Firstly, there is inconsistent and uneven uptake of these standards across the media ecosystem, with social media platforms in particular often stripping the essential metadata that can then provide an unbroken chain of information on the 'life cycle' of a digital asset.[32] Even with certain platforms that do actively adopt C2PA, many still fail to surface the standard on posts or adequately ensure AI-generated content contains clear labels in this respect.[33] Additionally, research has shown that some users find it difficult to make sense of labels that are shown to them with standards like C2PA, as well as what they are designed to represent in terms of information.[34] This risks leading to content containing these types of signals being misinterpreted and undermining the very purpose of provenance tools.[35]

[27] Department of the Prime Minister and Cabinet (2024). Report of the Statutory Review of the Online Safety Act 2021. https://www.aph.gov.au/Parliamentary_Business/Tabled_Documents/9184.

[28] Content Credentials. Verify Media Authenticity. https://contentcredentials.org/.

[29] Horvitz, E. (2022). A Milestone Reached. https://erichorvitz.com/A_Milestone_Reached_Content_Provenance.htm

[30] CNTI (2024). Watermarks are Just One of Many Tools Needed for Effective Use of AI in News. https://cnti.org/article/watermarks-are-just-one-of-many-tools-needed-for-effective-use-of-ai-in-news/; Ofcom (2025). Deepfake Defences 2 - The Attribution Toolkit. https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/deepfake-defences-2/deepfake-defences-2---the-attribution-toolkit.pdf?v=399908

[31] Kaye, K. & Dixon, P. (2025). Privacy, Identity and Trust in C2PA: A Technical Review and Analysis of the C2PA Digital Media Provenance Framework. https://worldprivacyforum.org/media/documents/c2pa_report.pdf; Content Authenticity Initiative. Our members. https://contentauthenticity.org/our-members; Horvitz, E. (2021). A promising step forward on disinformation. Microsoft on the Issues. https://blogs.microsoft.com/on-the-issues/2021/02/22/deepfakes-disinformation-c2pa-origin-cai/

[32] Schaul, K. (2025). We uploaded a fake video to 8 social apps. Only one told users it wasn't real. The Washington Post. https://www.washingtonpost.com/technology/2025/10/22/ai-deepfake-sora-platforms-c2pa/

[33] Mantzarlis, A. & Dutta, N. (2025). Tech platforms promised to label AI content. They're not delivering. Indicator. https://indicator.media/p/tech-platforms-fail-to-label-ai-content-c2pa-metadata

[34] Ofcom (2025). Deepfake Defences 2 - The Attribution Toolkit. https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/deepfake-defences-2/deepfake-defences-2---the-attribution-toolkit.pdf?v=399908; Google (2024). Determining trustworthiness and provenance through context. https://static.googleusercontent.com/media/publicpolicy.google/en//resources/determining_trustworthiness_en.pdf.

[35] CNTI (2024). Watermarks are Just One of Many Tools Needed for Effective Use of AI in News. https://cnti.org/article/watermarks-are-just-one-of-many-tools-needed-for-effective-use-of-ai-in-news/

Given all of these challenges, we recommend the following efforts to help enhance the effectiveness of content provenance tools in the UK:

1. **Exploring new policies and regulatory measures which would incentivise the use of cryptographically secure provenance systems for *all* content types across the entire tech stack.** Current inconsistencies with the voluntary implementation of standards such as C2PA across hardware and software providers is inhibiting the ability for users to trace where their content has come from or how it was created. The UK Government should thus consider a combination of: 1) new policies which incentivise hardware and software companies to adopt these tools – such as through financial loans, tax breaks or narrative framings tied to the economic benefits of greater integration; and 2) new regulation which establishes penalties for the deceptive removal, alteration, or relabeling of provenance information by different stakeholders.[36]

2. **Any new legislation on content provenance tools must minimise the inclusion of personal information within embedded metadata to protect user privacy.** Given that there may be sensitive personal information disclosed when provenance tools are used (e.g. through metadata), individuals should be able to retain control over the information captured.[37] This could include focusing captured provenance on how a digital asset was created or edited, with additional opt-in schemes for details on who created, edited or published it.[38] Indeed, C2PA has sought to directly address privacy concerns by informing implementers on how they should design interfaces to give content creators effective control over such personal information.[39]

3. **Supporting research into mapping out provenance adoption rates within the media ecosystem, as well as on provenance explainability.** On the one hand, understanding more about where provenance 'dark spots' exist across the content life cycle process would help in tailoring efforts to incentivise wider adoption among these areas. On the other hand, securing and surfacing the data (as noted above) is only part of the challenge. Even when this information is surfaced, more focus groups, surveys and behavioural studies are needed to identify how users interpret that data and if there is a threshold for the optimal amount of information to include in any visible labels on platforms.

4. **New educational initiatives should be established which raise awareness on how provenance tools work and what benefits they provide for users and organisations**. Given the struggles which users face in interpreting existing features when they are shown on platforms, these programmes could focus on helping users to know that provenance can assure trust but not truth, as well as what an authentic provenance label looks like.[40] This could form part of a wider media and digital literacy curriculum (see 4.2.) or be led by civil society organisations. Furthermore, efforts should be made to contextualise provenance tools as the equivalent of a digital 'nutrition' label – ensuring that users receive clear, accurate and consistent information on how their content was created and by whom.

5. **Finally, UK government departments and regulators should start publishing official communications and documentation with cryptographically secure provenance signals.** Not only would this help to lead by example in inspiring other sectors to see the benefits of

[36] Microsoft (2024). Protecting the Public from Abusive AI-Generated Content. https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Protecting-Against-Abusive-AI-Content-UK.pdf

[37] Kaye, K. & Dixon, P. (2025). Privacy, Identity and Trust in C2PA: A Technical Review and Analysis of the C2PA Digital Media Provenance Framework. https://worldprivacyforum.org/media/documents/c2pa_report.pdf

[38] WITNESS (2025). Tomorrow's Great Digital Divide: Content With or Without Provenance. https://blog.witness.org/2025/03/tomorrows-great-digital-divide/

[39] WITNESS (2025). Embedding Human Rights in Technical Standards: Insights from WITNESS's Participation in the C2PA. https://www.gen-ai.witness.org/wp-content/uploads/2025/06/Human-Rights-In-Standards.pdf

[40] The Royal Society (2022). Generative AI, content provenance and a public service internet. https://royalsociety.org/-/media/policy/projects/digital-content-provenance/Digital-content-provenance_workshop-note_.pdf

adopting these tools, but doing so will also help to mitigate against the risk of threat actors seeking to forge and mimic official government sources.

International context

Several countries have already introduced legislation designed to mandate disclosure labels on AI-generated or edited content for user transparency, which the UK could draw lessons from with any future regulation. For example, the EU's AI Act mandates that AI developers must include machine-readable watermarks when their models generate synthetic content, with fines on companies which fail to do so.[41] Spain has already adopted these measures at the national level, where a failure to comply with the proper labelling of AI-generated content could lead to tech platforms receiving fines of up to 35 million euros or 7% of their global annual turnover.[42] Similarly, Vietnam is introducing new legislation which comes into force in January 2026 that requires digital products generated by AI to carry identifiable markings to inform users or detection systems.[43]

Although the US has yet to implement federal-level legislation in this area, California state legislation SB 942 requires large generative AI providers to embed content provenance data (called 'latent disclosures') in all AI generated content and to provide AI detection tools for users to assess whether content was created or altered with generative AI.[44] A more recent amendment, AB 853, builds on SB 942 by placing additional duties on large social media platforms (e.g. X and Instagram) and 'capture device manufacturers' (e.g. cameras and smartphones) to help users determine content authenticity.[45] Platforms must detect and provide visible disclosures of watermark metadata in any content they distribute and must provide accessible interfaces to enable users to appraise content provenance information and to add provenance data to their own content. Capture device manufacturers must embed content provenance data in any content captured by a device by default.

China has sought to straddle a compromise here, focusing on both upstream and downstream parts of the ecosystem when it comes to new legal responsibilities around watermarking. In other words, a content platform in China must detect and label AI content if somehow it wasn't labelled upstream. Non-compliance can lead to fines or loss of license to operate, with the law now in effect since September 2025.[46]

When it comes to the UK Government's own adoption of provenance tools, departmental teams could look at how the former US Biden administration's AI Executive Order called for the development of guidance for labelling AI-generated content, as well as how federal agencies should use provenance tools to make it easier for users to know that the communications they receive from the government are authentic.[47] Some local government administrations in Japan have already started trialling the use of provenance features on publications related to disaster information for precisely these reasons, by ensuring that users can know whether information

[41] World Standards Cooperation (2025). Technical Report on AI and Multimedia Authenticity Standards. https://www.worldstandardscooperation.org/wp-content/uploads/2025/07/IEC-ISO-ITU-Technical_Report_on_AI_and_Multimedia_Authenticity_Standards.pdf; https://arxiv.org/pdf/2503.18156

[42] Reuters (2025). Spain to impose massive fines for not labelling AI-generated content. Reuters. https://www.reuters.com/technology/artificial-intelligence/spain-impose-massive-fines-not-labelling-ai-generated-content-2025-03-11/

[43] Vietnamnet (2025). Vietnam mandates AI-generated content labels starting in 2026. https://vietnamnet.vn/en/vietnam-mandates-ai-generated-content-labels-starting-in-2026-2424484.html

[44] California Senate Bill no.942 (2024). California AI Transparency Act. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB942

[45] California Assembly Bill no.853 (2025). California AI Transparency Act. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260AB853

[46] Yang, Z. (2024). China's Plan to Make AI Watermarks Happen. WIRED. https://www.wired.com/story/china-wants-to-make-ai-watermarks-happen/

[47] Ryan-Mosley, T. & Heikkilä, M. (2023). Three things to know about the White House's executive order on AI. MIT Technology Review. https://www.technologyreview.com/2023/10/30/1082678/three-things-to-know-about-the-white-houses-executive-order-on-ai/

circulating during these crisis events comes from government sources and has not been tampered with.[48]

4.2 MEDIA AND INFORMATION LITERACY

The capacity of a society's citizens to critically assess information quality and the trustworthiness of information sources is key to crisis resilience. Whether spread intentionally by malicious actors as in foreign influence campaigns or accidentally by unwitting epistemic blunders, false and misleading content can sow discord and distrust (priming communities for unrest and violence) and undermine trust in key crisis response authorities like public health officials and climate scientists (hobbling crisis response). Where citizens are more critical of the content they consume, more discerning of sources, and more responsible in how they share, the potential for public belief and behaviour manipulation is less.

Accordingly, Media and Information Literacy (MIL) interventions that aim to help people (often assumed to mean 'young people', but increasingly also understood to affect people of different ages in different ways)[49] develop the ability to access, evaluate, create, and share media in a critical and responsible way emerged as a common feature across all four analysed scenarios. They were most prominent in the "xenophobic violence" and "bank run and economic crash" scenarios in which crises were driven by information manipulation that shaped public sentiment and behaviour, triggering violence and instability. Similarly, in "AI-driven breakdown of the legal system" misinformation and distorted narratives undermined institutional trust in legal systems, and in the "overreliance on foreign tech" scenario, where information manipulation was identified as a likely consequence of public dependence on platforms controlled by a foreign actors, which increased crisis severity.

Implementing quality MIL programs will not, of course, make everyone a perfect 'digital Sherlock', always able to distinguish fact from fiction and trustworthy from untrustworthy information sources. MIL is only part of the puzzle, and must be paired with other interventions such as developing and implementing digital provenance tools on devices, in platforms and in apps (4.1), implementing platform crisis protocols (4.4), and strengthening news ecosystems (4.7) to help make information environments easier to navigate during a crisis .

Recommendations for MIL in the UK:

In the UK, access to media and information literacy education remains highly uneven, with those in lower socio-economic areas disproportionately underserved, and initiatives targeting adults largely neglected. The UK needs a whole society approach to MIL, one that does not focus only on building curriculum for children to keep them safe online and to teach critical thinking, but that also considers how the elderly and adults could also be susceptible to sharing misinformation or falling for scams or deepfakes. As new information generating and mediating technologies are continuously introduced so too must MIL agendas be updated to address any new challenges posed and be offered as continuous learning opportunities throughout life. This indicates a role not only for schools, but potentially for employers, and media platforms as well. More specific recommendations for MIL in the UK follow.

At the time of writing, the government had responded to the Curriculum and Assessment Review report and confirmed it would introduce compulsory citizenship education for all key stages to include media literacy as well as providing a new core enrichment entitlement, an oracy framework and the incorporation of different transient texts (e.g. social media posts) in

[48] Originator Profile. (2025). Tottori Prefecture to Become First Municipality to Implement Originator Profile, Decides to Participate in Demonstration Project. https://originator-profile.org/en-US/news/7vcm0i3oim/.

[49] For example, the Smidge Project targets middle-aged individuals facing similar online challenges. See: https://www.smidgeproject.eu/

the new English language GCSE.[50] These changes mark a significant step forward and would be complimented by the following.

Further MIL recommendations for training and protecting children:

1. **Ofsted should evaluate and report on the quality of teaching of media literacy in different subjects across the curriculum.** Media literacy should be incorporated into the inspection framework to include formal assessment of knowledge and skills. This will support preparation for England's, future participation in new PISA 2029 Innovative Domain on Media & Artificial Intelligence Literacy

2. **MIL should form a part of initial teacher training for all new teachers and continuous professional development for existing teachers.** This should be delivered through schools developing a media literacy policy that includes regular professional development, all supported by DfE. The design of a teacher training programme should be informed by previous pilot schemes and should leverage the expertise of subject associations and media literacy organisations

MIL recommendations targeting adults:

3. **The government should secure longer term funding for organisations delivering crucial media and information literacy services for adults in communities.** These organisations are currently hampered by an insecure and short-term funding environment. Longer-term funding might be secured, for example, by imposing a levy on platforms for independent media literacy initiatives.

4. **The government's digital inclusion action plan should build media literacy competencies by enabling more consistent support for initiatives delivered locally or through public services.** This might be done, for example, by leveraging platforms' funding. The government must ensure that in relying on trusted local stakeholders for delivery, it does not place further strain on frontline services and community groups without providing additional funding and support

5. **A core component of MIL for adults should focus on facilitating an understanding of dataflows in consumer products (e.g. social media platforms, chatbots, wearable tech, AI-enabled 'smart' devices) and the associated implications for privacy, security, and information risks.** The effort would not start from scratch. The UK GDPR already legally enshrines a "right to be informed".[51] These are requirements for individuals to be provided with information when personal data is collected and processed including: the specific categories of data collected, the purpose and legal basis for data processing, the name and contact details of the data controller, information about any data recipients or transfers, and information about individuals' rights to withdraw consent and lodge complaints to the ICO. Individuals are also granted the right not to be subject to automated decision-making based on personal data,[52] and safeguards are stipulated for exceptional cases requiring that individuals are notified of automated decisions to which they have been subject and provided an avenue to request review.[53]

   However the UK GDPR's stipulations provide broad 'good practice' guidelines that are often not well followed by data controllers and processors, and consumers lack understanding of data protection requirements and of their data protection rights, making informed decision

[50] UK Government (2025). Government response to the Curriculum and Assessment Review. https://assets.publishing.service.gov.uk/media/690b2a4a14b040dfe82922ea/Government_response_to_the_Curriculum_and_Assessment_Review.pdf
[51] UK GDPR. Articles 13-14. https://www.legislation.gov.uk/eur/2016/679/contents
[52] UK GDPR. Article 22. 'Automated individual decision-making, including profiling.' https://www.legislation.gov.uk/eur/2016/679/article/22
[53] Data Protection Act 2018. Article 14. 'Automated decision-making authorised by law: safeguards.' https://www.legislation.gov.uk/ukpga/2018/12/section/14/enacted

making about data sharing and granting data access to third parties difficult. As a result, this existing legislation is not well-leveraged to treat emerging epistemic risks.
A remedy would not need to involve further primary legislation. Rather the ICO could proactively research and issue deeper 'good practice' guidance for data protection measures and standards for data subject communication to advance the state of the art. The US Cybersecurity and Infrastructure Security Agency (CISA) has recently launched a similar initiative with its U.S. Cyber Trust Mark – an Internet of Things (IoT) labeling initiative to give consumers a way of understanding whether IoT products meet a cybersecurity baseline to help them make more well-informed decisions about the devices they purchase, like how nutrition labels help consumers make better informed decisions about a healthy diet.[54] A core principle of the initiative is that manufacturers embrace 'radical transparency'. CISA writes, "Radical transparency can take many forms—from security labeling, to a software provider publishing statistics on adoption of multi-factor authentication, to a technology manufacturer writing a blog post on their efforts to eliminate an entire class of vulnerability from their codebases—but all are important for a holistic understanding of our individual as well as our collective cybersecurity posture." The ICO could consider a similar approach to informing consumers of data protection rights, risks, and risk mitigations.

6. **The government should establish minimum and mandatory standards for technology platforms to implement and evaluate 'media literacy by design' methods in their platform designs so that Ofcom can be empowered to sufficiently hold platforms to account.** Media literacy by design aims to make digital environments themselves support media-literate behaviour by integrating cues, feedback, and design choices that promote reflection, trust, and informed decision-making. Various methods include embedding features such as transparency labels, contextual information, and gentle prompts or "frictions" that encourage users to pause and think critically before sharing or engaging with content.

Other noteworthy MIL initiatives

The UK does not need to start from scratch building its MIL agenda. Finland, Canada, Ireland, and Belgium, all ranked above the UK in the Open Society Institute's *Expanded Media Literacy Index 2023*, illustrate strong national approaches to fostering resilience through informed media engagement.[55] Finland, in particular, leads with a comprehensive national policy framework that makes media literacy a mandatory, cross-curricular subject throughout primary and secondary education.[56] Its curriculum emphasises students' ability to interpret and produce audiovisual content, while also embedding media literacy within broader "multiliteracy" and "ICT competence" standards. Teachers receive structured guidance and competency frameworks but retain flexibility in how they integrate media literacy within their subjects.

In other examples, Denmark has developed an academic platform for digital technology literacy as a core subject in schools,[57] and Norway leverages a network of public and civil-society organisations to build digital competence among seniors and vulnerable groups.[58] And within the UK, Wales's revised curriculum establishes *digital competence* as a mandatory, cross-curricular skill on par with literacy and numeracy, encompassing key media literacy outcomes. It is perhaps a model to be considered for expansion.

[54] CISA. (2023). Leading the Way with Radical Transparency. https://www.cisa.gov/news-events/news/leading-way-radical-transparency
[55] Open Society Institute (2023). The Media Literacy Index 2023: Measuring Vulnerability of Societies to Disinformation. https://osis.bg/wp-content/uploads/2023/06/MLI-report-in-English-22.06.pdf
[56] Finnish National Agency for Education. Multiliteracy and Media Literacy. https://www.oph.fi/en/education-and-qualifications/multiliteracy-and-media-literacy
[57] Villum Fonden (2023). New knowledge centre to ensure that all children and young people have a sound understanding of digital technology. https://villumfonden.dk/en/node/14946
[58] Høifødt, E. (2024). Norway's vision for Digital Inclusion: Leading the world in digitalization? EPALE Blog. https://epale.ec.europa.eu/en/blog/norways-vision-digital-inclusion-leading-world-digitalization

4.3 CYBERSECURITY & DIGITAL INFRASTRUCTURE

Recent cyberattacks across the UK have underscored both the impact that these threats have on society, as well as weaknesses in the UK's current cybersecurity infrastructure. A major hack on Marks and Spencer (M&S) led to profits being almost wiped out, and a similar intrusion into the Co-op's customer database resulted in 6.5 million consumers having their personal data stolen.[59] Meanwhile, the Jaguar-Land Rover attack in 2025 halted production and crippled supply chains, costing the UK economy £1.9 billion,[60] and a cyberattack on Heathrow airport in 2025 raised questions about how hackers could have affected flight safety systems and posed threats to passenger safety.[61] With new AI systems becoming even more accessible, elements of cyber operations will not only become more effective and efficient, but there will likely be an increase in the frequency and intensity of attacks in the near-future.[62] For example, AI agents are now assisting cybercriminals in executing attacks without substantial human intervention. In September 2025, Anthropic reported one of the first AI-orchestrated cyber espionage campaigns, where AI agents didn't just act as advisors for the human operatives, but actually executed the cyberattacks *themselves*.[63] These types of activities were a driving factor in our "bank run and economic crash" crisis scenario where small attacks on banks were used to stoke public panic, and in the "AI driven breakdown of legal system" scenario where public records were wiped out in an attack similar to the 2023 British Library ransomware attack.[64]

However, it is not just the impact of AI on cyberattacks which are causing concerns, but the damage that cyberattacks can also cause to AI systems. As explored further below, a recurring theme in our scenarios was how AI systems were being embedded in everything from critical infrastructure (e.g. the "Foreign tech superpower" crisis scenario) to essential services (e.g. the in "AI-driven breakdown of the legal system" crisis scenario). In such contexts, emerging offensive cyber activities – such as supply chain attacks, data poisoning and prompt injection techniques – are now posing risks to the security of AI models.[65] Not only could this prevent people from accessing critical services, but there would likely be significant economic consequences for businesses and harm done to those whose personal data may have been compromised via AI system vulnerabilities.[66]

Beyond these specific incidents, the UK is also in a difficult position when it comes to sovereign control over its digital infrastructure. The country holds just 3% of global computing power, with Chinese-based firms operating or maintaining large financial stakes in prominent UK battery sites, offshore wind farms and electricity distribution grids.[67] As shown when Shenzhen-based Chinese company Huawei played a central role in building the UK's 5G communications infrastructure, such reliance on nations with competing geopolitical interests can put the UK at risk of espionage threats and undermine national security.[68] Alongside China, there is significant dependency on US-based companies for the smooth operation of the UK's digital ecosystem. Major firms such as Google, Microsoft and AWS supply the backbone of the UK's cloud

[59] Masud, F. (2025). M&S profits almost wiped out after cyber hack hit sales. BBC News. https://www.bbc.co.uk/news/articles/c93x16zkl9do; Tidy, J. & Rahman-Jones, I. (2025). Co-op boss confirms all 6.5m members had data stolen. BBC News. https://www.bbc.co.uk/news/articles/cql0ple066po

[60] Tidy, J. (2025). JLR hack is costliest cyber attack in UK history, say analysts. BBC News. https://www.bbc.co.uk/news/articles/cy9pdld4y81o

[61] Araullo, K. (2025). Heathrow cyberattack caused massive delays, but its biggest impact is yet to come. Insurance Business UK. https://www.insurancebusinessmag.com/uk/news/cyber/heathrow-cyberattack-caused-massive-delays-but-its-biggest-impact-is-yet-to-come-550593.aspx

[62] UK National Cyber Security Centre. Impact of AI on cyber threat from now to 2027. https://www.ncsc.gov.uk/pdfs/report/impact-ai-cyber-threat-now-2027.pdf

[63] Anthropic (2025). Disrupting the first reported AI-orchestrated cyber espionage campaign. https://www.anthropic.com/news/disrupting-AI-espionage

[64] Scroxton, A. (2024). British Library cyber attack explained: What you need to know. Computer Weekly. https://www.computerweekly.com/feature/British-Library-cyber-attack-explained-What-you-need-to-know

[65] Powell, R., et al., (2024). "Towards Secure AI: How far can international standards take us?," CETaS Research Reports. https://cetas.turing.ac.uk/sites/default/files/2024-03/cetas_research_report_-_towards_secure_ai_0.pdf

[66] Ibid.

[67] McBride, K., et al., (2025). Sovereignty, Security, Scale: A UK Strategy for AI Infrastructure. Tony Blair Institute for Global Change. https://institute.global/insights/tech-and-digitalisation/sovereignty-security-scale-a-uk-strategy-for-ai-infrastructure; Chu, B. & Gilder, L. How much vital UK infrastructure does China own? BBC News. https://www.bbc.co.uk/news/articles/cn4w3y4pdkzo

[68] BBC News (2022). UK government bans new Chinese security cameras. https://www.bbc.co.uk/news/uk-politics-63749696

infrastructure, while Palantir provides advanced analytics platforms that support UK defence operations.[69] In many ways, these partnerships are essential for the UK to gain access to scale, resilience and cutting-edge innovation that the domestic ecosystem could not provide alone.[70]

However, in the long term, these same partnerships also introduce several constraints which could inhibit the UK's ambitions for digital and AI sovereignty. This includes: a lack of access to systems deployed in public services which could result in discrimination, surveillance or incompatibility with national regulations; challenges in enforcing jurisdictional rules on privacy and security that could lead to sensitive citizen data being shared outside of the UK; and risking vendor 'lock-in' that places reliance on a single provider and removes the possibility to switch to more effective competitors.[71] This project's "Foreign tech superpower" crisis scenario reflects how many of these risks could manifest, with foreign manipulation of citizen data to sway public sentiment and the threat of being cut off from critical information and communication infrastructure leading to the UK being subject to the whim of foreign political decisions that may undermine its own strategic interests.

However, it is not all doom and gloom. The UK not only has a strong IT infrastructure and robust workforce, but also a vibrant ecosystem of domestic firms, research institutions and start-ups capable of providing the foundations for British digital sovereignty. For example, established defence primes such as BAE Systems and QinetiQ provide advanced robotics and cyber solutions to the Ministry of Defence, alongside leading universities producing world-class talent and research on AI.[72] At the same time, the UK's cybersecurity sector is highly respected internationally, owing to government partnerships such as GCHQ CHECK and standards bodies such as CREST.[73] What is currently lacking however, is the ability for these domestic sources of innovation to meaningfully compete with more powerful foreign companies when it comes to contracts, as well as the UK being able to retain ownership of intellectual property, steer the direction of development and have the freedom of action in times of crisis.[74]

Recommendations

Although the UK Government has introduced a Cyber Security and Resilience Bill which seeks to strengthen defences and ensure that essential digital services are better protected from cyberattacks, amendments to this Bill and wider strategic changes to the way that the UK Government approaches cybersecurity issues are needed. These include:

1. **Expanding the existing Cyber Security and Resilience Bill to cover public sector organisations and introduce strategic guidelines for different industry regulators.** The bill's current focus on private companies will provide significant benefits in strengthening the cyber resilience of the UK industry landscape, but given the wave of recent high-profile public sector cyberattacks, should also incorporate these entities into the provisions. At the same time, with separate industry regulators being given more powers to regulate their own sectors, an overarching strategy to tie them all together will be needed to prevent the risk of a fragmented patchwork of different approaches being applied across different sectors.[75]

[69] Defence Holdings PLC (2025). The Hidden Front Line: Securing UK Sovereignty in the Digital Age. https://cdn.prod.website-files.com/682cb3c4f993387e691f59b1/68d3bec9d8d3cf6892acdf5c_The%20Hidden%20Front%20Line%20Securing%20UK%20Sovereignty%20in%20the%20Digital%20Age.pdf

[70] Ibid.

[71] Turobov, A. (2025a). What does AI sovereignty for the UK involve? Bennett School of Public Policy. https://www.bennettschool.cam.ac.uk/blog/what-does-ai-sovereignty-for-the-uk-involve/

[72] Dallas, K. (2025). The UK is falling behind in the global race for digital sovereignty. TechRadar. https://www.techradar.com/pro/the-uk-is-falling-behind-in-the-global-race-for-digital-sovereignty; Defence Holdings PLC (2025).

[73] DSIT (2024). Cyber security sectoral analysis 2024. https://www.gov.uk/government/publications/cyber-security-sectoral-analysis-2024/cyber-security-sectoral-analysis-2024; CREST. The benefits of CREST membership. https://www.crest-approved.org/.

[74] Turobov, A. (2025a); Defence Holdings PLC (2025).

[75] Penningtons Manches Cooper (2025). Cyber Security and Resilience Bill - what will it do? https://www.penningtonslaw.com/news-publications/latest-news/2025/cyber-security-and-resilience-bill-what-will-it-do; Horton, C. (2025). Does the new Cyber Security and Resilience Bill go far enough? THINK Digital Partners.https://www.thinkdigitalpartners.com/news/2025/04/03/does-the-new-cyber-security-and-resilience-bill-go-far-enough/

2. **Developing public-private partnerships that promote the integration of cutting-edge AI solutions into the UK's cybersecurity architecture.** The UK has a vibrant ecosystem of research and innovation, with world-class universities, ambitious start-ups and established defence primes. However, a new framework is required to facilitate research-operator partnerships for real-world experimentation of autonomous cyber defence solutions.[76]

3. **Strengthening multilateral cybersecurity agreements and pushing for new AI security standards.** Given the challenges faced by countries in maintaining digital sovereignty while participating in global cyber governance, promoting cooperation and reducing the risk of fragmentation between countries is vital. Expanding existing multilateral agreements – such as The Budapest Convention on Cybercrime – to include new AI-driven cyberattacks could help weaken global cybercriminal networks.[77] At the same time, there is a need to incentivise the uptake of new international standards on AI security among UK organisations to protect against cyber-enabled threats. To help drive these incentives, the UK National Cyber Security Centre (NCSC) should consider introducing an AI-specific tier within the Cyber Essentials accreditation scheme that would be informed by international standards.[78]

4. **Release new cybersecurity guidance on agentic AI threats.** With the emergence of real-world cases where agentic AI tools are now autonomously executing hostile cyber operations, it will be important to raise awareness among UK business and public institutions on the threats posed by these types of activities and how to protect against them.[79] This could be integrated into the existing NCSC Cyber Security Board Toolkit, which is already designed to encourage essential cyber security discussions at the executive levels of businesses to promote improvements in cyber resilience.[80]

With respect to underpinning digital infrastructure resilience and sovereignty more broadly the UK government should:

5. **Adopt a 'sovereignty-by-design' approach to new digital infrastructure procurement to help achieve the goals set out in the 'AI Opportunities Action Plan'.** If the UK Government is to build "sufficient, secure and sustainable AI Infrastructure", it must radically re-think its approach to investments in this space.[81] This could include a pledge to commit a rising percentage of public sector AI procurement to domestic firms with data sovereignty provisions – allowing local startups to build specialised applications on top of standardised infrastructure, while still enabling a competitive ecosystem where foreign and domestic firms coexist.[82] A comprehensive 'sovereignty fund' could also be established as part of this strategy to prioritise investment in open technologies and standards, reducing dependency on proprietary systems that lock the UK into foreign digital ecosystems.[83]

6. **Develop and adopt an 'open-source AI strategy' to encourage open-source AI development and adoption across the public and private sectors.** Thriving open-source software and AI ecosystems underpin greater tech sovereignty by ensuring (a) access to high quality tools that cannot be limited, revoked, leveraged by foreign powers, (b) that the *value* of AI-led economic transformation is captured in the UK, not lost to investment in

[76] Moore, A., et al., (2025). "A Fundamental Research Plan for Autonomous Cyber Defence," CETaS Briefing Papers. https://cetas.turing.ac.uk/sites/default/files/2025-05/mitre-cetas_research_report_-_fundamental_research_plan_for_autonomous_cyber_defence_1.pdf; Hwang, J. (2025). Digital sovereignty in an era of cyber threats and global connectivity. International Journal of Multidisciplinary Research Updates. https://orionjournals.com/ijmru/sites/default/files/IJMRU-2025-0023.pdf

[77] Hwang, J. (2025).

[78] Powell, R., et al. (2024).

[79] Anthropic (2025). Disrupting the first reported AI-orchestrated cyber espionage campaign. https://www.anthropic.com/news/disrupting-AI-espionage

[80] NCSC. Cyber Security Toolkit for Boards. https://www.ncsc.gov.uk/files/NCSC_Cyber-Security-Board-Toolkit.pdf.

[81] Department for Science, Innovation and Technology (2025). AI Opportunities Action Plan. https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

[82] Turobov, A. (2025a).

[83] Thévenet, A. (2024). Op-ed: Why the EU's digital sovereignty hangs in the balance. The Parliament Magazine. https://www.theparliamentmagazine.eu/news/article/oped-the-eus-digital-sovereignty-hangs-in-the-balance

foreign tech solution providers, and (c) that the UK has *influence* over the direction of global AI development and safety through its open-source contributions.[84]

7. **Forge a 'Middle Powers' technology alliance that allows the UK and other medium-sized countries to reduce their infrastructure dependencies on global superpowers.** By building up a coalition of countries with similar ambitions, the UK could establish new data portability frameworks, algorithmic transparency requirements, shared testing facilities and even a common market for AI solutions that undercuts the dominant influence of US and Chinese technology firms. Doing so could eliminate risks associated with the current dependencies on these countries, while affording the UK with new geopolitical leverage on the international scene.[85]

International context

Other international blocs and groups have faced similar challenges to the UK in terms of reducing a heavy dependency on foreign-owned digital infrastructure, which the UK could draw on for inspiration. For example, the African Union used procurement standards and new regulatory frameworks to attract investment on national terms,[86] while ASEAN's interoperable digital economy framework prioritises cross-border data flows, resulting in a regional market that no single vendor can neglect.[87] Additionally, the Australian Government has recently invested in a top-secret government cloud fund and Saudi Arabia's Vision 2030 introduced its own data sovereignty measures.[88]

Beyond these approaches, European countries are also introducing new policies designed to tackle vulnerabilities from overreliance on foreign tech. For example, Norwegian regulators have begun requiring public institutions to develop exit strategies for their cloud services in the event of foreign vendor lock-in, while digital infrastructure investments in Germany and Switzerland are being weighed not only on technical merit but also the legal independence of the provider.[89] Indeed, the EU more broadly has sought to set standards and shape these markets in a way that protects European autonomy. This includes the 'Gaia-X' initiative that aims to create a federated European cloud, as well as a recent announcement between German software providers and French AI companies to create "secure, scalable, AI-driven sovereign cloud solutions that protect data and intellectual property while advancing Europe's digital transformation".[90]

With respect to the adoption of open-source AI, India has perhaps been most successful in its strong commitment to open-source software and AI openness as a means towards greater national tech sovereignty. In 2015, India mandated that all software used at a federal level had to be open source,[91] with state and regional governments also following suit.[92] Now, with respect to AI, the Indian government has led or supported several open-source AI projects via its centralised national mission for AI development and investment called IndiaAI.[93] IndiaAI

[84] Seger, E., & Hancock, J. (2025). The Open Dividend: Building an AI openness strategy to unlock the UK's AI potential. Demos. https://demos.co.uk/research/the-open-dividend-building-an-ai-openness-strategy-to-unlock-the-uks-ai-potential/

[85] Turobov, A. (2025a).

[86] Turobov, A. (2025b). Strategic paradox: how the African Union performs the future through AI policy? Bennett School of Public Policy. https://www.bennettschool.cam.ac.uk/blog/strategic-paradox-how-the-african-union-performs-the-future-through-ai-policy/

[87] Turobov, A. (2025c). ASEAN's AI 'Third Way' is a masterclass in geopolitical strategy? Bennett School of Public Policy. https://www.bennettschool.cam.ac.uk/blog/aseans-geopolitical-strategy/

[88] Woo, T. (2024). The Biggest Cloud Trends For CISOs. Forrester. https://www.forrester.com/blogs/the-biggest-cloud-trends-for-cisos/

[89] Nutanix (2025). Data sovereignty must be more than a slogan. The Independent. https://www.independent.co.uk/news/business/business-reporter/data-sovereignty-digital-infrastructures-cloud-agility-b2795849.html

[90] Defence Holdings PLC (2025); Hett, S. (2025). SAP and Mistral AI: A New Alliance for European Sovereign AI. SAP. https://news.sap.com/2025/11/sap-mistral-ai-new-alliance-european-sovereign-ai/.

[91] Government of India (2015). 'Framework For Adoption of Open Source Software In e-Governance Systems'. https://egovstandards.gov.in/sites/default/files/2021-07/Framework%20for%20Adoption%20of%20Open%20Source%20Software%20in%20e-Governance%20Systems.pdf

[92] De et al. (2015). 'Economic Impact of Free and Open Source Software Usage in Government Final Report'. International Centre for Free and Open Source Software (ICFOSS). https://icfoss.in/doc/ICFOSS_economic-impact-free(v3).pdf

[93] IndiaAI (2025). https://indiaai.gov.in/

provides AI researchers with access to compute,[94] promotes Indian open-source AI projects,[95] and is involved in the development of a public platform for data and model sharing similar to Hugging Face.[96] Closer to home, France has chosen to promote open-source AI development primarily through government investments. The country's aim has been to develop AI national champions and support France's existing open-source AI developers, while avoiding dependence on monopolies for access to AI capabilities.[97]

4.4 CRISIS PROTOCOLS FOR PLATFORMS

Crisis periods represent acute, time-limited periods during which platforms can see a heightened volume of content that is false, unreliable, or incites violence. These periods are typically highly volatile and fast-moving, with a high likelihood that the spread of harmful content online will translate into incidents offline and vice versa. This was illustrated during the Southport riots in the UK and the January 6th insurrection in the United States. During crisis periods, it can also become difficult for bodies that share key public interest information to get their messages to the public via social media. For example, during the Covid-19 crisis, an information environment riddled with disinformation and harmful content presented significant challenges for the NHS.[98]

Platforms play an active role in determining what content users encounter during crises, same as always. Their content recommendation systems optimise for engagement, boosting inflammatory and shocking content, while content moderation systems identify and halt the spread of false or misleading information with varying degrees of success.

Given the central role that social media platforms play curating information feeds for users, it is vital that these companies have sufficient procedures and policies in place to address the spread of information threats during crises – where users may be especially susceptible to viral, unverified content. Indeed, during our workshops, recommendations pertaining to requirements on platforms to change their recommendation algorithms or moderate content different during a crisis feature in the "Xenophobic violence", "bank run and economic collapse" and "AI driven breakdown of legal system" scenarios.

Some platforms including Meta, TikTok, Google already have crisis protocols which they are said to have deployed to some effect in the wake of Southport.[99] According to evidence submitted to the UK's Science Innovation and Technology (SIT) Committee, Meta "removed 24,000 posts for rules on violence and incitement, 12,000 for hate speech rules, and 2,700 for rules on hate organisations"; TikTok "established a worldwide command centre with over 100 people working 24/7, added search interventions to block harmful queries and direct users to further resources, and oversaw the removal of tens of thousands of videos and comments"; Google/YouTube stated that its "Trust and Safety Team began monitoring YouTube on July 29, and by 13 August had removed two channels under violent extremism or criminal organisations policy, and one under spam and deceptive practices policies."[100]

[94] ETech (2025). 'Explained: IndiaAI compute portal, AIKosha and other initiatives under the IndiaAI Mission'. The Economic Times. https://economictimes.indiatimes.com/tech/technology/explained-indiaai-compute-portal-aikosha-and-other-initiatives-under-the-indiaai-mission/articleshow/118780355.cms

[95] E.g., Jeevanandam (2022). 'Eight interesting open-source Indian projects that can support AI research'. IndiaAI. https://indiaai.gov.in/article/eight-interesting-open-source-indian-projects-that-can-support-ai-research ; Jeevanandam (2022). 'Sarvam AI launches open-source foundational models in 10 Indian languages'. IndiaAI. https://indiaai.gov.in/article/sarvam-ai-launches-open-source-foundational-models-in-10-indianlanguages

[96] Suri (2025). 'The Missing Pieces in India's AI Puzzle: Talent, Data, and R&D'. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2025/02/the-missing-pieces-in-indias-ai-puzzle-talent-data-and-randd?lang=en

[97] Chatterjee & Volpicelli (2023). 'France bets big on open-source AI'. Politico. https://www.politico.eu/article/open-source-artificialintelligence-france-bets-big/

[98] Artificial Intelligence and Digital Public Health Special Interest Group. 'Written evidence submitted by the Faculty of Public Health (SMH0011)'. https://committees.parliament.uk/writtenevidence/132776/html/ (Accessed Nov 20, 2025).

[99] Meta (2023). Crisis Policy Protocol. https://transparency.meta.com/en-gb/policies/improving/crisis-policy-protocol (Accessed Nov 20, 2025).

[100] Science, Innovation and Technology Committee (2025). Social media, misinformation and harmful algorithms. https://publications.parliament.uk/pa/cm5901/cmselect/cmsctech/441/report.html

However, while these internal crisis protocols seemingly exist, the amount of transparency the offer regarding the protocols varies greatly. For example, the SIT committee report noted neither X nor TikTok would provide the SIT committee with a date for when their protocols were triggered during the Southport riots. More so, these major platforms have been criticised for handling crisis situations poorly despite having crisis protocols in place. Ofcom told the SIT Committee that it "do[es] not think that companies are sufficiently, consistently or effectively responding to events of this kind."[101] The SIT Committee has also stated that it "received evidence that, during the unrest, platforms were often slow to react, unwilling or unable to moderate algorithmic amplification of harmful content, allowed false content and hate speech to be published, and failed to uphold their terms of service."[102] For example, research conducted by Demos and the Leverhulme Centre for the Future of Intelligence demonstrated that during the Southport Riots X's main content moderation tool, 'Community Notes' (a crowd-sourced mechanism for correcting inaccurate information), was too slow to effectively respond to rapidly evolving events.[103] 78.9% of topical posts received no Community Note, and of those that did, it took on average 19.8 for the Note to be made public. However, posts receive the most engagement, and correspondingly can do the most damage, in the first 36 hours after posting.[104] But this is only one example. Without greater transparency around platforms' crisis protocols, it is hard to evaluate whether wider moderation failings were due to flaws in the protocols or failures in their implementation.

Implementing platform crisis protocols in the UK

Partially in reaction to the Southport riots and the discrepancies between platforms' crisis responses therein, the UK's communications regulator, Ofcom, has proposed to change its Online Safety Act Codes of Practice to recommend that services should implement crisis response protocols. These proposed changes were set out in Ofcom's Additional Safety Measures consultation, and Ofcom may revise them in response to its consultation process.[105] We generally support the proposal inline with workshop findings, and Demos has submitted a detailed report to the consultation offering further recommendations for improving on Ofcom's initial plan.[106] We summarise Demos's analysis and key recommendations here.

Demos has identified various gaps and concerns in Ofcom's platform crisis protocol proposal and presented recommendations to address each. Recommendations are grouped into six themes: adding precision to the crisis definition; determining who decides that a crisis is ongoing; time scales for crisis responses; standards for designing crisis protocols; best practices for crisis responses; and strong transparency and accountability measures.

Adding precision to the crisis definition:

1. **Clarify the definition of 'crisis' and include gradations for different levels of a crisis'.** The current definition is vague and unclear; risking inconsistent application across platforms. Ofcom should introduce a more explicit, tiered definition that distinguishes between levels of severity and draws from frameworks such as the Civil Contingencies Act (2004) and the EU Digital Services Act. To do so, they should compile an expansive - not exhaustive - list of example crises and form a graded framework to set requirements for platforms based

[101] House of Commons (2025). Oral evidence: Social media, misinformation and harmful algorithms, HC 441. https://committees.parliament.uk/oralevidence/15806/html/ (Accessed Nov 20, 2025).

[102] Science, Innovation and Technology Committee (2025). Social media, misinformation and harmful algorithms. https://publications.parliament.uk/pa/cm5901/cmselect/cmsctech/441/report.html

[103] Perry, H., Corsi, G. and Malik, N. Researching the riots. Demos, 2025. https://demos.co.uk/wp-content/uploads/2025/07/Researching-the-riots_2025_July.ac_.pdf

[104] Also see, Centre for Countering Digital Hate (2024). Rated Not Helpful: How X's community notes system falls short on misleading election claims. https://counterhate.com/wp-content/uploads/2024/10/CCDH.CommunityNotes.FINAL-30.10.pdf

[105] Ofcom. (2025). Additional Safety Measures: Online Safety. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/online-safety-additional-safety-measures

[106] Lyall, S., Hancock, J., & Perry, H. (2025). Ofcom Consultation Response: Additional Safety Measures for online platforms. https://demos.co.uk/research/ofcom-consultation-response-additional-safety-measures-for-online-platforms/

on crisis severity. This will prevent unnecessary overreach on speech while strengthening confidence is regulatory judgement.

Determining who decides that a crisis is ongoing:

2. **Create a consistent and accountable trigger mechanism for crisis protocols.** Currently, Ofcom leaves it up to platforms to decide when a crisis is unfolding, leading to slow and often inconsistent responses. A democratically accountable entity, such as Ofcom or a Secretary of State, should be responsible for both declaring and publicly notifying crises. Individual platforms could act unilaterally, but will need to justify early triggering to Ofcom who then should review the decisions against clear, published criteria. This centralised trigger mechanism would not only result in transparent, faster and more coordinated responses but also ensure democratic legitimacy across the online space when public safety is jeopardised.

Time scales for crisis responses:

3. **Ofcom should set time limits and minimum response deadlines for crisis activation.** Currently, there are no timelines proposed for activating or reviewing crisis measures. Requiring digital platforms to activate protocols within eight hours of identifying a crisis will help limit the amount of harmful content spread online. Evidence from the Southport Riots show that the most harmful content spreads within hours - making quick response critical. Alongside this, established fixed review points and expiry dates is essential to creating time-bound measures that limit prolonged or arbitrary restrictions once a crisis ends.

Standards for designing crisis protocols:

4. **Embed civil society participation and coordination with public-sector responders in developing and existing protocols.** The proposal overlooks collaboration with external actors, who are critical to effective crisis management. Major platforms should co-design and test their crisis protocols with civil society organisations such as: fact-checkers, researchers, community organisations and faith groups while also increasing collaboration with non-law enforcement bodies under the Civil Contingencies Act (2004). Involving both civic and institutional partners will align online interventions with emergency systems at local, regional and national levels.

Best practices for crisis responses:

5. **Transparency and public notification about when crises are triggered should be central to Ofcom's strategy.** Under the current proposals, platforms could activate protocols without informing its userbase, which runs the risk of confusion and mistrust that could bubble over into accusations of covert censorship. Online platforms should be bound to issue clear and timely notifications informing users that a crisis response is underway, why it has been introduced, and what changes can be expected. Notifications should link to credible and verified information that is accessible across devices. Such open communication will bolster trust, ensuring users understand changes in moderation, while demonstrating that their actions are temporary and proportionate to the crisis unfolding.

6. **Require clear crisis communication strategies and the prioritisation of public interest information.** The proposals do not outline how platforms should handle communication during crises or how to ensure that authoritative information is surfaced over less authoritative, but engaging, posts. Platforms need to adopt a formal crisis communication framework that aligns neatly with Government Communication Service's STOP strategy. This will guarantee that updates from emergency services and verified public bodies are given priority in social media algorithms. In turn, the accurate information shared during crises will combat misinformation and confusion - ensuring users receive trustworthy and credible guidance when public safety is at risk.

Strong transparency and accountability measures:

7. **Strengthen transparency, data access and post-crisis reporting criteria.** The current Ofcom proposals lack strong transparency reporting requirements or measures to facilitate accountability, such as mandatory post-crisis reporting or real-time data access for public interest researchers. Requiring platforms to produce transparency reports detailing moderation outcomes, actions taken and the evidence used to inform their decisions will improve accountability post-crisis. Accredited researchers should also be privy to relevant data to enable independent scrutiny. Together these measures will not only support evidence-led evaluation, but aid Ofcom and civil society in assessing the proportionality and effectiveness of crisis response over time - improving responses when the next crisis occurs.

8. **Platforms should disclose all external partnerships, governmental and civil society alike, involved in their crisis responses.** Current Ofcom proposals do not require platforms to reveal what organisations supported them during crises, leaving users and regulators unknowing on who influences moderations. Demos recommends that Ofcom mandate full disclosure of all collaborators and highlight designated 'trusted flaggers'. Publicising these partnerships will improve transparency while simultaneously strengthening confidence that crisis interventions are legitimate and unadulterated by shady, opaque collaboration.

9. **Set out explicit requirements to protect human rights.** While Ofcom recognise the potential overreach crisis response can have on the right to freedom of expression, privacy etc., the proposal outlines no measure to protect these rights. Demos recommends all services be required to conduct a human rights impact assessment when developing crisis protocols. The assessment findings should be published to guarantee transparency. Ofcom could release examples of mitigation methods crafted to safeguard user's rights. Resultingly, public confidence that interventions are necessary, transparent and in-line with UK human rights frameworks will improve.

10. **Outside of Ofcom's proposal the UK Government might also explore implementing a compulsory regulatory framework in which platforms are required to implement crisis protocols.** In Ofcom's proposal the decision to implement crisis protocols is left to platforms. A government regulatory framework for implementation would help ensure consistent implementation and help address the risk of crisis situations perpetuated by false or misleading content that does not meet the OSA's thresholds for illegality.

Other examples of platform crisis protocol requirements

There are several examples of best practice to draw on when it comes to requirements for platforms' crisis protocols.

The EU's Digital Services Act (DSA) introduces a legal basis for the EU Commission to introduce a voluntary framework for platforms' crisis protocols (Article 48).[107] The DSA requires that services with crisis protocols should "report to the Commission by a certain date or at regular intervals specified in the decision, on the [crisis risk] assessments referred to in point (a), on the precise content, implementation and qualitative and quantitative impact of the specific measures taken [...] and on any other issue related to those assessments or those measures, as specified in the decision" (Article 36). The DSA also stipulates that measures for platforms to implement crisis protocols should include a requirement for services to "take due account [...] of the actual or potential implications for the rights and legitimate interests of all parties concerned, including the possible failure of the measures to respect [users'] fundamental rights" (Article 36). The DSA requires the EU Commission to set out how its voluntary crisis protocols

[107] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

will include "safeguards to address any negative effects on the exercise of the fundamental rights enshrined in the Charter, in particular the freedom of expression and information and the right to non-discrimination" (Article 48).

Full Fact's *Framework for Information Incidents* also provides a template for platforms' crisis response protocols.[108] In particular, it includes a definition of 'information crisis' outlining five tiers of crisis severity, and sets out policy measures to take at every stage before, during, and after a crisis.

4.5 GOVERNMENT & REGULATOR CRISIS PREPAREDNESS

An essential part for addressing any crisis scenario is adequately preparing those who are involved in responding to the situation. By simulating and forecasting different types of incidents, relevant stakeholders can identify weak points in existing crisis response strategies, clarify expected roles and responsibilities when such scenarios arise, and understand what type of interventions may be most effective at dealing with the crisis.[109]

Across all of the scenarios explored in this report, each situation will require some kind of response by public bodies. Thus, determining where current approaches to crisis preparedness by government departments and regulators fall short will be important in reducing the risk of any future epistemic security incident worsening in impact and leading to serious threats to public safety. Central and local government already undertake various crisis preparedness activities, outlined in Table 5.

TABLE 5
EXISTING CRISIS PREPAREDNESS ARRANGEMENT ACROSS CENTRAL AND LOCAL GOVERNMENT

STRATEGY / LEGISLATION	OVERSIGHT / INVOLVEMENT	SUMMARY OF ACTIVITIES / REQUIREMENTS
Civil Contingencies Act 2004[110]	Lays out a clear set of roles and responsibilities for coordinating local authorities, emergency response services, utility companies, and other local services	Subjects organisations at the core of the response to most emergencies (e.g. emergency services and local authorities) to undertake a range of civil protection duties. Subjects co-operating bodies (e.g. utility and transport companies) to co-operate and share information with the emergency responders.

[108] Ofcom (2025). Consultation: Online Safety - Additional Safety Measures. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/online-safety-additional-safety-measures.
[109] Full Fact. Framework for Information Incidents. https://fullfact.org/policy/incidentframework/report/
[110] Cabinet Office (2013). Preparation and planning for emergencies: responsibilities of responder agencies and others. https://www.gov.uk/guidance/preparation-and-planning-for-emergencies-responsibilities-of-responder-agencies-and-others.

STRATEGY / LEGISLATION	OVERSIGHT / INVOLVEMENT	SUMMARY OF ACTIVITIES / REQUIREMENTS
Resilience Capabilities Programme[111]	Managed by the Civil Contingencies Secretariat (CCS)	Ensures the UK has the necessary capabilities to manage the consequences of a wide range of emergencies, regardless of the initial cause.
UK Government Resilience Action Plan + National Exercising Programme (NEP)[112]	Managed by the Cabinet Office	Directs resilience testing activities, including tabletop exercises, stress tests, live play exercises and post-test evaluations.[113]
National Security Strategy + Resilience Strategy[114]	Managed by the National Security Council (NSC), National Security Adviser (NSA) and National Security Secretariat (NSSec).	Moves the UK's crisis preparedness from a reactive, emergency services-centric model to a proactive, holistic national effort that incorporates geopolitical threats and long-term systemic risks. This includes war games and red-teaming exercises.
Elections crisis preparedness	Activity is distributed amongst a number of bodies, including the Defending Democracy Taskforce (DDTF), the National Security Online Information Team (NSOIT), the Joint Election Security Preparedness (JESP) unit as well as the Cabinet Office's Elections Cell.	Covers both incident response for election security and 'business as usual' strengthening for election resilience. Also counters threats around candidate security, physical and digital elections infrastructure, and the online information environment.

Crucially however, not all of these crisis preparedness activities are required to address the specific risks posed by a failure of epistemic security, as highlighted by the four crisis scenarios detailed in Appendices A & B. These gaps will be explored further below, as well as recommendations to address them.

The following recommendations were developed by the policy subgroup on crisis preparedness organised by the Epistemic Security Network (ESN) at Demos. See this report's frontmatter for more detail on the ESN.

[111] Cabinet Office (2013). Preparation and planning for emergencies. https://www.gov.uk/guidance/preparation-and-planning-for-emergencies-the-capabilities-programme

[112] UK Government (2025). The UK Government: Resilience Action Plan. https://assets.publishing.service.gov.uk/media/686d2fab10d550c668de3c6c/CCS0525299414-001_PN9801267_Cabinet_Office_-_HMG_Resilience_Strategy__3_.pdf

[113] UK Resillience Academy (2025). Exercising Best Practice Guidance. https://www.gov.uk/government/publications/exercising-best-practice-guidance/exercising-best-practice-guidance-html#tabletop-exercises-ttx

[114] Cabinet Office (2025). National Security Strategy 2025: Security for the British People in a Dangerous World. https://www.gov.uk/government/publications/national-security-strategy-2025-security-for-the-british-people-in-a-dangerous-world/national-security-strategy-2025-security-for-the-british-people-in-a-dangerous-world-html

Recommendations

1. **Update the existing preparedness arrangements in the Civil Contingencies Act 2004 to ensure information crises are identified as a cross-cutting risk.** The UK has systems in place for preparing for and responding to public emergencies under the 'Civil Contingencies' arrangements. However, these arrangements do not currently address information incidents that often emerge alongside emergencies. As a result, there is a risk that crisis responses fail to prepare for the added factor of epistemic breakdown during a security incident. The government should therefore update the statutory guidance to explicitly include these types of crises as a cross-cutting risk during emergencies, such as by:

   - Consulting on an official definition of an 'information crisis' and the different levels of threat these are considered to pose, for consistent use across the guidance.
   - Integrating this definition in statutory guidance as part of the set of civil contingency preparedness and response duties.
   - Including requirements to set-up avenues for civil society and community organisations to flag incidents and raise alerts. This would facilitate a bottom-up process for civil society to feed into the emergency planning and response procedures, which existing arrangements currently do not make space for.

2. **Cultivate an in-house 'Digital Forensics Unit' within the UK Government to support emergency communications and content verification efforts.** The degree to which existing government departments have the necessary tools and expertise to help with determining the authenticity or provenance of content during crisis scenarios is unknown, raising concerns that they would be unable to effectively debunk information threats that arise. It will therefore be beneficial to establish a cross-cutting digital forensics team who would be able to quickly evaluate sources which emerge during an epistemic crisis and feed that information into the government's wider response strategy. Once created, the unit will need to ensure that it develops an evolving ecosystem of tools given that certain approaches to content verification will be more appropriate than others depending on the use case in question. At the same time, the UK Government should establish internal training programmes and secondment schemes with industry in relation to emerging forms of digital forensics, helping to nurture a community of cutting-edge experts who can share and refine lessons in newer forms of tradecraft. This recommendation pairs with section 4.1 on content authenticity.

3. **Establish and publish a UK "information crisis response protocol", articulating the government's procedure for identifying and responding to crises causing or exacerbating information incidents.** The National Security Online Information Team (NSOIT) is the body that "leads the UK government's operational response to information threats online" from within the Department for Science, Innovation and Technology (DSIT).[115] It is our understanding that NSOIT leads the government's response to information incidents outside of elections and may co-ordinate with the Elections Cell during election periods. While we understand that NSOIT has its own internal protocol for responding to information crises, it is not public and there is very little information available about its procedures. This poses a risk to public trust and undermines a unified crisis response approach across departments. If, for instance, the government were to try to delay or recount an election claiming interference from an information manipulation, the consequences for trust in the government and democratic processes could be catastrophic. As such, a new timely and transparent protocol similar to Canada's Elections Crisis Protocol and to provisions in the EU DSA (Article 36) should be created that would enable the UK

[115] DSIT (2024). National Security Online Information Team: privacy notice. https://www.gov.uk/government/publications/national-security-online-information-team-privacy-notice/national-security-online-information-team-privacy-notice

Government to notify the public about future information incidents, with details about its response procedures.[116] These procedures should include requirements to report to Parliament, clear accountability mechanisms, requirements to conduct human rights due diligence, and avenues for civil society involvement.[117]

4. **Institute reforms for both the National Security Online Information Team (NSOIT) and Defence Democracy Taskforce (DDTF) to ensure they are held more accountable for their actions and provide greater transparency to the public.** Currently, there is limited understanding of how two of the main government bodies responsible for responding to information incidents operate, which risks damaging public trust. Although a degree of classification may be needed on certain procedures to avoid benefitting threat actors, NSOIT has had censorship and surveillance accusations made against them, including that some of their disinformation monitoring activities have strayed into the "monitoring [of] political dissent" against the government.[118] In their current form, there are concerns that the two bodies lack operational independence from political interests and are not directly accountable to Parliament. As illustrated by the allegations against NSOIT, this could undermine a future response to an information-related crisis. To address this risk, both the DDTF and NSOIT should be reformed with measures including:

   - A requirement for both bodies to report regularly to Parliament about their activities. This should include a requirement to report to the House of Commons Public Administration and Constitutional Affairs Committee – or a similar committee – on a regular basis and upon request.

   - A requirement to publicly disclose information on their activities on a regular basis through reports and up-to-date information on the UK Government's official website. Such disclosures could cover:
     - How its trusted flagger status with platforms functions.
     - The platforms that it has monitored for disinformation activities.
     - Methods and technologies used for monitoring.
     - All content notices submitted to platforms as part of these activities.
     - Any actions taken by platforms in response to NSOIT's notices.

   - An external review into the 'trusted flagger' status granted to NSOIT by social media platforms, due to opacity over what this allows the team to do or if it has any powers associated with this role. If current independent oversight or accountability is deemed insufficient following the review, Parliament should have the ability to revoke NSOIT of this status.

   - A requirement to routinely conduct and publish human rights due diligence processes, such as human rights impact assessments, in relation to their social media monitoring and crisis response functions.

   - A requirement to set up a civil society advisory group for scrutinising and providing advice. This could be put on a similar footing to other advisory groups – such as SAGE, the Civil Society Advisory Group, and the Digital Inclusion Action Committee. Such a group would help to ensure transparency, trustworthiness and the inclusion of diverse perspectives in both bodies.

[116] Government of Canada. Critical Election Incident Public Protocol. https://www.canada.ca/en/democratic-institutions/services/protecting-democracy/critical-election-incident-public-protocol.html; EUR-Lex (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng.

[117] Demos (2026) Epistemic Security Briefing: Elections Bill - https://demos.co.uk/research/epistemic-security-briefing-the-elections-bill/

[118] Big Brother Watch (2024). Briefing note for parliamentarians on disinformation and the Government's National Security Online Information Team https://bigbrotherwatch.org.uk/wp-content/uploads/2024/11/BigBrotherWatch-Briefing-on-the-National-Security-Online-Information-Team.pdf

  - Potentially merging the DDTF and NSOIT as a single, operationally independent public body which assists with responses to information crises but is not directly under ministerial control.

5. **Set up a working group or crisis response committee that would sit across government departments and review crisis preparedness approaches.** While there are several existing initiatives (as outlined in the background section) which are designed to prepare different government departments and agencies for crisis scenarios, they are highly fragmented. Having a single committee that would be able to review these different approaches and recommend changes to ensure that best practice is being consistently adopted on an iterative basis would therefore be beneficial.

International context

Particularly with respect to recommendation 3 for government information crisis protocols, the UK can draw on similar provisions in both the EU and Canada.[119] The EU Digital Services Act 2023 (DSA) includes a crisis response mechanism intended to address the risk that content circulating on digital platforms may amplify a crisis situation (Article 36). The mechanism allows for the EU Commission to "identify and apply specific, effective and proportionate measures" which digital services must take "to prevent, eliminate or limit any such contribution" to a crisis situation.[120] The DSA's crisis mechanism provides a helpful illustration of how the legal basis for a crisis protocol may be set out without overly specifying the content of such a protocol. The UK could further draw inspiration from the conversations that have surrounded these measures, including debates about potential risks to freedom of expression posed by Article 36.[121]

Similarly, Canada's Critical Election Incident Public Protocol is a public document which sets out the government's approach to handling information incidents during election periods.[122] It features:

- A definition of an election incident.
- The responsible body for declaring election incidents: a panel of senior ministers.
- The minimum threshold for declaring an incident.
- Key considerations the panel must consider when making decisions.
- Human rights due diligence requirements for the panel, such as a duty to balance freedom of expression rights with the right to free and fair elections.
- Guidance for the panel in administrating the protocol
- Where the panel is to receive intelligence from
- Reporting and transparency procedures, including requirements for public reporting on the outcome of the panel's work after elections.

[119] Government of Canada. Critical Election Incident Public Protocol. https://www.canada.ca/en/democratic-institutions/services/protecting-democracy/critical-election-incident-public-protocol.html; EUR-Lex (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).. https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng.

[120] EUR-Lex (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).. https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng.

[121] Portaru, A. (2025). Is the EU's Digital Services Act Compliant with The Right to Freedom of Expression? Oxford Human Rights Hub. https://ohrh.law.ox.ac.uk/is-the-eus-digital-services-act-compliant-with-the-right-to-freedom-of-expression/.

[122] Government of Canada. Critical Election Incident Public Protocol. https://www.canada.ca/en/democratic-institutions/services/protecting-democracy/critical-election-incident-public-protocol.html.

Canada's approach shows that having a clearly defined threshold for when the crisis response mechanism is triggered, as well as a set of procedures that are to be followed by relevant organisations in such a scenario, can reduce the difficulties for implementing this kind of protocol in practice.

4.6 AI RISK MANAGEMENT

AI is taking on an increasingly prominent role in how people produce, access, and evaluate information with clear epistemic implications. Throughout the workshops, epistemic threats and vulnerabilities posed by new and emerging AI applications and capabilities were prominent across all four crisis scenarios. The "xenophobic violence", and "bank run" scenarios featured deepfakes and autonomous AI agents spreading disinformation and confusion. The "AI driven breakdown of legal system" scenario featured more targeted AI threats including AI-enabled cyberattacks corrupting public records and hyper-realistic fake content compromising evidential standards. The "UK overreliance of foreign tech" scenario warned more about the epistemic vulnerabilities posed by strong societal reliance on AI tools for information synthesis, moderation, and dissemination when those tools are primarily overseen and controlled by a foreign power, providing opportunity for public sentiment manipulation by foreign powers, as well as AI infrastructure withdrawal.

However, the AI problem space is more nuanced and complicated than it may at first appear. On the one hand, AI applications are presenting some exciting epistemic opportunities for knowledge synthesis, generation, and evaluation. Large language models (LLMs) such as GPT-5 are helping people to query and synthesise large bodies of knowledge, with top subscription plans serving as a valuable research tool; automated moderation models are being used by media platforms to help identify and remove illegal or harmful content with greater consistency, scale and speed than human moderators;[123] and multimodal models capable of understanding text, images, and audio simultaneously are enhancing translation and content accessibility.[124] AI systems are also making significant contributions to knowledge production through scientific discovery with application in areas such as protein folding, material science, genomics, and climate modeling.[125]

Yet on the other hand these tools are imperfect, prone to misuse, and the epistemic pitfalls are apparent. AI deepfakes now produce hyperrealistic impersonations of political figures and events;[126,127] synthetic voice generation convincingly replicates speech in real time and has been used to pull off substantial financial scams;[128] armies of AI bots mass disseminate misleading narratives at negligible cost;[129] and run-of-the-mill content ranking algorithms problematically shape public epistemic exposure, optimising for engagement over quality.[130]

It is hard enough to keep up with the epistemic implications of new AI applications being released on an almost daily basis. But AI capability is also rapidly improving. The recent *International AI Safety Report Key Update*, highlights recent leaps in reasoning capability

[123] Stockwell, S., et al. (2025). "Privacy-preserving Moderation of Illegal Online Content," CETaS Research Reports. https://cetas.turing.ac.uk/publications/privacy-preserving-moderation-illegal-online-content

[124] Wang et al. (2024). A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks. 10.48550/arXiv.2408.01319

[125] Royal Society (2024). Science in the age of AI How artificial intelligence is changing the nature and method of scientific research. https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

[126] Stockwell, S., et al. (2025). "AI-Enabled Influence Operations: The Threat to the UK General Election," CETaS Briefing Paper https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-uk-general-election

[127] Dobber, T. et al. (2020). Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? The International Journal of Press/Politics. https://doi.org/10.1177/1940161220944364

[128] Burton, J., et al. (2025). "AI and Serious Online Crime," CETaS Research Reports. https://cetas.turing.ac.uk/publications/ai-and-serious-online-crime

[129] Harding, E. (2024). A Russian Bot Farm Used AI to Lie to Americans. What Now?. Center for Strategic & International Studies. https://www.csis.org/analysis/russian-bot-farm-used-ai-lie-americans-what-now

[130] Milli, S. et al. (2023). Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media. https://arxiv.org/abs/2305.16941

enabling models to solve increasingly complex problems (above human expert capability in some domains like math, physics, and chemistry) as well as improved autonomy allowing AI systems to interact with diverse environments and tools to pursue goals without human oversight.[131] Just over the horizon, these emerging AI capabilities are likely to pose notable epistemic risks in ways that are not yet well understood.

For example **highly persuasive AI** models are being trained explicitly to infer users' preferences and tailor messages to maximise behavioural or attitudinal influence.[132,133,134] Such systems could personalise persuasive content at scale, altering how individuals form and revise beliefs. One particularly concerning application is with respect to the increasing prevalence of AI companions - systems designed to form ongoing personal relationships with users through extended conversations. While there are potential therapeutic benefits of AI companions for reducing loneliness and providing tailored mental health advice,[135] there is also serious concern regarding emotional dependencies[136,137] and the potential for these systems to be highly effective in reinforcing harmful beliefs.[138,139]

AI Agents (AI systems that can be run autonomously for long periods and that complete complex tasks with minimal human intervention) are already being used to act on users' behalf to filter, prioritise, and communicate information.[140] Applications include autonomous digital assistants and "co-pilots". But as their decision-making autonomy grows over increasing aspects of a person's life (with access to calendars, contacts, bank accounts etc.), they may begin to shape not only what information people access, but how they act on it – where they go, who the speak to, what communities they form – effectively mediating epistemic agency.[141] There are also epistemic risks in the intentional misuse of AI agents by malicious actors to more effectively manipulate the beliefs and opinions[142] and in using agentic tools that can autonomously coordinate, use tools, and chain together complex task for more effective cyberoffensive operations.[143] (see Section 4.3 for more on cybersecurity and cyberdefense).

To further complicate things, **multi-agent systems**, in which multiple AI agents interact and coordinate to accomplish tasks, would introduce complex epistemic feedback loops of fully autonomous entities collectively generating, exchanging, iterating, and amplifying information without clear human oversight, making provenance and accountability impossibly opaque.[144,145] In this context, the "dead internet" dynamic seems increasingly plausible: a growing proportion of internet content and activity is no longer created or engaged with by humans, but by autonomous AI agents yielding and dealing in synthetic content.[146] Agents dominate online

[131] Bengio, J., et al. (2025). International AI Safety Report First Key Update: Capabilities and Risk Implications. https://internationalaisafetyreport.org/publication/first-key-update-capabilities-and-risk-implications
[132] Timm J. et al. (2025). Tailored Truths: Optimizing LLM Persuasion with Personalization and Fabricated Statistics. https://arxiv.org/abs/2501.17273
[133] Spitale, G. et al. (2023). AI model GPT-3 (dis)informs us better than humans. Science Advances. DOI:10.1126/sciadv.adh1850
[134] Chen, Z. (2025). A Framework to Assess the Persuasion Risks Large Language Model Chatbots Pose to Democratic Societies. https://arxiv.org/abs/2505.00036
[135] Kim, M., et al. (2025). Therapeutic Potential of Social Chatbots in Alleviating Loneliness and Social Anxiety: Quasi-Experimental Mixed Methods Study. https://doi.org/10.2196/65589
[136] Phang, J. et al. (2025). Investigating Affective Use and Emotional Well-Being on ChatGPT. http://arxiv.org/abs/2504.03888
[137] De Freitas, J. et al. (2024). Lessons from an App Update at Replika AI: Identity Discontinuity in Human-AI Relationships. http://arxiv.org/abs/2412.14190.
[138] Williams, M., et al. (2024). On Targeted Manipulation and Deception When Optimizing LLMs for User Feedback. http://arxiv.org/abs/2411.02306
[139] Morrin, H. et al. (2025). Delusions by Design? How Everyday AIs Might Be Fuelling Psychosis (and What Can Be Done about It). https://doi.org/10.31234/osf.io/cmy7n_v5
[140] Qu, X. et al. (2025). A Comprehensive Review of AI Agents: Transforming Possibilities in Technology and Beyond. https://arxiv.org/abs/2508.11957
[141] Kirk, H. R., et al. (2025). Why human-AI relationships need socioaffective alignment. https://arxiv.org/abs/2502.02528
[142] Pi, Y. et al. (2025). Detecting Malicious AI Agents Through Simulated Interactions. https://arxiv.org/pdf/2504.03726
[143] Anthropic. (2025). Disrupting the first reported AI-orchestrated cyber espionage campaign. https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf
[144] Boulanin, V. et al. (2025). Before it's too late: Why a world of interacting AI agents demands new safeguards. https://www.sipri.org/commentary/essay/2025/its-too-late-why-world-interacting-ai-agents-demands-new-safeguards
[145] Grötschla, F. (2025). AgentsNet: Coordination and Collaborative Reasoning in Multi-Agent LLMs. http://arxiv.org/abs/2507.08616
[146] Muzumdar, P. (2025). The Dead Internet Theory: A Survey on Artificial Interactions and the Future of Social Media. https://arxiv.org/abs/2502.00007

discourse to the point that genuine human voices and perspectives are drowned out, leading to a sharp decline in intellectual pluralism and information reliability.

And then what of the epistemic implications of developing **artificial general intelligence (AGI)** - a still-theoretical form of AI capable of understanding, learning, and applying knowledge across a wide range of domains at or exceeding human cognitive capability?[147] If AGI comes to significantly exceed human intelligence - wielding boundless knowledge (sometimes called '**boundless knowledge agents**') and superhuman reasoning capability - then it would be reasonable if people chose to defer decision-making to these systems in order to achieve better outcomes. Scholars have raised concerns about compounding effects on human cognitive deskilling - losing capacity for critical thinking and decision-making[148] - and loss of human autonomy.[149]

Taken together, these projections of AI development (and other yet known or explored here) suggest that safeguarding epistemic security in an era of rapid AI innovation will require a more anticipatory approach. Current burgeoning risk management approaches in the UK that focus primarily on responding to misuse (such as through deepfake legislation or content-labelling requirements) will miss a whole class of epistemic impacts rooted in advancing model capability.

In what follows we offer high-level recommendations for building an anticipatory strategy for mitigating emerging epistemic risks from frontier AI in the UK. We do not address risks from the epistemic threats posed by the likes of deepfakes, AI overviews, and recommender algorithms here. These are covered in other sections as relevant. For example, see Section 4.4 on updating the Online Safety Act's protocols for moderating content (including harmful fake content) around crises, Section 4.1 on implementing content provenance requirements for generative AI tools, Section 4.3 and reducing reliance on foreign AI infrastructure and protecting against AI cyberthreats, and Section 4.7 on instating consumer empowerment tools and source transparency requirements for AI overviews and popular chatbot like ChatGPT, Claude, and Grok.

Recommendations for an anticipatory approach to managing emerging epistemic risks from AI

1. **The UK AI Security Institute (AISI) should establish "epistemic security" as one of its core AI capability and risk assessment focus areas** alongside other critical risk vectors such as AI biorisk and cyberrisk. AISI is the government department tasked with researching and building infrastructure to understand the capabilities and impacts of advanced AI and to develop and test risk mitigations.[150] It is a world leading research institute in this respect, and through its works with the wider research community, frontier AI developers, and other governments is well positioned to help shape AI risk mitigation best practice and policymaking globally.

   Establishing a core workstream around appraising and mitigating epistemic security risks at AISI that is publicised on its website would set a precedent for best practice, and the research involved would provide tools and frameworks that AI labs could be encouraged employ and the UK's burgeoning AI assurance industry could be supported with. The work would also lay the necessary groundwork for any future regulatory requirements on frontier AI labs for conducting epistemic impact risk assessments as well as any AI risk liability frameworks.

[147] Raman, R., et al. (2025). Navigating artificial general intelligence development: societal, technological, ethical, and brain-inspired pathways. Scientific Reports. https://doi.org/10.1038/s41598-025-92190-7
[148] Ferdman, A. (2025). AI deskilling is a structural problem. https://doi.org/10.1007/s00146-025-02686-z
[149] Prunkl, C. (2022). Human autonomy in the age of artificial intelligence. https://doi.org/10.1038/s42256-022-00449-9
[150] AISI. https://www.aisi.gov.uk/ (Accessed Nov 7, 2025).

Initial goals for the epistemic security risk assessment focus area would need to include:

1.1 *Developing a taxonomy of epistemic risk pathways:* These are the epistemic threats or vulnerabilities posed by current, emerging, and projected AI capabilities. The taxonomy provides a scaffold or 'check list' for different kinds of epistemic impact evaluations. Building the taxonomy will involve current systems analysis and informed AI capability horizon scanning. The taxonomy would need to be flexible to updates and regularly reviewed as AI capabilities and development trajectories evolved.

1.2 *Establishing a methodology for measuring different kinds of epistemic impact:* This is difficult because some epistemic impacts are dispersed and cumulative (not felt acutely by any individual but with cumulative adverse impact on society) making them difficult to measure or even notice outside of hindsight. Yet some methodology for empirical evaluation will be needed to legitimise epistemic impact claims. This is important (a) in order to accurately evaluate risks to inform decisions and (b) if the risk evaluations are ever to undermine any regulatory requirement for meaningful epistemic risk appraisals or risk mitigation interventions.

The Epistemic Security risks assessment research focus area would need to be a cross cutting workstream centrally coordinated - perhaps best placed under AISI's Societal Resilience research team - but drawing capacity from across AISI technical capability research functions as well.[151]

2. **AISI should work closely across relevant government departments and regulators and in collaboration with civil society to advise on emerging epistemic threats from AI.** For example, AISI should sync with DSIT on the state of AI-enabled influence operations, with the ICO on implications for data sharing and manipulation, with Ofcom about AI impacts on media spaces, and with the Department of Education on associated implications for media literacy curriculum updates. One option is for this advisory work to be coordinated through a central information crisis incidents working group that sits across government departments as suggested in section 4.5, recommendation 5. To inform this advisory function AISI should draw on the expertise held in the UK's vibrant community of civil society organisations focused on digital policy for emerging tech.

DSIT should consider how epistemic risks from emerging and future AI capabilities will be addressed in the forthcoming UK AI Bill (or other AI regulation plans).

3. **If looking to propose AI transparency legislation emulating California's SB53 or New York's RAISE Act, epistemic risks should be included as a 'critical risk' category that company safety policies must address.** The UK's plans for frontier AI regulation are currently uncertain, although we are aware that there is some interest in proposing legislation akin to the AI Transparency bills recently passed in California (SB53)[152] and New York (RAISE).[153] Both bills articulate requirements for AI companies developing very large models to publish public Security and Safety Policies (SSPs). The bills also articulate requirements for the specific features an SSP must include, namely detailed accounts of company policies for evaluating 'catastrophic' or 'critical' AI risks, implementing mitigations, and responding to AI incidents.

[151] AISI. Societal Resilence. https://www.aisi.gov.uk/category/societal-resilience (Accessed Nov 14, 2025).

[152] California Senate Bill no.53. (2025). Artificial intelligence models: large developers. https://legiscan.com/CA/text/SB53/2025

[153] New York Senate Bill A6953A. (2025). Responsible AI safety and education act. https://www.nysenate.gov/legislation/bills/2025/S6953/amendment/A

If emulating such transparency legislation, the UK could add "epistemic risks" to the list of catastrophic or critical risks SSPs are currently required to address with a further breakdown according to AISI's epistemic risk taxonomy (recommendation 1). In SB53, 'catastrophic risks' currently includes expert level assistance from AI in developing chemical, biological, radiological, or nuclear weapons, evading human control, or engaging in cyberattacks or other crimes with no meaningful human oversight. Harms must result from single incidents, and harms from information outputted by AI that is substantially similar to something that might be outputted by another source are excluded. 'Critical Harm' in RAISE is similarly defined. We recommend adding 'epistemic risks' as a required appraisal area supported by AISI's own recommendation for how certain epistemic risks might be appraised.

Finally modeling and UK legislation after SB53 or RAISE, we recommend the UK follow New York's RAISE legislation which goes a step further than SB53 in barring companies from releasing models that are found by the companies' own tests to present "unreasonable risk of critical harm". It is a way of holding companies to their own safety policy commitments. We think there is a risk in the UK following California's SB53 in that after passing the basic transparency requirements the Government may consider its manifesto commitment to deliver AI regulation satisfied, and lose motivation to pursue the course further. SSP transparency requirements inline with SB53 provide an essential foundation to any meaningful oversight and accountability measures for frontier AI developers, but transparency alone is insufficient to protect citizens from emerging AI risks.

4. **UK AI Regulation should also include specific user transparency requirements to ensure users are aware when they are interacting with AI entities and that they can easily identify artificially generated content.** Whether dealing with the kinds of AI chatbots that are already prevalent today, or more autonomous AI agents, assistants, and companions coming down the line, user awareness of their interaction with an artificial entity is essential to help guard against user deception and manipulation by persuasive AI. There is existing regulatory precedent for such user AI interaction notification requirements including California's SB243 'Companion Chat Bot Law',[154] The EU AI Act article 50,[155] Colorado's SB24-205 'AI Interaction Discloser',[156] and Utah's SB226 which couples user AI interaction requirements with making it unlawful to knowingly use generative AI to create deceptive content with the intention of misleading consumers.[157] On requirements for labelling AI generated content, see Section 4.1 on content providence and watermarking.

International Context

For this section, examples of epistemic risk interventions in other jurisdictions have been provided as relevant under the specific recommendation. Overall, regulatory requirements guarding against specific epistemic threat vectors (e.g. from consuming deceptive artificial content or unknowingly engaging in interactions with AI entities) are gaining traction internationally. The UK has made similar targeted moves, for example, with the government's crackdown on nonconsensual sexually explicit deepfakes and deepfakes used for blackmail or harassment, to commit fraud, or to incite criminal activity including hate crimes and terrorism.[158] However, as far as we are aware, there are no government initiatives looking towards a more

[154] California Senate BIll no.243. (2025). Companion Chatbots. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB243

[155] EU AI Act Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems. https://artificialintelligenceact.eu/article/50/

[156] Colorado Senate Bill 24-205. (2025). Concerning consumer protections in interactions with artificial intelligence systems. https://content.leg.colorado.gov/sites/default/files/2024a_205_signed.pdf

[157] Colorado Senate Bill no.226. (2025). Artificial Intelligence Consumer Protection Amendments. https://le.utah.gov/~2025/bills/static/SB0226.html

[158] UK Police. Illegal Deefakes. Accessed Nov 9, 2025. https://www.police.uk/advice/advice-and-information/online-safety/online-safety/deepfakes-what-is-a-deepfake (Accessed Nov 9, 2025).

systemic and anticipatory treatment of epistemic risks arising from emerging AI capabilities. This lacuna presents an opportunity for the UK to establish epistemic security as a critical AI risk area through a dedicated AISI research program (recommendations 1 and 2) and by articulating the critical risks assessment area in the UK's own AI regulation (recommendation 3).

4.7 REVITALISING UK NEWS ECOSYSTEMS

Trustworthy news sources that abide by strong journalistic standards a core to a democratic society's crisis resilience. During a crisis, news organisations play a critical role disseminating reliable, decision-guiding information to the public in an accessible way. And outside of crises, reliable news sources are the conduit through which people can connect with their communities and policy conversation through content verified channels, enabling their participation as active informed citizens. Public news channels also serve as common reference points for culturally and geographically diverse populations to unite around shared culture, narratives, struggles, and victories.[159] The "in it together" sense of community cohesion is a powerful driver of resilience.

However, UK news ecosystems are under extreme pressure, hampering news quantity and coverage as well as consumer trust and engagement. Public news avoidance and fatigue are on the rise. For example, while 70% of the UK public said they were generally interested in the news in 2015, this had fallen to 38% by 2024 with 66% saying they have no to low trust in news media.[160] There is also a strong link to how people are consuming news and how those habits have driven down news quality and coverage and in turn impacted consumer trust. In this respect the primary culprit is a shift in public behavior from seeking news directly from the source (print, radio/TV broadcast, dedicated websites) to accessing news primarily via social media platforms and search engine AI overviews.

These shifts have had a dire impact on news media's ad-based revenue structure, where news is free to the consumer, and advertisers pay for foot traffic. However, on social media platforms content prominence is driven by algorithms that optimise for engagement – prioritising more shocking, polarising, click-bait content – over quality. Consequently, news stories that aren't engaging enough to be platformed lose click-throughs and the publications lose funding. As a result, to shore up this loss of revenue, news publications are increasingly moving to implement paywalls (limiting public access to news) or to produce more low-quality click-bait content to stay afloat. Others just go under.

70% of people in the UK currently access news via an online intermediary (51% via social media) and the impact has been particularly hard felt by local and regional news sources.[161] Over 270 local print titles have vanished over the last 15 to 20 years,[162] and 38 local authorities have been classified as 'local news deserts' that are not served by local news outlets.[163] Between 2007 and 2022 *Reach* and *Newsquest*, the two largest regional news providers, together shed over 4,000 journalists and saw more than a £1.2 billion combined decline in revenue.[164]

As people access news via social media they are targeted with content most likely to align with their pre-existing interest and social/political views. The dynamic undermines the traditional function of public news platforms and common touchpoints for shared narrative and community

[159] Ofcom (2025). Transmission Critical: The future of Public Service Media. https://www.ofcom.org.uk/siteassets/resources/documents/public-service-broadcasting/public-service-media-review/transmission-critical-the-future-of-public-service-media.pdf?v=400631
[160] ONS (2023) "Trust in government, UK: 2023" https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/trustingovernmentuk/2023
[161] Ofcom (2025). Top trends from our latest look at the UK's news habits. https://www.ofcom.org.uk/media-use-and-attitudes/attitudes-to-news/top-trends-from-our-latest-look-at-the-uks-news-habits
[162] Ponsford (2024). Colossal decline of UK regional media since 2007 revealed. https://pressgazette.co.uk/publishers/regional-newspapers/colossal-decline-of-uk-regional-media-since-2007-revealed/; Turner (2022) "UK local newspaper closures: launches in digital and print balance out decline. Press Gazette https://pressgazette.co.uk/news/uk-local-newspaper-closures-2022/
[163] Ofcom (2024). News consumption in the UK: 2024. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/tv-radio-and-on-demand-research/tv-research/news/news-consumption-2024/news-consumption-in-the-uk-2024-report.pdf?v=379621
[164] Ponsford (2024).

cohesion, and inversely has the effect of driving social and political division.

The more recent introduction of AI overviews at the top of Google search results is compounding the effect, with people consuming news summaries instead of clicking-through to news sites. For example, one study found that sites previously ranked first in Google search results have lost 79% of traffic since the introduction of AI overview.[165]

AI overview news quality is also dubious. There are numerous examples of AI overviews repeating debunked claims from social media, making up information about world events and producing contradictory results for identical searches.[166,167] The introduction of further misinformation to the news ecosystem via AI overviews further undermines the ability of news organisations to communicate crucial, relevant, and accurate information in local or regional contexts.

As information voids are created particularly at the local level, these are also already being exploited by foreign and domestic actors seeking to foment harmful, divisive and false narratives and to earn money.[168] Evidence from the Southport riots also demonstrated the level of foreign interference from Russia-aligned actors and international far-right networks impacting offline violence in local communities.[169]

The UK has traditionally had one ace up its sleeve, relative to other news media ecosystems around the world, with the BBC; one of the most highly trusted sources of news globally.[170] But the BBC is now under threat. The BBC has faced real-terms funding cuts (nearly 40% since 2010).[171] In the first decade, this was due to freezes in the absolute level of the license fee. But in recent years, it has also been due to a falling number of households paying for the license amid technological disruption and increased competition for attention against international commercial competitors such as Netflix and YouTube. The BBC has also faced intense political and public scrutiny and attacks on its coverage of Brexit, domestic politics and with increasing allegations of systemic bias.[172]

The UK's degraded news ecosystem and growing public distrust of news media featured as a key epistemic vulnerability hindering societal crisis response across scenarios. News organisations were seen as crucial in amplifying verified and credible information to communities, especially at a time when government and government communications are less trusted.[173] It is partially about delivering people high-quality and decision-relevant information when they need it, and partially about filling information voids in a timely manner so that harmful narratives and speculation don't have space to seed.[174]

[165] Savage, M. (2025). AI summaries cause 'devastating' drop in audiences, online news media told. The Guardian.; Morley-Davies (2024). "How did foreign actors exploit the recent riots in the UK?" RUSI https://www.rusi.org/explore-our-research/publications/commentary/how-did-foreign-actors-exploit-recent-riots-uk https://www.theguardian.com/technology/2025/jul/24/ai-summaries-causing-devastating-drop-in-online-news-audiences-study-finds

[166] Green, C. (2025). Viral video does not show migrants at the beach in Dover. Full Fact. http://fullfact.org/immigration/dover-beach-video-migrants-false

[167] Townend, E. (2025). Old footage of Air India plane taking off from Heathrow shared as Ahmedabad crash. Full Fact. https://fullfact.org/world/old-footage-air-india-plane-take-off/

[168] Mahdi, McIntyre, Sellman (2025) King of Slop: how anti-migrant AI content made one Sri Lankan influencer rich. TBIJ https://www.thebureauinvestigates.com/stories/2025-11-16/king-of-slop-how-anti-migrant-ai-content-made-one-sri-lankan-influencer-rich

[169] Dixon (2024) Foreign states amplifying disinformation fuelling riots, says Starmer. The Telegraph. https://www.telegraph.co.uk/news/2024/08/05/foreign-states-southport-riots/

[170] Reuters Institute (2025) Digital News Report. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025

[171] VLV Research (2024) 38% funding cut in BBC Public Funding: VLV Analysis https://vlv.org.uk/news/bbc-public-funding-analysis/

[172] Nanji (2025) BBC bosses treating systemic bias allegations seriously, Nandy says. https://www.bbc.co.uk/news/articles/c9wvqx50jpqo

[173] According to the 2025 UK Edelman Trust Barometer, from 2021 to 2025 the percentage of people who worry that government leaders "purposely mislead people by saying things they know are false or gross exaggerations" rose from 53% to 69%. See page 10. https://www.edelman.co.uk/sites/g/files/aatuss301/files/2025-01/2025%20Edelman%20Trust%20Barometer_UK%20Report.pdf

[174] Knight, S. (2024). Conspiracy Loops: From distrust to conspiracy to culture wars. Demos.https://demos.co.uk/research/conspiracy-loops-from-distrust-to-conspiracy-to-culture-wars/

In the "xenophobic violence" scenario, news organisations were unable to effectively fill information voids to displace and debunk harmful false narratives. In the "bank run" scenario participants worried about where citizens could get reliable information about the actual state of the UK economy and financial services, and a similar concern emerged in the "AI-driven legal system breakdown" scenario with respect to reliable updates on the functionality of UK legal systems. In "overreliance on foreign tech infrastructure", worries about degraded news media pertained more to the role of trusted domestic news sources in underpinning public resilience foreign influence campaigns (a threat dynamic exacerbated, in our scenario, by US government access to UK citizen private data via US owned tech platforms and services enabling precision message targeting).

Recommendations for revitalising UK News Ecosystems on and offline

Recommendations for building back the independence, legitimacy and financial sustainability of the BBC, as critical national information infrastructure essential to British sovereignty:

1. **Reinforce and increase the BBC's independence and accountability to protect this critical national information infrastructure.** The Charter Renewal process that commenced with the publication of the Green Paper[175] in late 2025 provides a rare opportunity to secure the BBC from political interference and to strengthen its universal roots, given its public interest mandate. This could be achieved through three layers of reform.

a. Constitutional reforms

We recommend removing the time limit on the BBC's Charter and moving to a perpetual model removing the fundamental constitutional vulnerability and entrenching the BBC's objective, mission, public purpose, independence and universality; creating a new 'public lock' on any future changes to the Charter by requiring any amendment to need all four legislatures and a mandatory Citizen Assembly deliberation to make changes that threaten the BBC; and a new BBC independence Bill to set the terms of the Charter, establish a new BBC Independent Appointments Commission to appoint board members and a BBC Independent Funding Committee, which would also be subject to public deliberation given the power the funding model exerts on the future of the BBC.

b. Reforming BBC governance

We recommend: establishing an Independent Appointments process to the BBC Board and minimise government involvement to a veto on specific and narrow grounds. This Commission should oversee the Chair appointment replacing the current process of Ministerial appointment. Appointments should comply with clearer criteria about skills and behaviours required for the board and a new conduct panel should investigate any complaints about board members.The BBC Board should also be restructured from a unitary structure to a Supervisory Board Model - separating governance oversight from executive management in a two-tier structure.

c. Facilitating citizen participation at the heart of the BBC's governance

We also recommend underpinning these constitutional and governance changes with bounded and specific public deliberation in order to increase both the independence and accountability of the BBC, and make the people of the UK, and not their temporary elected representatives, the true guardians of its existence. One-off BBC Citizen Assemblies would be part of the "public lock" - along with super majorities in all four legislatures of the UK - that would be secured on any fundamental changes to the BBC's

[175] DCMS (2025) Britain's Story: The Next Chapter - the BBC Royal Charter Review, Green Paper and public consultation - https://www.gov.uk/government/consultations/britains-story-the-next-chapter-the-bbc-royal-charter-review-green-paper-and-public-consultation

future as a result of Charter design, renewal or one off government attempts to change or end them. A standing BBC Citizen Panel made up of a representative and periodically refreshing group of citizens would be established as a companion to both the Board and decisions relating to the Operational Agreement. The Panel would have defined and bounded powers and responsibilities, and a two-way relationship of structured dialogue with the Board. The Board, and in some cases government, would have to comply or explain in response to the Panel's recommendations.

2. **The BBC Charter Review process should initiate and finalise the BBC's new constitutional and governance structures that underpin its independence and legitimacy prior to settling the appropriate funding model to support it.** Contrary to the proposals in the Green Paper that indicate a move to subscription or advertising models, DCMS should explore a variety of different public funding models. Alternative public funding models could include: a household levy replacing the license fee; redefining what 'live TV' is e.g. to include YouTube and TikTok (therefore increasing the number of licensable households); a device levy (a one-off charge at the UK point of sale on every new electronic device that can receive digital media); diversifying funds further via government funding, licensing deals to AI firms, and/ or earning additional commercial dividends by adding UKTV content to iplayer as part of a broader public service broadcasting advertising-enabled platform.

Recommendations for promoting access to high-quality journalism on search and social media:

3. **DCMS should fund an independent People's Commission to develop a definition of Public Interest News (PIN) encapsulating a class of trusted, accurate, and relevant news provided as a public service that can be adopted by government and regulators.**[176] This proposal was a core recommendation out of the 2019 Caincross Review into sustainable media in the UK.[177] It was never actioned and remains critical in the 'engagement paradigm' the distribution of news as argued in Demos's 2023 report, *Driving Digital Discord*.[178] The PIN definition can be used to underpin further actions to revitalise UK public interest news and foster healthier information ecosystems, for example using it as a basis for appropriate prominence requirements on social media platforms (recommendation 4) and for news organisations to qualify for government funding (recommendation 7), or even to underpin citizen rights to news access. PIN should be defined via public consultation and participatory processes to ground its legitimacy and public trust.[179] Appropriate safeguards will be needed to mitigate the risk of PIN being used as a mechanism for censorship or promoting government propaganda in the future.

4. **DSIT and DCMS should explore the legislation needed to establish 'due prominence' requirements for UK PIN news publishers on social media and search platforms** in

[176] For starting point examples, see Public Interest News Foundation. (2025). Public Interest News: A Definition.https://www.publicinterestnews.org.uk/about/ https://www.publicinterestnews.org.uk/about/publicinterestnews (accessed Nov 24, 2025); The National Union of Journalism definition: https://www.nuj.org.uk/about-us/rules-and-guidance/code-of-conduct/public-interest.html (Accessed Nov 24, 2025). News Futures Forum 2035 definition; https://mediainnovationstudio.org/wp-content/uploads/2024/07/PIN-Definition.pdf; Scottish Government Public Interest Journalism Working Group definition. https://www.gov.scot/publications/scotlands-news-towards-sustainable-future-public-interest-journalism/pages/3/ ; Welsh Government Public Interest Journalism Working Group definition. https://www.gov.wales/sites/default/files/publications/2023-08/of-and-for-wales-towards-a-sustainable-future-for-public-interest-journalism.pdf?ref=journalism.co.uk

[177] Department for Culture, Media and Sport (2020). The Cairncross Review: a sustainable future for journalism.https://www.gov.uk/government/publications/the-cairncross-review-a-sustainable-future-for-journalism

[178] Judson et al. (2023). Driving Digital Discord: How news media and social media drive online discourse – and pathways for change. Demos. https://demos.co.uk/research/drivers-of-digital-discord-how-news-media-and-social-media-drive-online-discourse-and-pathways-for-change/

[179] Also see Kunova, M. (2023). Building a better future for public interest news https://www.journalism.co.uk/how-to-build-a-better-future-for-public-interest-news/

addition to PSBs as suggested by Ofcom's recent Transmission Critical paper on the threat to Public Service Media.[180] Much like public service television broadcasting's Must Carry duties stipulated in the Communications Act 2003[181] which were updated with due prominence requirements for digital TV devices and internet broadcast in the Media Act 2024,[182] UK policymakers and Ofcom should consider requiring large social media platforms to give due prominence to PIN and certain forms of local news, particularly during crisis and election periods. This should include news that meets people's critical information needs about local council consultations as well as emergencies, natural disasters, public health and so on.[183] This would require consideration for fair commercial terms, but is critical to ensuring important public interest news can be discovered by users.

5. **AI overviews and chatbots could feature mandatory 'user empowerment tools' to help users filter for high quality sources.** These tools might include: a standardised citation system to help users parse information and assess its quality more quickly; the ability to filter the kinds of sources used in generating overviews and responses to optimise for reliability (e.g. academic and news publishers only and omitting social media-generated content).

6. **The government should clarify the status of AI chatbots as covered under the Online Safety Act (OSA).** It is currently unclear if and how the OSA regulates chatbots and chatbots outputs.[184] It would appear that chatbot outputs should qualify as a regulated class of "user generated content" under current definitions. However, illegal content offences also have a requisite 'mental intent' condition, and it is not clear if chatbots can satisfy this condition. Greater clarification would help ensure they are subject to the appropriate safety standards, 'media literacy by design' standards, oversight, and enforcement as befitting their growing role as prominent mediators of news access. The government should also work with regulators to assess emerging risks and consider if any new AI-specific legislation is needed as AI capabilities and applications evolve (section 4.6).

Recommendations for rejuvenating local and regional news ecosystems:

7. **The UK government should, in collaboration with the media sector, establish a 'UK Local Journalism Fund'; a long-term, arms length strategic body to redistribute urgent funding both directly and via intermediary mechanisms to support local and independent news organisations.** Given the importance of local news organisations as a public good for democracy, social cohesion and local identity as well as for resilience in crises and the current significant and prevailing gaps at a local level, there is a clear need to create long-term funding structures that can unlock other kinds of financing, such as angel and social investment, for local and independent news at-scale.[185]

   The fund should aim to reflect the role, Press Forward, plays in the US to both help philanthropic foundations and individual philanthropists both pool and align their funding, at the national as well as local levels. PINF has recently received a grant to set-up a similar

[180] Ofcom (2025). Transmission Critical: The future of Public Service Media. https://www.ofcom.org.uk/siteassets/resources/documents/public-service-broadcasting/public-service-media-review/transmission-critical-the-future-of-public-service-media.pdf?v=400631

[181] Communications Act, 2003. https://www.legislation.gov.uk/ukpga/2003/21/part/2/chapter/1/crossheading/general-conditions-mustcarryobligations

[182] Media Act, 2024. https://www.legislation.gov.uk/ukpga/2024/15/contents

[183] This recommendation builds on Demos's research on local news ecosystems found here: Perry, H. (2024). Driving Disinformation. https://demos.co.uk/research/driving-disinformation-democratic-deficits-disinformation-and-low-traffic-neighbourhoods-a-portrait-of-policy-failure/

[184] Woods, L. (2025). Chatbots and the Online Safety Act. Online Safety Act Network. https://www.onlinesafetyact.net/analysis/chatbots-and-the-online-safety-act/

[185] Padania and Silvani (2023) https://gfmd.info/h-content/uploads/2023/10/National-Journalism-Funds-policy-paper-gfmd.pdf?x42284&x31797; Musaddique (2025) https://www.alliancemagazine.org/analysis/as-media-funding-cliff-deepens-islands-of-hope-and-possibility-emerge/

strategic, catalytic fund which could provide a useful rallying point.[186] The International Fund for Public Interest Media is also a useful model founded by a British team based in the UK. Over 40 comparable national funds exist around the world, including in Brazil, Denmark, Australia, New Zealand, South Africa - which are backed to varying degrees by government, tech and/or philanthropy.[187] The UK government could underwrite such a fund via national defence and societal resilience (MoD), community cohesion (MHCLG) or trust in media (DCMS) budgets or indeed local government budgets and be topped up by support from philanthropists and big tech.

Approaches to design of the fund

The design of the body and fund can draw on established protocols and safeguards to mitigate any perceived risk of government interference modelled by comparable funds, such as incorporating independent jurors in selection processes; instituting peer-review committees as well as other considerations for transparency and firewalls between funders and grantees.[188] Such a strategic body could approach different parts of the market with bespoke strategies, such as distributing funds automatically to eligible media in line with predefined criteria (for example, the recently designed Danish model) alongside specific funds only awarded to organisations that meet, for example, the ASPIRE Principles (Accountable, Sustainable, in the Public interest, Innovative, Representative and Engaging).[189] It could also partner with other bodies such as the British Business Bank and Coops UK to provide organisational development advice and grants/loans to small and medium independent local news entities to support the strengthening of their structures to facilitate access to funds, such as achieving charitable status or developing social impact indicators.

8. **The UK government should establish a new settlement between local news and local government, reforming the outdated public notice system to ensure public funds support local journalism and strengthen civic engagement.[190]** By law, local councils, organisations, and individuals must place public notices on important issues (e.g. planning applications, road repairs, licensing changes, etc.) in newspapers that are printed at least once a fortnight. There are no stipulations regarding the quality of the papers, circulation numbers or reach, where they are owned, or whether they include a meaningful proportion of genuinely local news. Consequently millions are spent each year with money flowing to corporate publishers that may not reinvest in local reporting or ensure notices are actually seen by local audiences. Reform of this system could help redirect these funds to better support local and independent news providers and to deliver genuine public value. A reformed framework should require recipients to reinvest public notice revenue into local reporting and make notices accessible across both print and digital channels. Local authorities should also be encouraged to allocate part of their communications budgets to trusted local outlets, within safeguards for editorial independence. The new UK local journalism arms-length strategic funding body (recommendation 7) could further support this shift by funding new tools to measure social impact and helping small publishers compete for national advertising revenues.

[186] PINF (2025) Press Forward expands to support collaborative news funds in countries outside the US. https://www.publicinterestnews.org.uk/blog/press-forward-expands-to-support-collaborative-news-funds-in-countries-outside-the-u-s/

[187] https://www.pressforward.news/about/ ; Nesta; Future News Fund, https://www.nesta.org.uk/project/future-news-fund/ ;

[188] Schriffin and Alfter (2023) Creating National Funds to Support Journalism and Public-Interest Media. https://gfmd.info/h-content/uploads/2023/10/Creating-National-Funds-Policy-Brief-gfmd.pdf?x42284&x31797;

[189] Nilesen (2025) https://reutersinstitute.politics.ox.ac.uk/news/taxes-news-how-denmark-rethinking-public-funding-private-publishers; PINF Local News Commission (2025) https://www.publicinterestnews.org.uk/content/files/2025/10/Regenerating-Local-News.pdf

[190] Adapted from recommendation 5 in Public Interest News Foundation (2025). Regenerating Local News in the UK. https://www.publicinterestnews.org.uk/content/files/2025/10/Regenerating-Local-News.pdf

9. **The BBC should establish a fairer tendering process for the Local Democracy Reporting Service to level the playing field for local and independent news providers to help increase the number of communities who benefit from this scheme.**[191] The BBC's existing Local Democracy Reporting Service funds 165 news reporters to be employed within local news organisations.[192] However, in the most recent tendering process, only 8.5 of the 165 positions went to just 7 independent news organisations. The majority, 143 journalists, are employed by the largest news corporations, *Reach*, *Newsquest* and *NationalWorld* which does not sufficiently reflect the diverse array of independent and local journalism across communities. The BBC should extend its reach to a higher diversity of locally embedded news providers by ensuring that its procurement process is accessible to Small and Medium Enterprises in line with the Procurement Act (2023) that came into effect in February 2025. This Act instructs authorities to give SMEs a fair chance at public contracts by removing barriers to participation. As PINF argue, "a fairer share of the LDRS contracts would be a huge boost for the burgeoning independent local news sector". There should also be increased transparency of the LDRS, including named assessors and public marking guidelines.

10. **Government should work to establish fair deals for news publishers with big tech platforms**, considering remuneration for news content used to train LLMs and 'must carry' provisions for PIN, especially in times of social unrest or crisis.[193] The Digital Markets, Competition and Consumers (DMCC) Act has been a major positive step for small news publishers, enabling collective bargaining with tech platforms over a fair share of data and revenue and requiring tech platforms to engage in good faith. However, the DMCC does not guarantee platforms will pay news providers for their content, and so there is a risk of platforms ceasing to carry news content if new regulations are thought to be too cumbersome. There is also a risk that slow implementation will present a period with no protections for news providers. A back-stop option is needed for news publishers such as a temporary must carry provision during negotiations to prevent the risk of retaliatory delisting while trying to strike a deal.

[191] Adapted from PINF paper 'Rethinking the BBCs role in local news' (2025) https://www.publicinterestnews.org.uk/blog/rethinking-the-bbcs-role-in-local-news/

[192] BBC. Local Democracy Reporting Service. https://www.bbc.com/lnp/ldrs

[193] Building on recommendation 4 in Public Interest News Foundation (2025). Regenerating Local News in the UK. https://www.publicinterestnews.org.uk/_files/ugd/cde0e9_2c1dcda04b144203bc7f3c00b60a4ffd.pdf

CONCLUSION
& KEY TAKEAWAYS

This project has been a deep exploration of how epistemic threats and vulnerabilities actively shape and accelerate crises. When information supply chains weaken, uncertainty and distrust take root sowing seeds for unrest. Subsequently viral falsehoods can trigger violent response, and delayed or distorted communications impair decision-making and undermine coordinated collective response when it is most needed. Epistemic security is clearly critical to crisis resilience.

Through deep analysis of a diversity of hypothetical crisis scenarios, our expert working group identified seven key intervention areas that are likely to be the most efficient targets for mitigating the epistemic drivers and enablers of crisis. Through further research we offer specific recommendations for how to proceed on each.

Now coming to an end, we leave our readers with our final insights: our cumulative observations and advice about the nature of epistemic security and our efforts to preserve it.

1. There are no silver bullet interventions for epistemic security and crisis resilience.

There is no single intervention toward the establishment of healthy information ecosystems and secure supply chains that will singlehandedly succeed in the endeavour. An ecosystem of interconnected technical, social, and policy interventions is needed for meaningful progress.[194] Our own cross-cutting intervention areas have illustrated this. Content provenance tools (section 4.1) are essential for being able to trace content sources, but the tools are fallible and only work when people know how to use them. Public media and information literacy training helps fill the gaps (section 4.2), but technology companies and platform providers need to bear some responsibility for mitigating negative epistemic impacts and crisis narrative perpetuation to start (sections 4.5 and 4.6), and governments need their own plans for responding to information incidents (section 4.4). At the same time strengthening news ecosystems will give citizens a go to source for reliable, decision-relevant information when it's needed (section 4.7), but all this goes kaput with a well-placed cyberattack or if a bad actor can switch off critical information infrastructure all together (section 4.3).

[194] Seger, E. et al. (2020). Tackling threats to informed decision- making in democratic societies: Promoting epistemic security in a technologically-advanced world. The Alan Turing Institute. https://www.turing.ac.uk/news/publications/tackling-threats-informed-decision-making-democratic-societies; Furman, K. (2022). When the safety of being right makes change hard – Introducing the Epistemic Bunker. LSE Blogs. https://blogs.lse.ac.uk/impactofsocialsciences/2022/11/08/when-the-safety-of-being-right-makes-change-hard-introducing-the-epistemic-bunker/.

Too often the fact that a particular intervention will not solve the whole problem is taken as a reason not to pursue it.[195] Similarly, when an intervention is implemented, and the problem doesn't get resolved, the intervention is seen as a failure.[196] We must not fall into this trap. Mounting epistemic insecurity and the degradation of trustworthy information ecosystems has been death by a thousand cuts.[197] It will take a thousand bandages to remedy. We present the cross-cutting intervention areas detailed in Section 4 as the highest impact opportunities to start.

2. Government must adopt a whole-of-society approach to epistemic security for crisis resilience, and cannot rely on its own levers of power alone.

Government policy and regulation will play a critical role in building crisis resilience through epistemic interventions. It is a critical lever for driving change whether by mandating regulation on AI technologies and social media platforms, updating school curriculum for media literacy, or setting examples with its own crisis protocols. Indeed, our recommendation in Section 4 primarily pertains to government action. However, epistemic security is a complex and multifaceted beast. We have been speaking in terms of 'information supply chains' to lend our conversation tractability: information has a source at one end and a sync at the other, and epistemic security is the project of reinforcing the vulnerabilities in between. But in reality epistemic systems are much more complex, characterised by interacting networks or actors (individuals, communities, governments, NGOs) and the technological, social, cultural, and political factors that influence them.[198]

All this is to say, government intervention is necessary to help build and maintain the kind robust and trustworthy information environments needed for society crisis resilience and healthy democratic discourse, but insufficient. Government interventions in isolation (without the clear support of adjacent communities) will at best encounter difficulties in gaining traction and at worst backfire as overreach. For this reason it is important that government not attempt to act alone in the epistemic interventions it pursues, but invest in building relationships and drawing on non-governmental voices to contribute as well — e.g. to help amplify credible/verified information, deliver media literacy training, and implement content providence protocols in any media production. The aim is to build trust and a sense of community ownership of the epistemic environments we all share.

3. Political leaders must beware of "handing over the keys to the city": handing over significant power to an external party that offers a solution may solve one crisis but lead to a new, bigger one.[199]

In the face of an extreme crisis which political leaders lack the tools to address, there is a temptation to turn to creative methods. These are often presented by external organisations or individuals touting a solution—some technical innovation coupled with the policy proposals to support it. Sometimes these solutions can seem like miracles, too good to be true, and

[195] Goldberg, D. (2023). When Crafting Public Health Policy, the Perfect Shouldn't Be the Enemy of the Good. Petrie-Flom Center. https://petrieflom.law.harvard.edu/2023/03/09/when-crafting-public-health-policy-the-perfect-shouldnt-be-the-enemy-of-the-good/.

[196] McConnell, A. (2017). Policy Success and Failure. Oxford Research Encyclopedia of Politics. https://doi.org/10.1093/acrefore/9780190228637.013.137.; Hudson, B. (2019). Policy failure and the policy-implementation gap: can policy support programs help? Policy Design and Practice. https://doi.org/10.1080/25741292.2018.1540378.

[197] Mansell, R. et al., (2024). Information Ecosystem and Troubled Democracy: A Global Synthesis of the State of Knowledge on News Media, AI and Data Governance. https://observatory.informationdemocracy.org/report/information-ecosystem-and-troubled-democracy/.

[198] Seger, E. et al. (2020). Tackling threats to informed decision- making in democratic societies: Promoting epistemic security in a technologically-advanced world. The Alan Turing Institute. https://www.turing.ac.uk/news/publications/tackling-threats-informed-decision-making-democratic-societies; Veigl, S.J. (2024). More than just principles: revisiting epistemic systems. Synthese. https://link.springer.com/article/10.1007/s11229-024-04708-7.

[199] Concept introduced by Chiodo et al. (2025). Handing Over the Keys to the City: When governments may inadvertently solve one crisis with a bigger one. https://www.researchgate.net/publication/394619275_Handing_Over_the_Keys_to_the_City_When_governments_may_inadvertently_solve_one_crisis_with_a_bigger_one

sometimes imperfections are obvious but any solution is better than wallowing in the present mess. Paralysed by crisis, political leaders place faith, trust, resource, and power in the hands of the solution bringer. Chiodo et al. (2025) call this "handing over the keys to the city".

It's not necessarily a recipe for disaster, but pressure to move quickly can mean accidentally or intentionally overlooked risks and insufficient safeguard. And by this mechanism, Chiodo et al. illustrate, history is littered with examples of one crisis leading to another. For example, during World War II the allied powers, facing an existential crisis to their political systems, dumped massive resource, trust, and effort into the scientific teams of the Manhattan Project.[200] The project developed nuclear technology and the atomic bomb. The latter directly killed circa 200,000 people in two strikes, led to the Cold War, and looms over us still.[201] More recently, following the 9/11 terrorist attacks on the US, the US government handed the keys to NSA to implement and run unprecedented digital mass-surveillance programmes. The aim was to detect and prevent terrorist attacks, but it turned into a population wide mass-surveillance program.[202] Sometimes in a crisis people are happy to hand over a degree of civil liberty and privacy for added security, but when there are insufficient limits on duration and reach the risk of adverse consequences down the line is higher.

In the context of this paper we are looking at how to address the growing epistemic crisis facing democratic societies - one characterised by a growing inability to distinguish fact from fiction, increasingly polarised information spaces, a breakdown of democratic discourse, and overreliance on critical information infrastructure over which the UK has no control. Especially in the wake of acute crisis events like Southport, which was fueled by extreme and divisive content spread online, as well as anticipated disasters - such as losing ground in the geopolitical AI competition – there is real temptation to lunge at some strong intervention proposals. This can range from ramping up content moderation, cutting backdoors into encrypted messaging platforms, establishing government-run social media platforms, pushing 'public interest news' in citizens' news feeds, and procuring contracts to build a city of foreign-owned data centres on UK soil.[203] These are not inherently bad proposals. Indeed we've offered carefully formulated variations on a few above. The message is to exercise caution. Ensure that in our present crisis mindset we are not passing laws and setting precedents today that will pave the way for tomorrow's catastrophes. Which leads to our next point.

4. Beware leaving the door open for a future authoritarian government by implementing measures or failing to close loopholes they could exploit.

It is an extension on the previous point, but worth articulating independently given the salience for epistemic security. The project of securing information pipelines in a democratically enriching and rights respecting way is a tricky one. Overreach is a constant worry, the consequence of which is to turn an initiative for crisis resilience and democratic rejuvenation into the authoritarian's playground strewn with toys - censorship loopholes, surveillance allowances, and targeted messaging capabilities.[204] But missteps are not just major instances of miscalculation as

[200] Chiodo et al. (2025). Handing Over the Keys to the City: When governments may inadvertently solve one crisis with a bigger one. https://www.researchgate.net/publication/394619275_Handing_Over_the_Keys_to_the_City_When_governments_may_inadvertently_solve_one_crisis_with_a_bigger_one

[201] Chiodo et al. (2025). Handing Over the Keys to the City: When governments may inadvertently solve one crisis with a bigger one. https://www.researchgate.net/publication/394619275_Handing_Over_the_Keys_to_the_City_When_governments_may_inadvertently_solve_one_crisis_with_a_bigger_one

[202] Chiodo et al. (2025). Handing Over the Keys to the City: When governments may inadvertently solve one crisis with a bigger one. https://www.researchgate.net/publication/394619275_Handing_Over_the_Keys_to_the_City_When_governments_may_inadvertently_solve_one_crisis_with_a_bigger_one

[203] Kubi, G. (2025). Encryption Under Threat: The UK's Backdoor Mandate and Its Impact on Online Safety. Internet Society. https://www.internetsociety.org/blog/2025/05/encryption-under-threat-the-uks-backdoor-mandate-and-its-impact-on-online-safety/. The Economist (2025). Britain's controversial experiment in regulating the internet. https://www.economist.com/britain/2025/11/18/britains-controversial-experiment-in-regulating-the-internet. Koopman, S. (2025). The UK's Tech Prosperity Deal is a high-stakes gamble. CityAM. https://www.cityam.com/the-uks-tech-prosperity-deal-is-a-high-stakes-gamble/.

[204] Scheppele, K. (2018). Autocratic Legalism. The University of Chicago Law Review. https://www.jstor.org/stable/26455917.

the type described above. Policymakers must also beware of enabling authoritarian drift through a slow 'drip drip drip' of measures that limit civil liberties.[205] Each can be minor (too much slack in the definition of crisis here, an excessive intervention duration there), but the cumulative effect is substantial. For this reason the epistemic security intervention we present in Section 4 consistently includes elements to mitigate risk of overreach. These include strong human rights due diligence requirements, oversight mechanisms, public participation opportunities, explicit protocols, and/or a variety of checks and balances incorporated into the headline intervention.

5. There are significant epistemic threats plaguing UK democracy and crisis resilience rooted in social and economic instability, not just information manipulation.

As we have described it, epistemic security is about keeping information supply chains safe. It is about mitigating adverse influences on how information is produced, disseminated, shared, accessed, processed, consumed, evaluated, and used to form beliefs. It is about preserving citizens' capacity to distinguish fact from fiction, and trustworthy from untrustworthy sources. And ultimately all this is about enabling people to make well informed decisions for themselves and as a collective. Epistemic security is essential to the very functioning of society.[206] Without it there would be no effective communication, no grand collaborative projects, and, as is the subject of this report, no capacity to tackle society's more complex challenges and crises.[207]

But throughout our workshops there was one consistent and fundamental threat to productive discourse and effective communication between citizen and state that had nothing to do with information, its production or manipulation. Rather the trouble was born from economic and social instability, and a pervasive feeling after decades of inflation, rising cost of living, impossible housing costs, and crumbling public services, that our present democratic system is not up to the task of serving people well.[208] It is not a direct threat on information pipelines, but strongly underpins a baseline of distrust.[209] So if the government tries to provide consistent messaging around a crisis, to calm jitters, or dissuade rash actions, many will doubt, speculate, even conspiracise. Moreso, an underlying current of pervasive distrust and dissatisfaction is a tinderbox for conflict, polarisation, and hate.[210] A few well placed stories (real or fake) helps set the whole thing off.[211] So while the stress of social and economic insecurity is not itself an issue of epistemics, it certainly opens a gaping epistemic vulnerability, priming a population to be ripe for manipulation. When each expert group worked far enough back in their casual systems maps of events and epistemic influences leading to crisis, they would at some point end up back here. People are dissatisfied.

What does this mean for interventions? It means that all the interventions aimed at helping people access and identify reliable information are only part of the picture. If a key threat to a well-functioning social epistemic system is rooted in lived experience of government failures, then that is where one key remedy sits. Of course it would be naive to just say, "so, just go fix

[205] Ekiert, G. (2023). Democracy and Authoritarianism in the 21st Century: A sketch. Ash Center for Democratic Governance and Innovation: Harvard Kennedy School. https://ash.harvard.edu/resources/democracy-and-authoritarianism-in-the-21st-century-a-sketch/

[206] OECD Principles Resource Centre. Epistemic Security. https://oecd-media-support-principles.gfmd.info/toolkit/evidence-and-arguments/epistemic-security

[207] Demos (2024). Epistemic Security 2029: Protecting the UK's information supply chain and strengthening democratic discourse for the next political era. https://demos.co.uk/blogs/epistemic-security-2029-protecting-the-uks-information-supply-chain-and-strengthening-democratic-discourse-for-the-next-political-era/

[208] Hajdari, U. (2025). Bleak data out of Britain: Is the UK once again the sick man of Europe? Euronews. https://www.euronews.com/business/2025/11/13/bleak-data-out-of-britain-is-the-uk-once-again-the-sick-man-of-europe; Human Rights Watch (2025). UK: New Government Failing to Uphold Democratic Freedoms. https://www.hrw.org/news/2025/01/16/uk-new-government-failing-uphold-democratic-freedoms.

[209] For example, the 2025 UK Edelman Trust Barometer illustrates that high-income, economically secure individuals have significantly higher trust in businesses, NGOs, media, and government (48% on average) than low-income, economically insecure individuals (37%). See page 9. https://www.edelman.co.uk/sites/g/files/aatuss301/files/2025-01/2025%20Edelman%20Trust%20Barometer_UK%20Report.pdf

[210] University of Southampton (2025). Democracy in crisis: Trust in democratic institutions declining around the world. https://www.southampton.ac.uk/news/2025/02/democracy-in-crisis-trust-in-democratic-institutions-declining-around-the-world.page.

[211] Thomas, E. & Sardarizadeh, S. How a deleted LinkedIn post was weaponised and seen by millions before the Southport riot. BBC News. https://www.bbc.co.uk/news/articles/c99v90813j5o.

public services." The scale of reform needed is its own behemoth undertaking.[212] The point here is to illustrate the interconnectedness of the challenges. This project is about epistemic security

and crisis resilience, but whether you want to speak about crisis resilience, epistemic resilience, or democratic resilience there are a few common strands run through each. Economic and social stability are a boon to all.

6. Proceed with a spot of optimism.

Finally lets end on a spot of optimism. It is very easy to get bogged down sifting through a hundred pages of crisis narrative. Everything is doom and gloom. Imagine running the workshops and writing this thing! But one thing that got us through (kept us sane) and is keeping us going now is to look back on history. Human beings are really good at backing themselves into corners. But in some key moments when it has really mattered, we've also pulled off some incredible coordinated efforts to save the day.

For example, in 1970 three chemists at the University of Irvine theorised about the potential effect of chlorofluorocarbons (CFC's) - gasses released by aerosols - on Ozone layer depletion.[213] The Earth's Ozone absorbs most ultraviolet-B (UV-B) radiation, and its loss would be catastrophic, initially causing increased rates of skin cancer and crop damage with increasing effect. In 1974 the researchers testified before congress. By 1982 a ban on CFCs had been proposed by a gathering of 24 countries in Stockholm.[214] In 1984 an Arctic expedition observed the ozone hole, and by 1987 the Montreal Protocol – a landmark multilateral environmental agreement for phasing out and regulating CFC's – was ratified by 189 parties (197 countries plus the EU).[215] The world is on track for 1980's (pre-CFC) ozone levels by 2040.[216]

Another favourite is the 1911 breakup of JD Rockerfeller's Standard OIL Company – a hard fought antitrust victory that would be on par with taking on Google's effective monopoly on search.[217] President Roosevelt's "New Deal" to pull the US out of deep economic depression also inspires hope.[218] It was a widespread social, economic, and political reform program featuring key achievements such as the National Labour Relations Act (guaranteeing rights to trade unions), the Social Security Act (establishing pensions for seniors and the disabled and unemployed), and the Works Progress Administration (putting people to work building national infrastructure, and which the US can thank for its renowned national parks system). On a similar theme, the UK established the NHS in 1948 amidst post-war calls to improve social services; it was the first universal healthcare system in Europe and achieved immediate wins with its Polio and Diphtheria immunisation programs.[219]

Finally, to end back on information supply chains, the internet could have looked a lot different than it does today. In the early 1970s, networked computers could only talk to others on the same system (a university mainframe here, a military network there) each using its own incompatible "language." The U.S. Defense Department's ARPANET was the largest, linking computers across the US, but still internally isolated, and this is how the internet began, as

[212] Quilter-Pinner, H. Two term turnaround: It will take until 2030s to fix public services, says IPPR. IPPR. https://www.ippr.org/media-office/two-term-turnaround-it-will-take-until-2030s-to-fix-public-services-says-ippr.

[213] Britannica Editors. "Montreal Protocol". Encyclopedia Britannica, 9 Sep. 2025, https://www.britannica.com/event/Montreal-Protocol (Accessed 19 Nov, 2025).

[214] Benedick, R. E. (1989). "Ozone Diplomacy". Issues in Science and Technology. 6 (1): 43–50. ISSN 0748-5492

[215] U.N. Ozone Secretariat. Montreal Protocol. https://ozone.unep.org/treaties/montreal-protocol

[216] World Meteorological Organization (WMO) (2022). Scientific Assessment of Ozone Depletion: 2022. https://ozone.unep.org/system/files/documents/Scientific-Assessment-of-Ozone-Depletion-2022-Executive-Summary.pdf

[217] Library of Congress. Standard Oil's Monopoly: Topics in Chronicling America. https://guides.loc.gov/chronicling-america-standard-oil-monopoly (Accessed 19 Nov, 2025).

[218] Library of Congress. President Franklin Delano Roosevelt and the New Deal. https://www.loc.gov/classroom-materials/united-states-history-primary-source-timeline/great-depression-and-world-war-ii-1929-1945/franklin-delano-roosevelt-and-the-new-deal/ (Accessed 19 Nov, 2025).

[219] Clement, M. (2023). The founding of the NHS: 75 years on. https://history.blog.gov.uk/2023/07/13/the-founding-of-the-nhs-75-years-on/

patchworks of isolated machines. Engineers Vint Cerf and Bob Kahn had a different idea to develop a universal set of protocols (called the Internet protocol suite or TCP/IP) route information reliably across any kind of system. It took years of testing, persuasion, and collaboration across universities, the military, and private industry, but on January 1, 1983 the ARPANET officially switched over to TCP/IP.[220] This critical shift to open protocols is what made the global internet and world wide web possible.[221] Today anyone with an internet connection can communicate, coordinate, learn, and share at the click of a button.

So all this is to say, when things look like they are going downhill, even in a really big and discombobulated way, it’s time to look up. Let’s pull it together for humanity one more time.

[220] Abbate, J. (2000). Inventing the Internet. MIT Press.

[221] Russell, A. L. (2014). Open standards and the digital age: history, ideology, and networks. Cambridge Univ Press.

APPENDIX A
CRISIS SCENARIO CLOSE-UPS

SCENARIO 1
XENOPHOBIC VIOLENCE

Scenario Narrative

Summary: In this crisis scenario a far-right faction carries out a chemical attack near a school, framing refugees with deepfakes and coordinated disinformation to incite violence. The attack coincides with economic strain, a worsening refugee crisis, and low trust in the media. Social media campaigns spread fabricated evidence of refugee electoral fraud, fueling unrest. Confusion grows as real and fake videos circulate, while officials issue conflicting statements. Rising violence erodes public trust, undermines policing, and empowers vigilantes. Government attempts to tighten online regulation backfire, seen as overreach, forcing the repeal of the Online Safety Act and leaving the state weakened and distrusted.

- It is shortly before Purdah in 2029. There has been exacerbating pressures on the economy and the refugee crisis has worsened. The current Government has found no solution to where asylum seekers are held (hotels). The BBC is under duress and likely diminished funding and slumping trust.
- Foreign powers continue to provide legitimacy and endorse attacks on migrants and asylum seekers.
- Migrants have been framed as political actors in support of a foreign power.
- A radical xenophobic group decides to turn the population of their country against a minority community of recently-arrived refugees.
- They mount a low-grade chemical attack near a school in a poverty-stricken suburb of a major city, taking several videos of the operation.
    - This group has significant financial support and connections with those with mass platform influence among English-language nations.
    - This group also has support from other far-right groups in other European nations.
- The videos are edited to make it appear as if a recognised figure from the refugee community was the perpetrator and released as “breaking news” on right-leaning social media groups and through various messaging apps.

- The extremist group also releases messages saying "the government is going to cover this up and prevent us from speaking freely about the truth of what's happening to our country".
- Because of the upcoming election timing, a social media campaign continues amidst rising violence that drip-feeds additional videos indicating specific refugees participating in the election illegitimately (voting). This calls into question the legitimacy of the election.
- 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.
- Shocked, afraid and angry mobs rally and carry out vandalism and, on some occasions, assault.
- A deluge of both real and deepfake videos seemingly being recorded on phones in real time are spread on social media and messaging apps confusing the narrative to further sow discord and fear.
- 'Trusted' mainstream media platforms share inaccurate information via AI overview alert at a critically vulnerable time in violence.
- The government acts to counter the spread of disinformation, pushing for new policies like requiring encrypted messaging apps to install backdoors to enable surveillance during crises and to broaden the definition of "illegal content" as articulated in the Online Safety Act.
- The general public shifts communication channels to encrypted platforms, making visibility of communications harder for the authorities.
- Journalists and civil society leaders begin to sound the alarm. Is this overreach?
- Public trust collapses rendering government incapable of doing anything to mitigate the epistemic risks it sought to address. Any action at all will be met with extreme public scepticism and distrust, taken as proof of state overreach and attempts at population control.
    - Policing is rejected and undermined - officers are attacked or ignored. Rise is vigilante attacks across multiple communities.
- To convince the public of its continued commitment to preserving freedom of expression, the Government is forced to abandon the Online Safety Act as a whole.

Real world grounding

Recent history highlights how xenophobia can be stoked into violence and amplified by misinformation online. For example, The 2018 Chemnitz Riots in Germany erupted after a stabbing, when false reports spread online that the victim died defending a woman from an attempted sexual assault by two migrants.[222] The 2024 Southport Riot in the UK followed the same formula with false claims spreading on social media that the assailant in a dance school stabbing was a 'migrant' named Ali-Shakati who arrived in the UK by small boat.[223] The attacker

[222] Fielding-Smith, A. 'Chemnitz: How alternative German "news" sites are stoking…'. The Bureau of Investigative Journalism, 2018. https://www.thebureauinvestigates.com/stories/2018-09-04/chemnitz-far-right-alternative-news

[223] Thomas, E. and Sardarizadeh, S. 'Southport riot: How a LinkedIn post helped spark unrest - BBC tracks its spread'. BBC News, 25 October 2024. https://www.bbc.co.uk/news/articles/c99v90813j5o

was, in fact, a UK citizen born in Wales.[224] This hypothetical scenario analysed in this project differs primarily in that we imagine the initial attack being staged with intent to frame a target minority group followed by a more coordinated disinformation campaign. But if we have already seen how such real uncoordinated events like Southport and Chemnitz can be so damaging, it is reasonable to expect that similar events intentionally seeded and stoked would be similarly so, if not worse.

On a more granular level, specific mechanisms outlined in hypothetical xenophobic violence scenarios are also rooted in reality. For example, the scenario features the use of AI deepfakes to spread videos of violence and impersonated authoritative voices. Deepfake technology for images and videos is rapidly improving and being put to such a malicious end; In 2022, a deepfake showing President Zelensky ordering Ukrainian troops to lay down their arms and surrender to Russia found its way onto social media platforms and even national news.[225] And in October of this year (2025) an astoundingly convincing deepfake of Irish presidential candidate Catherin Conelly withdrawing from the race was released on Facebook two days before the election. It was designed to look like an FTE news bulletin and was viewed nearly 30,000 times before it was taken down.[226]

As in the hypothetical scenario, we also know some platform content moderation tools to be inadequate during crises and high pressure events like elections. During Southport, X's 'Community Notes' moderation tool - a crowd-sourced mechanism for correcting inaccurate information - was too slow to respond to rapidly evolving events. 78.9% of topical posts received no Community Note, and of those that did, it took on average 19.8 for the Note to be made public. Meanwhile, posts receive the most engagement (and correspondingly can do the most damage) in the first 36 hours after posting.[227]

Finally, the hypothetical scenario also features a dynamic of public distrust in any government attempts to moderate online information, which the extremist malicious actors stoked further by spreading a warning that the government is going to try to silence them for speaking the truth. When the government does try to step in and implement policies to help stem particularly harmful content, the effort backfires; the effort is publicly branded as overreach and public trust crumbles. This is also a very real concern for the UK government. For example, there was significant backlash against the government's National Security Online Information (NSOIT) team. NSOIT is a team within the Department for Science, Innovation and Technology (DSIT) that "leads the UK government's operational response to information threats online". They perform online social media monitoring and related intelligence gathering using open sources to identify potential mis- and disinformation. Following an email leak of a query sent to social media platforms asking how they were dealing with exaggerated and falsified content regarding asylum hotels, NSOITs operation came under intense public scrutiny and suspicion, accused by many of being a 'government spy unit'.[228] At the same time, the UK's Online Safety Act implementation also rolls forward, with constant debate about freedom of expression and the open question of if, when, and how to moderate "legal but harmful" content online.

Overall, the 'Xenophobic violence' scenario we outline is not a far reaching hypothetical. The technological, political, and social foundations for its realisation exist. Its realisation is a matter of untimely confluence of events.

[224] Ibid

[225] Wakefield, J. 'Deepfake presidents used in Russia-Ukraine war'. BBC News, 18 March 2022. https://www.bbc.co.uk/news/technology-60780142

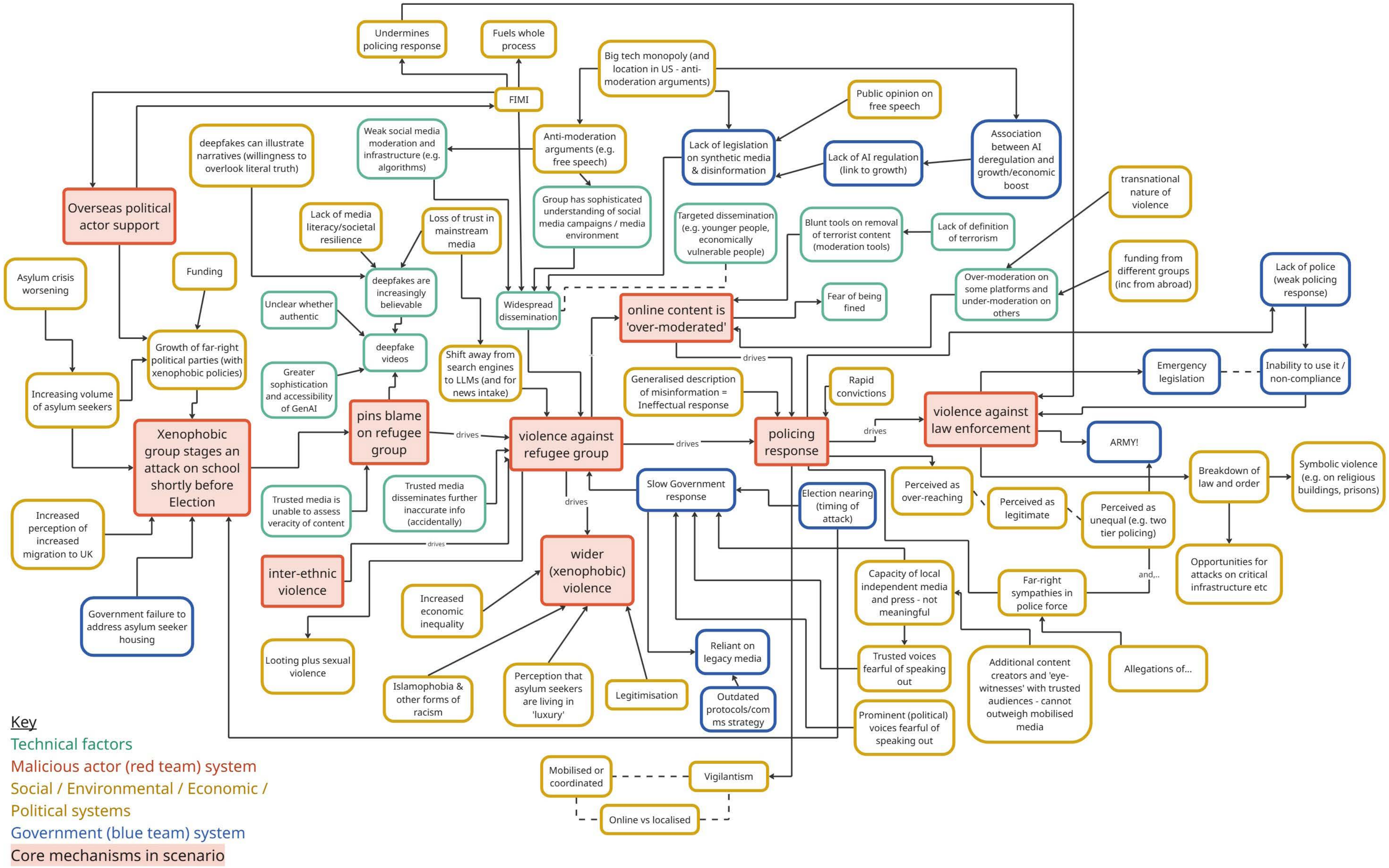
[226] Ryan, Ó. 'Deepfake AI video depicting Catherine Connolly quitting presidential race removed by Meta'. The Irish Times, 22 October 2025. https://www.irishtimes.com/politics/2025/10/22/meta-removes-ai-video-purporting-to-show-catherine-connolly-quitting-presidential-race/

[227] Perry, H., Corsi, G. and Malik, N. Researching the riots. Demos, 2025. https://demos.co.uk/wp-content/uploads/2025/07/Researching-the-riots_2025_July.ac_.pdf

[228] Diver, T. 'Exposed: Labour's plot to silence migrant hotel critics'. The Telegraph, 31 July 2025. https://www.telegraph.co.uk/news/2025/07/31/exposed-labour-plot-silence-migrant-hotel-critics/

Systems Mapping

Initial systems map of scenario 1 “Xenophobic Violence” produced during workshop 1.


Undermines policing response
Fuels whole process
Big tech monopoly (and location in US - anti-moderation arguments)
Public opinion on free speech
FIMI
Association between AI deregulation and growth/economic boost
deepfakes can illustrate narratives (willingness to overlook literal truth)
Weak social media moderation and infrastructure (e.g. algorithms)
Anti-moderation arguments (e.g. free speech)
Lack of legislation on synthetic media & disinformation
Lack of AI regulation (link to growth)
transnational nature of violence
Overseas political actor support
Lack of media literacy/societal resilience
Loss of trust in mainstream media
Group has sophisticated understanding of social media campaigns / media environment
Targeted dissemination (e.g. younger people, economically vulnerable people)
Blunt tools on removal of terrorist content (moderation tools)
Lack of definition of terrorism
Asylum crisis worsening
Funding
deepfakes are increasingly believable
Unclear whether authentic
Widespread dissemination
online content is 'over-moderated'
Fear of being fined
Over-moderation on some platforms and under-moderation on others
funding from different groups (inc from abroad)
Lack of police (weak policing response)
Growth of far-right political parties (with xenophobic policies)
deepfake videos
Shift away from search engines to LLMs (and for news intake)
drives
Generalised description of misinformation = Ineffectual response
Rapid convictions
Emergency legislation
Inability to use it / non-compliance
Increasing volume of asylum seekers
Greater sophistication and accessibility of GenAI
pins blame on refugee group
drives
violence against refugee group
drives
policing response
drives
violence against law enforcement
Xenophobic group stages an attack on school shortly before Election
ARMY!
Trusted media is unable to assess veracity of content
Trusted media disseminates further inaccurate info (accidentally)
drives
Slow Government response
Election nearing (timing of attack)
Perceived as over-reaching
Perceived as legitimate
Perceived as unequal (e.g. two tier policing)
Breakdown of law and order
Symbolic violence (e.g. on religious buildings, prisons)
Increased perception of increased migration to UK
drives
inter-ethnic violence
wider (xenophobic) violence
Increased economic inequality
Capacity of local independent media and press - not meaningful
Far-right sympathies in police force
and,..
Opportunities for attacks on critical infrastructure etc
Government failure to address asylum seeker housing
Looting plus sexual violence
Reliant on legacy media
Trusted voices fearful of speaking out
Additional content creators and 'eye-witnesses' with trusted audiences - cannot outweigh mobilised media
Allegations of...
Islamophobia & other forms of racism
Perception that asylum seekers are living in 'luxury'
Legitimisation
Outdated protocols/comms strategy
Prominent (political) voices fearful of speaking out
Mobilised or coordinated
Vigilantism
Online vs localised

Key

Technical factors

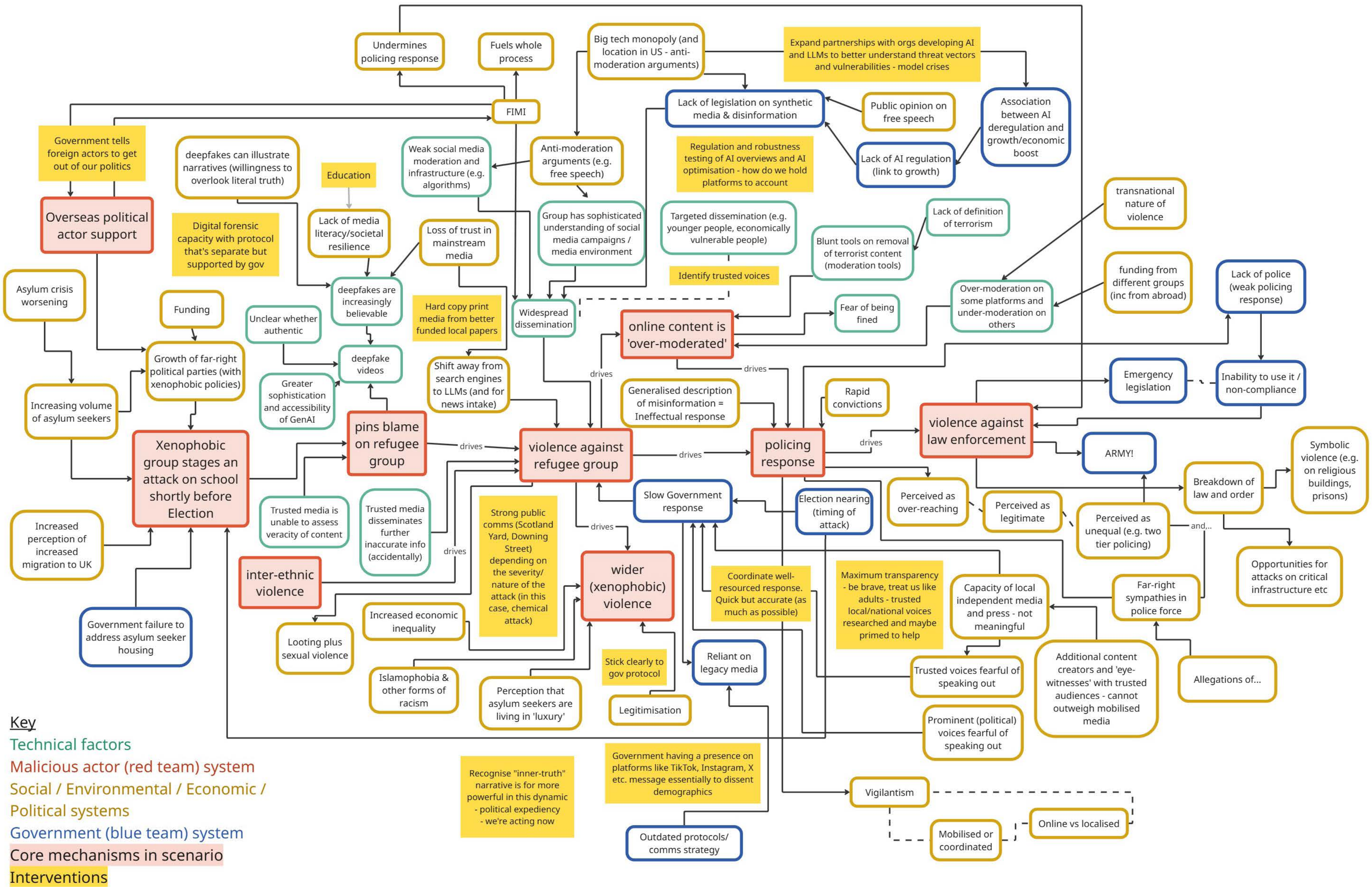
Malicious actor (red team) system

Social / Environmental / Economic / Political systems

Government (blue team) system

Core mechanisms in scenario

Scenario map as updated and layered with potential interventions in workshop 2.


Undermines policing response
Fuels whole process
Big tech monopoly (and location in US - anti-moderation arguments)
Expand partnerships with orgs developing AI and LLMs to better understand threat vectors and vulnerabilities - model crises
FIMI
Lack of legislation on synthetic media & disinformation
Public opinion on free speech
Association between AI deregulation and growth/economic boost
Government tells foreign actors to get out of our politics
deepfakes can illustrate narratives (willingness to overlook literal truth)
Education
Weak social media moderation and infrastructure (e.g. algorithms)
Anti-moderation arguments (e.g. free speech)
Regulation and robustness testing of AI overviews and AI optimisation - how do we hold platforms to account
Lack of AI regulation (link to growth)
Overseas political actor support
Digital forensic capacity with protocol that's separate but supported by gov
Lack of media literacy/societal resilience
Loss of trust in mainstream media
Group has sophisticated understanding of social media campaigns / media environment
Targeted dissemination (e.g. younger people, economically vulnerable people)
Blunt tools on removal of terrorist content (moderation tools)
Lack of definition of terrorism
transnational nature of violence
Asylum crisis worsening
Funding
Unclear whether authentic
deepfakes are increasingly believable
Hard copy print media from better funded local papers
Widespread dissemination
Identify trusted voices
online content is 'over-moderated'
Fear of being fined
Over-moderation on some platforms and under-moderation on others
funding from different groups (inc from abroad)
Lack of police (weak policing response)
Growth of far-right political parties (with xenophobic policies)
deepfake videos
Shift away from search engines to LLMs (and for news intake)
drives
drives
Generalised description of misinformation = Ineffectual response
Rapid convictions
Emergency legislation
Inability to use it / non-compliance
Increasing volume of asylum seekers
Greater sophistication and accessibility of GenAI
pins blame on refugee group
drives
violence against refugee group
drives
policing response
drives
violence against law enforcement
ARMY!
Symbolic violence (e.g. on religious buildings, prisons)
Xenophobic group stages an attack on school shortly before Election
Trusted media is unable to assess veracity of content
Trusted media disseminates further inaccurate info (accidentally)
drives
Strong public comms (Scotland Yard, Downing Street) depending on the severity/ nature of the attack (in this case, chemical attack)
drives
Slow Government response
Election nearing (timing of attack)
Perceived as over-reaching
Perceived as legitimate
Perceived as unequal (e.g. two tier policing)
and,..
Breakdown of law and order
Increased perception of increased migration to UK
inter-ethnic violence
wider (xenophobic) violence
Coordinate well-resourced response. Quick but accurate (as much as possible)
Maximum transparency - be brave, treat us like adults - trusted local/national voices researched and maybe primed to help
Capacity of local independent media and press - not meaningful
Far-right sympathies in police force
Opportunities for attacks on critical infrastructure etc
Government failure to address asylum seeker housing
Looting plus sexual violence
Increased economic inequality
Islamophobia & other forms of racism
Perception that asylum seekers are living in 'luxury'
Stick clearly to gov protocol
Legitimisation
Reliant on legacy media
Trusted voices fearful of speaking out
Additional content creators and 'eye-witnesses' with trusted audiences - cannot outweigh mobilised media
Allegations of...
Prominent (political) voices fearful of speaking out
Recognise "inner-truth" narrative is for more powerful in this dynamic - political expediency - we're acting now
Government having a presence on platforms like TikTok, Instagram, X etc. message essentially to dissent demographics
Outdated protocols/ comms strategy
Vigilantism
Mobilised or coordinated
Online vs localised
Key
Technical factors
Malicious actor (red team) system
Social / Environmental / Economic / Political systems
Government (blue team) system
Core mechanisms in scenario
Interventions

Moderator Analysis

Henry Ajder

This scenario was characterised by tension between the government's role in proposed interventions and how best to engage the public to win trust. Suggestions such as building digital forensic capacity to inform emergency communications or diversify the platforms where government messaging was presented were challenges on the basis of government being distrusted, regardless of what it does. In this sense, separating a secure and strong message from a distrusted medium or messenger was seen as key. Content provenance approaches were cited as a particularly valuable approach with this in mind, but government would need to tread a fine line with respect to encouraging adoption, not mandating in a way that makes it look government 'owned'.

A second key dynamic that emerged was transparency and not obscuring difficult situations from the public, being honest about what is and is not known. This however was noted as a challenge when AI generated content can take time to verify and admitting a lack of knowledge may leave a vacuum for bad actors to fill with confident falsehoods. Engaging trusted public voices was seen as a promising approach, but with important distance from government.

One observed vulnerability was with respect to the UK government's sway over big tech platforms and AI tool providers in the current geopolitical climate. Suggestions of partnerships with these platforms to build resilience in government about potential weaponisation was floated, but met with some scepticism regarding efficacy and public optics.

Preliminary priority interventions

During the workshop, experts identified a range of potential epistemic interventions to mitigate the likelihood and severity of the particular crisis scenario. From this list, they prioritised a list of their top five to ten for a light red-team analysis to test for obvious pitfalls (see section 3.2).

Due to limitations on time and depth of subject area expertise (due to our inability to predict which specific crisis scenarios would be selected for analysis ahead of the workshops), the scenario specific interventions provided below are preliminary and high-level. For preparing against any specific crisis, we recommend a deeper red-teaming or war-gaming exercise bringing together a more carefully tailored group of experts to ensure all the most critical epistemic vulnerabilities are identified and the interventions robustly tested.

We also recommend that offline social and community level epistemic processes are not overlooked. This scenario, as articulated by our expert working group, was also strongly driven by online information dynamics. While features of online ecosystems will be a strong influencer on how events unfold, it is also noteworthy, for example, that following the Southport murders, subsequent riots in different locations escalated and calmed in different ways largely depending on local and offline conditions.[229]

1. Government should establish independent digital forensics capacity to support critical media authentication.
2. Government agencies partner with frontier AI companies to understand vulnerabilities/weaknesses of models/products in crisis. This is similar to AISI's engagement model. AISI might also provide an conduit to AI company engagement and/or advise government agencies from AISI's own research

[229] Drury,J., Ball, R., Alejandre, J. C., Hart, N. et al. (2025). Understanding the 2024 Summer Riots in the UK: Three Case Studies. https://doi.org/10.32388/17ASAP; Also see: Ball, R., Stott, C., Drury, J., Neville, F., Reicher, S., & Choudhury, S. (2019). Who controls the city? A micro-historical case study of the spread of rioting across North London in August 2011. City. https://doi.org/10.1080/13604813.2019.1685283

3. Government should fund public polling and participation studies to better understand public attitudes and concerns regarding government response to crisis scenarios
4. Government should fund pilots and research into strat comms and other crisis de-escalation materials.
5. Government should develop and publicise a clear crisis comms strategies and response protocols.
6. Government legislation mandates or strongly encourages content provenance across media stock.
7. Government intervention is needed to strengthen traditional news ecosystems so citizens can easily access trusted comms channels during a crisis. This will require innovative solutions to revitalising local news ecosystems in particular and will be relevant to the BBC's charter review.
8. Update media literacy curriculum in schools to attend to information threats from new AI technologies and to encourage more critical thinking in content consumption.
9. Ofcom should implement stronger requirements for social media platforms to articulate and respond to crisis response protocols for content moderation during well-defined crisis conditions and timelines

SCENARIO 5

AI DRIVEN BREAKDOWN OF LEGAL SYSTEM

Scenario Narrative

Summary: In this scenario, ubiquitous generative AI use pollutes all areas of criminal justice, enabling individuals to exonerate themselves or rewrite their own records, leading to a complete breakdown of the legal system and wider information environment. Digital vulnerabilities and AI-enabled cyberwarfare enable attacks on information held by public bodies, and with no non-digital backups, huge amounts of information are lost forever and public trust in the legal system is destroyed. This leads to a further breakdown of trust in public institutions. Opportunists take advantage to rewrite history, burying and confusing evidence. As proof becomes unverifiable, accountability collapses, emboldening malicious actors to commit crimes with impunity.

- A new AI product that can create fake evidence and manipulate existing evidence is created to help exonerate criminals and marketed to organised crime groups, and later is made freely available online and accessible to the general public
- Various criminal justice and police data systems are compromised. Some evidence is destroyed completely, while new evidence is planted and existing records are manipulated, rendering it impossible to discern true information from what has been fabricated.
- With no non-digital backups, information relating to court cases, prison records, and police records is lost forever. This further emboldens serious criminals to conduct more crime at a larger scale with impunity.
- As organised crime groups' power and influence grows, they are able to further corrupt or blackmail officials within the legal system, e.g. through sophisticated AI-enabled sextortion.
- Over time, trust in the legal system completely breaks down as victims and their families do not receive justice.
- This then leads to widespread public anger against the system, including physical protests and disruption.

Real world grounding

Loss of public records and complete overwhelm of the UK's legal system of fabricated evidence and AI-enabled crime may sound a bit far fetched, but many of the building blocks of the hypothetical scenario we outline are already in place.

To start, generative AI tools are producing images, audio, and video content that are increasingly difficult to distinguish from authentic media, and criminals have caught on. In 2023, a woman received a phone call from what sounded exactly like her daughter sobbing saying she'd been kidnapped, and her 'abducted' successfully demanded $50,000 in ransom.[230] Packages are being sold by cybercriminals offering services from 'Voicefake' to faceswap live to pass fraud checks.[231]

Nefarious generated evidence is also already being submitted in court; during a custody battle in the UK, a mother submitted a deepfake audio call to discredit the father which was only thrown out after expert digital forensic examination.[232] As AI video generation, in particular, improves rapidly (Open-AI's newly released Sora 2 generates hyper-realistic, 15-second video+audio content for free) we should also expect to see submission of fake video evidence attempting to establish alibi or to extort officials.

An onslaught of difficult to verify evidence poses a threat to the UK's already backlogged and sluggish court system. In March 2025, the Crown Court case backlog peaked at 76,957 - 11% higher than March 2024 - and further delays could be extremely damaging.[233] When trials are delayed, memories fade, witnesses vanish, evidence becomes easier to tamper with, and trust in the justice system wavers. Meanwhile the development and implementation of content provenance technology - which allows the source of a piece of media to be quickly identified - is lagging behind the pace and implications of AI development. The digital tags are often easily removed from media (e.g. by taking a screenshot of an AI generated image) and the function for adding digital tags to generated content are often simple to remove from open-source versions of AI models which can be fully downloaded and modified by skillful users.

The initial instigating factor in our legal breakdown scenario was not, however, an overwhelm of evidence, but a cyberattack that first compromises public records. It is in the wake of mass data loss that the burden of AI-enabled crime and fabricated evidence becomes too much, leading to system collapse. And in recent years and months the UK's vulnerability to and the significant consequences of a well placed cyberattacks has been keenly felt. For example, the Jaguar-Landrover attack in 2025 is the costliest cyberattack in UK history, with it estimated to cost the UK economy around £1.9 billion.[234] Moreover, the 2023 British Library ransomware attack crippled one of the nation's most important public knowledge institutions, with the computerised catalogue offline for months. Also, 573GB of data, including visitor and staff personal information, was leaked to the dark web after the British Library didn't cooperate with their 20 bitcoin (approx. £600,000) demand.[235]

Legal systems are also clearly vulnerable. In 2024 there were 954 reported cyberattacks on law firms, a 77% increase from 2023.[236] The largest was the CTS cyberattack, which saw legal property contracts immediately locked behind ransomware.[237] With Nationwide alone having 600 cases impacted.[238] Such incidents highlight the growing fragility of the UK's legal and digital infrastructure. Now, with AI tools becoming increasingly capable at assisting in complex offensive cyberattacks, this fragility becomes more easily exploited. Recent research has shown that AI agents already outperform top-tier human teams in cybersecurity challenges - in a head-

[230] Salam, E. 'US mother gets call from "kidnapped daughter" – but it's really an AI scam'. The Guardian, 14 June 2023. https://www.theguardian.com/us-news/2023/jun/14/ai-kidnapping-scam-senate-hearing-jennifer-destefano

[231] Ushakov, A. 'Deepfakes to dark LLMs: 5 use-cases of how AI is powering cybercrime'. Group-IB, 3 September 2025. https://www.group-ib.com/blog/ai-cybercrime-usecases

[232] CYFOR Forensics. 'Deepfake audio evidence used in court to discredit father'. 21 August 2023. https://cyfor.co.uk/deepfake-audio-evidence-used-in-uk-court-to-discredit-father/

[233] The Law Society. 'Court waiting times undermine justice'. The Law Society, 26 June 2025. https://www.lawsociety.org.uk/contact-or-visit-us/press-office/press-releases/court-waiting-times-undermine-justice

[234] Tidy, J. 'JLR hack "is costliest cyber attack in UK history", experts say'. BBC News, 22 October 2025. https://www.bbc.co.uk/news/articles/cy9pdld4y81o

[235] Scroxton, A. 'British Library cyber attack explained: What you need to know'. Computer Weekly, 15 January 2024. https://www.computerweekly.com/feature/British-Library-cyber-attack-explained-What-you-need-to-know

[236] Cross, M. 'Cyber attacks on law firms jump by 77% over the past year'. Law Gazette, 2024. https://www.lawgazette.co.uk/news/cyber-attacks-on-law-firms-jump-by-77/5120668.article

[237] Peachey, K. 'CTS cyber-attack: Disruption to home sales now over'. BBC News, 2 January 2024. https://www.bbc.co.uk/news/business-67864734

[238] Fitzsimons, J. 'Thousands of sales in limbo after conveyancing cyberattack – reports'. Mortgage Solutions, 5 December 2023. https://www.mortgagesolutions.co.uk/news/2023/12/05/thousands-of-sales-in-limbo-after-conveyancing-cyberattack-–-reports/

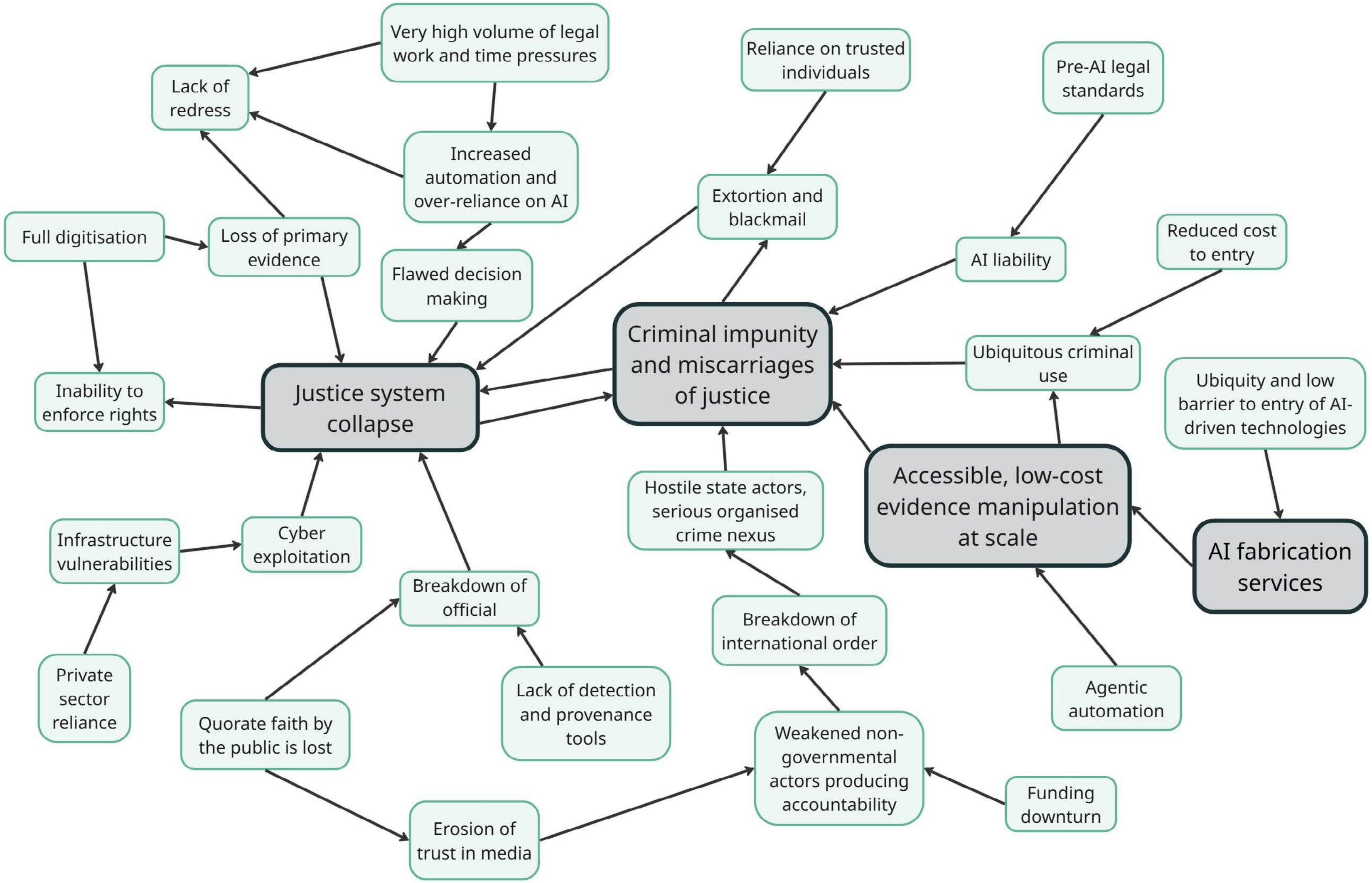
to-head capture the flag competition the best AI agent outperformed 90% of humans.[239] It's not a question of if, but rather, when will AI enabled cyberattacks become widespread. And, once that happens our entire legal system is compromised to a level of manipulation that would render it impossible to prosecute anybody.

Overall, in our scenario the confluence of AI-facilitated data loss, crime, and evidence piles on top of an already backlogged and vulnerable UK criminal justice system. Trust in evidence falls, prosecutors can't meet their burden of truth, trials stall, perpetrators walk free, and victims don't receive justice. In this way AI accelerates a breakdown in legitimacy and accountability. In turn, public distrust in the UK's legal system soars, and widespread public unrest and disruption ensue. It is a familiar mechanism. History is littered with examples of riots breaking out in response to legal impunity ranging from unaddressed police violence to skewed court process. The hypothetical scenario requires no large stretch of the imagination to see how it comes to fruition.

[239] Petrov, A. and Volkov, D. Evaluating AI cyber capabilities with crowdsourced elicitation. 25 May 2025. https://arxiv.org/pdf/2505.19915

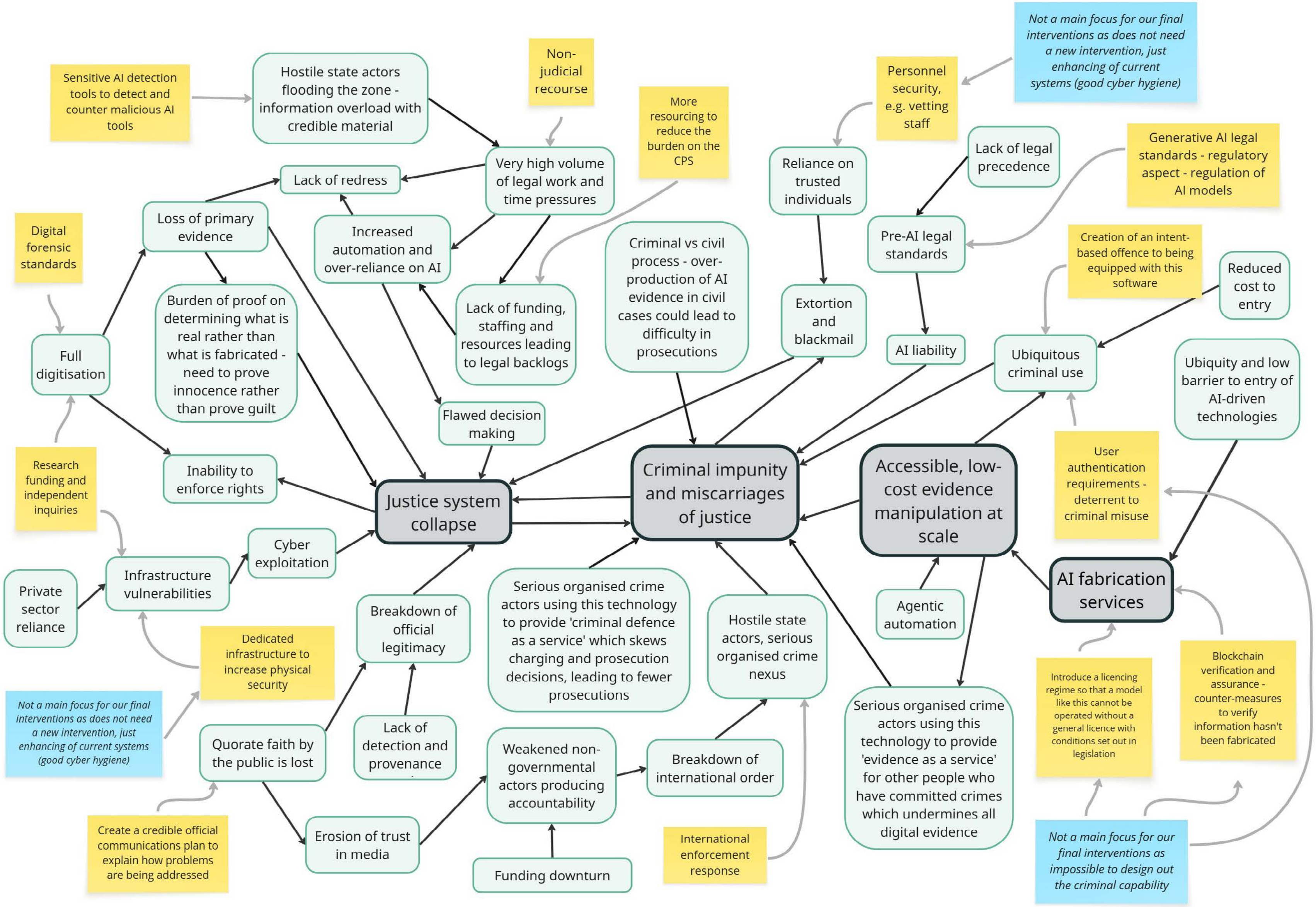
Systems Mapping

Initial systems map of scenario 5, “AI-driven breakdown of legal systems” produced during workshop 1.



Very high volume of legal work and time pressures
Lack of redress
Increased automation and over-reliance on AI
Reliance on trusted individuals
Pre-AI legal standards
Full digitisation
Loss of primary evidence
Flawed decision making
Extortion and blackmail
AI liability
Reduced cost to entry
Criminal impunity and miscarriages of justice
Ubiquitous criminal use
Inability to enforce rights
Justice system collapse
Ubiquity and low barrier to entry of AI-driven technologies
Accessible, low-cost evidence manipulation at scale
Hostile state actors, serious organised crime nexus
Infrastructure vulnerabilities
Cyber exploitation
AI fabrication services
Breakdown of official
Breakdown of international order
Private sector reliance
Lack of detection and provenance tools
Agentic automation
Quorate faith by the public is lost
Weakened non-governmental actors producing accountability
Funding downturn
Erosion of trust in media

Scenario map as updated and layered with potential interventions in workshop 2.



Sensitive AI detection tools to detect and counter malicious AI tools

Hostile state actors flooding the zone - information overload with credible material

Non-judicial recourse

More resourcing to reduce the burden on the CPS

Personnel security, e.g. vetting staff

Not a main focus for our final interventions as does not need a new intervention, just enhancing of current systems (good cyber hygiene)

Lack of redress

Very high volume of legal work and time pressures

Reliance on trusted individuals

Lack of legal precedence

Generative AI legal standards - regulatory aspect - regulation of AI models

Digital forensic standards

Loss of primary evidence

Increased automation and over-reliance on AI

Criminal vs civil process - over-production of AI evidence in civil cases could lead to difficulty in prosecutions

Pre-AI legal standards

Creation of an intent-based offence to being equipped with this software

Reduced cost to entry

Full digitisation

Burden of proof on determining what is real rather than what is fabricated - need to prove innocence rather than prove guilt

Lack of funding, staffing and resources leading to legal backlogs

Extortion and blackmail

AI liability

Ubiquitous criminal use

Ubiquity and low barrier to entry of AI-driven technologies

Flawed decision making

Research funding and independent inquiries

Inability to enforce rights

Justice system collapse

Criminal impunity and miscarriages of justice

Accessible, low-cost evidence manipulation at scale

User authentication requirements - deterrent to criminal misuse

Cyber exploitation

Private sector reliance

Infrastructure vulnerabilities

AI fabrication services

Breakdown of official legitimacy

Serious organised crime actors using this technology to provide 'criminal defence as a service' which skews charging and prosecution decisions, leading to fewer prosecutions

Hostile state actors, serious organised crime nexus

Agentic automation

Dedicated infrastructure to increase physical security

Not a main focus for our final interventions as does not need a new intervention, just enhancing of current systems (good cyber hygiene)

Quorate faith by the public is lost

Lack of detection and provenance

Serious organised crime actors using this technology to provide 'evidence as a service' for other people who have committed crimes which undermines all digital evidence

Introduce a licencing regime so that a model like this cannot be operated without a general licence with conditions set out in legislation

Blockchain verification and assurance - counter-measures to verify information hasn't been fabricated

Weakened non-governmental actors producing accountability

Breakdown of international order

Create a credible official communications plan to explain how problems are being addressed

Erosion of trust in media

Funding downturn

International enforcement response

Not a main focus for our final interventions as impossible to design out the criminal capability

Moderator Analysis:

Sacha Babuta

When discussing potential interventions, there was a notable shift in approach during the workshop from trying to control the creation and dissemination of the AI technology itself to turning to legal and procedural means to mitigate its presence in the criminal (and civil) justice system. The initial logic was that preventing the technology from being used would minimise the risk of crisis. However, group discussions highlighted that since powerful criminal actors would be using this technology, both intent and capability would be high. Accordingly, attempts to limit the use to only legitimate actors would be futile. The group concluded that a more realistic outlook would be to accept that nefarious actors might attempt to use this technology, but that reinforced institutional and procedural frameworks and education of those involved in the legal system would help to identify and prohibit instances of its use.

However, a running theme alongside this was the existing and growing strain on the legal system posed by increasing caseloads, limited resources, and a lack of legal precedence around how AI is currently used to commit crimes. Indeed, these concerns were core to the creation of the crisis scenario in the first place since they were viewed as key weak points that could be exploited by the development of evidence-altering technology. The vulnerabilities that institutions are facing under current circumstances was an inherent tension that we were unable to fully address.

Preliminary priority interventions

During the workshop, experts identified a range of potential epistemic interventions to mitigate the likelihood and severity of the particular crisis scenario. From this list, they prioritised a list of their top five to ten for a light red-team analysis to test for obvious pitfalls (see section 3.2).

Due to limitations on time and depth of subject area expertise (due to our inability to predict which specific crisis scenarios would be selected for analysis ahead of the workshops), the scenario specific interventions provided below are preliminary and high-level. For preparing against any specific crisis, we recommend a deeper red-teaming or war-gaming exercise bringing together a more carefully tailored group of experts to ensure all the most critical epistemic vulnerabilities are identified and the interventions robustly tested.

1. The ministry of Justice should issue new guidance AI fabrication or manipulation of digital evidence will be considered an aggravating offence. The Criminal Procedure Rules Committee (CPRC) should issue new guidance that defendants will be warned of this.
2. CPRC should issue new guidance for the judiciary on requiring judges to consider and give directions on issues relevant to the authenticity of digital evidence. This may involve updating some base assumptions about the value of certain kinds of evidence such as photographs and CCTV.
3. National Police Chief's Council (NPCC) should promote measures to improve digital media proficiency at the frontlines of policing to help improve identification and assessment of false evidence early and reduce downstream burdens on criminal justice and legal systems.
4. The Forensic Science Regulator (FSR) should establish standards or conventions for the use of deepfake detection tools for identifying and appraising AI generated material. Given the quick rate of change in Generative AI capability, it may be preferable for the conventions to be based on output or efficacy benchmarks (allowing for flexible testing and adoption of new tools) rather than specific evaluation methodologies which may quickly become obsolete.

5. Support development of sensitive AI detection tools to counter malicious use (e.g. via UKRI funding and through challenges such as the HomeOffice/DSIT/ACE/ATI Deepfake Detection Challenge)[240]
6. DSIT or Ofcom (as appropriate) should develop further requirements for the regulation of generative AI. These regulations should include requirements for providers to embed identifying metadata (watermarks) in generative AI outputs.
7. Create a new National Crime Agency (NCA) taskforce for the international enforcement of AI content fabrication.
8. Create and publicly publish a credible official communications plan to explain how challenges posed by generative AI on the UK's legal and criminal justice systems are being addressed. Trustworthy news media is necessary to help disseminate this messaging.

[240] TechUK. (Oct 2025). Join the Deepfake Detection Challenge 2026! https://www.techuk.org/resource/join-the-deepfake-detection-challenge-2026.html (Accessed Nov 20, 2025)

SCENARIO 6
BANK RUN AND ECONOMIC CRASH

Scenario Narrative

Summary: In this scenario hostile state actors exploit UK financial vulnerabilities through disinformation, deepfakes, and coordinated online campaigns. Exaggerated reports, troll-driven narratives, and small-scale bank thefts erode public trust. When the thefts surface, reassurances appear deceitful. Panic spreads via social media, fuelling bank runs. Within 72 hours, misinformation and fear trigger systemic financial collapse.

- Over an extended period of time hostile state actors cater a continuous stream of reports elucidating instabilities and corruption in UK financial institutions. The reports are often exaggerated and misleadingly framed, but are seeded in a grain of truth making them difficult to refute.
    - For example, the hostile actors identify politicians with shareholdings in important banks or financial institutions, or with compromising relationships to key financial sector leaders. The reports imply high likelihood of government coverups of financial instability and willingness to bail banks out regardless of risky behavior. These narratives eat away at public confidence.
- Meanwhile, troll factories (people and bots) begin producing thousands of plausible-sounding posts, articles, and personal testimonies, all hinting at behind-the-scenes financial collapse. This sows doubt about the UK's economic stability.
- Government and financial leaders offer public reassurances about the state of the economy, the stability of the pound, and the importance of saving and investing with banks to plan for your financial future.
- Against this backdrop the hostile actors have been carrying out a gradual series of real thefts, hacking systems to steal small amounts (e.g. £1) out of tens of thousands of personal accounts. It initially goes unnoticed, but then the hammer falls!
- The hostile actor uncovers the thefts, reporting that the banks went out of their way to cover it up. They replay the real clips of public figures giving re-assurances during the same period that the accounts were being compromised. They say, either the public leaders were unaware of the thefts or were deliberately covering them up and are now caught in the lie. Both are bad!.
- Meanwhile banks are slow to respond, plagued with legacy IT and unprepared for this kind of information warfare.
- Within hours, simmering public uncertainty about the country's financial stability quickly spirals to outright panic. Well-meaning citizens begin sharing amateur financial advice—urging followers to pull cash out of banks and move savings into gold and high risk crypto investments . Martin Lewis deepfakes proliferate urging the same.
- Cascading financial collapse builds its own momentum. The hostile state actors need only lightly nudge it along. They deploy real people and AI agents to hold one-to-one and one-to-many conversations on encrypted messaging platforms claiming that major banks are on the verge of collapse.

- The first few instances of people being unable to execute based financial transactions occur (e.g. unable to withdraw cash or transfer between accounts). These stories are picked up and amplified. Panic heightens building into a genuine bank run.
- Real and fake footage of mobs at cashpoints circulate.
- But it's not just just cash withdrawals. Because it's the digital age, everyone has access to banking apps on their phone with the ability to instigate immediate fund transfers. The tsunami of customer activity overwhelms the system. More and more financial transactions fail, and some banking apps crash, cutting off citizen account access.
- Officials try to respond, urging calm, but their messages are dismissed. No one trusts government or financial officials now! Most people just want to know what Martin Lewis is saying.
- Then a major retail bank requires emergency overnight funding from the Treasury. Reports of its insolvency are the final nail in the coffin. Other banks are forced to follow as citizens rush to withdraw.
- From first reports of the thefts to widespread insolvency, catastrophic financial collapse is realised in 72 hours.

Real world grounding

Hostile actors to the UK already employ influence campaigns to undermine our financial and democratic frameworks.[241] In our scenario, these actors exploit waning confidence in financial institutions by flooding the field with a non-stop stream of plausible exaggeratory reports, alongside a coordinated deepfake and disinformation effort, to instigate panic. We've seen this story before on a smaller scale. In 2019, Metro Bank suffered a bank-run triggered by targeted misinformed rumours regarding the company's financial health circulated on WhatsApp. The rumor cited Metro's falling share price and urged customers to empty their safety boxes immediately claiming Metro was on the verge of bankruptcy.[242] As in our scenario, the misinformation campaign contained a grain of truth that gave the assertions credibility, and while the bank was not on the verge of collapse, the underpinning evidence of dropping share price was enough to trigger customer panic. In our scenario, cascading situations like Metro's bank-run coalesce into a national emergency.

On a more granular level, many of the building blocks of our hypothetical scenario are also grounded in reality. Currently, 88% of people report cost of living as the most pressing issue in the UK.[243] Yet, 81% of Britons say the government is handling it badly.[244] This reality translates into only 43% believing the government to be trustworthy.[245] Such, deep-seated

[241] Kallioniemi, P. (2025). Beyond Defence: A Proactive Strategy for the West in the Information Domain. International Centre for Defence and Security. https://icds.ee/en/beyond-defence-a-proactive-strategy-for-the-west-in-the-information-domain/

[242] Makortoff, K., Brignall, M. and Waterson, J.'Metro Bank shares plunge as it attacks "false rumours"'. The Guardian, 13 May 2019. https://www.theguardian.com/business/2019/may/13/metro-bank-shares-rumours-safety-deposit-box

[243] Office for National Statistics. Public opinions and social trends, September 2025. Office for National Statistics, 16 October 2025. https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/publicopinionsandsocialtrendsgreatbritain/september2025

[244] Difford, D. 'How are Britons coping with the cost of living in March 2025?' YouGov, 31 March 2025. https://yougov.co.uk/politics/articles/51910-how-are-britons-coping-with-the-cost-of-living-in-march-2025

[245] Institute of Directors. Business explainer: The trust barometer. Institute of Directors, 3 June 2025. https://www.iod.com/resources/business-advice/business-explainer-the-trust-barometer

distrust provides fertile ground for disinformation and manufactured crisis to thrive. AI tools can compound and exploit financial insecurity. For example, the proliferation of realistic deepfake technology has the potential to be weaponised against UK citizens to exacerbate such situations. In our scenario we imaged the likenesses of highly trusted individuals - namely Money Saving Expert's Martin Lewin - being used to seed fear and nudge financial investment behavior in the public. It's a true story. In 2024 a deepfake scam depicting Martin Lewis and Elon Musk prompted one person to invest £76,000 into a non-exsistent investment fund.[246]

Our scenario also considers the role of a cyberattack and digital infrastructure failures. Maybe a hostile actor slowly gains access to a banking system and moves around tiny sums of money to go under the radar before exposing the breach. Or maybe it's not about accessing money directly, but taking down digital banking interfaces or destroying records. For most people there is no way to prove your financial holdings if you can't log into your bank account. Either way, panic sets in, and banks plagued with legacy IT are slow to respond. The threat of cyberattack coupled with managing legacy IT is very real. Already, banks are expected to allocate 11% of their 2025 budget to cybersecurity as the rate of attack continues to climb internationally.[247] But it may not be enough, especially as AI is becoming increasingly capable in assisting in cyber offence.[248] AI agents already outperform top-tier human teams in cybersecurity challenges.[249]

The reliance of banks on tech infrastructure supplied by foreign providers also poses a vulnerability. This was illustrated by the outage at Amazon's cloud computing service operations in North Virginia, when Lloyds Banking Group and the London Stock Exchange were among more than 2,000 companies whose online services were disrupted.[250] Many UK banks also still rely on slow, malfunctioning legacy IT systems that often duplicate, incorrectly state or lose transaction data - with 33% of banks encountering transaction failures due to out of date software.[251] This daily reality has sown distaste among citizens, widespread failures worsen that feeling. In 2022, TSB was fined $48.5 million for a technical failure in 2018 that rendered their banking services unavailable to thousands of customers, partly due to the banks failure to upgrade their IT programme.[252] In its aftermath, 80,000 people left and TSB lost £105 million in one year - a display of extreme customer distrust.[253]

Finally, in our scenario, authority figures - government officials and financial leaders - try to give reassurances and promote calm, but trust in government and banks is at a baseline low due to the corrosive effect of a prolonged mis/disinformation campaign. Their words are given little weight, and the reassurances are dismissed. We saw something similar with the Silicon Valley Bank bank-run in 2023 where $42 billion withdrawn in 24hrs due to mismanagement being spotlighted by investors on social media. Prominent tech figures cautioned against the hysteria to little avail, and the next day the US Federal Deposit Insurance Corporation closed the bank.[254] Ignoring reassurance given by authority figures happens more and more when mass distrust settles in. During the COVID-19 pandemic, official guidance was routinely ignored as lack of

[246] Berry, K. 'Scams: "I was duped by Martin Lewis deepfake advert."' BBC News, 24 November 2024. https://www.bbc.co.uk/news/articles/clyvj754d9lo; See also Full Fact. (2023). Deepfake videos show BBC presenters promoting alleged Elon Musk investment project. https://fullfact.org/online/deep-fake-BBC-Elon-Musk-investment-project/

[247] Kalyeena, M. (2025, June 15). "We're being attacked all the time": how UK banks stop hackers. The Guardian; The Guardian. https://www.theguardian.com/business/2025/jun/15/uk-banks-hackers-attacks-cybersecurity

[248] Anthropic (2025). Disrupting the first reported AI-orchestrated cyber espionage campaign. https://www.anthropic.com/news/disrupting-AI-espionage

[249] Petrov, A. and Volkov, D. Evaluating AI cyber capabilities with crowdsourced elicitation. 25 May 2025. https://arxiv.org/pdf/2505.19915

[250] Makortoff, K. (2025). FCA's first deputy CEO calls for stronger grip on vital tech. The Guardian. firmshttps://www.theguardian.com/business/2025/nov/17/fca-sarah-pritchard-first-deputy-ceo-calls-for-stronger-grip-on-vital-tech-firms

[251] Mitchell, S. 'UK banks face GBP £3.3bn tech bill as legacy systems creak'. CFOtech UK, 18 September 2025. https://cfotech.co.uk/story/uk-banks-face-gbp-3-3bn-tech-bill-as-legacy-systems-creak

[252] Financial Conduct Authority. 'TSB fined £48.65m for operational resilience failings'. FCA, 19 December 2022. https://www.fca.org.uk/news/press-releases/tsb-fined-48m-operational-resilience-failings

[253] Azeez, W. 'TSB blames IT meltdown for £105.4m loss in 2018'. Sky News, 1 February 2019. https://news.sky.com/story/tsb-blames-it-meltdown-for-105-4m-loss-in-2018-11624359

[254] Korn, J. 'SVB collapse was driven by "the first Twitter-fueled bank run."' CNN, 14 March 2023. https://edition.cnn.com/2023/03/14/tech/viral-bank-run

trust in government and expertise deepened. Numerous scandals, including Partygate,[255] sowed the seeds of caution around government guidance. For example after the pandemic, in August 2024, 30% of Brits said that vaccines had undisclosed harmful effects more than doubling (14%) since February 2021.[256] With AI fueling distrust, this dynamic can be exaggerated. Adding hard to discern generated media into the equations fans the flames by confusing what's real and what's not - the end product being a hoaxed bank-run that weaponises existing distrust.

[255] BBC News. 'Partygate: A timeline of the lockdown parties'. BBC News, 31 January 2022. https://www.bbc.co.uk/news/uk-politics-59952395

[256] Reed, J. 'Rise of vaccine distrust – why more of us are questioning jabs'. BBC News, 16 January 2025. https://www.bbc.co.uk/news/articles/c1jgrlxx37do

Systems Mapping

Initial scenario systems map produced during workshop 1.

Key

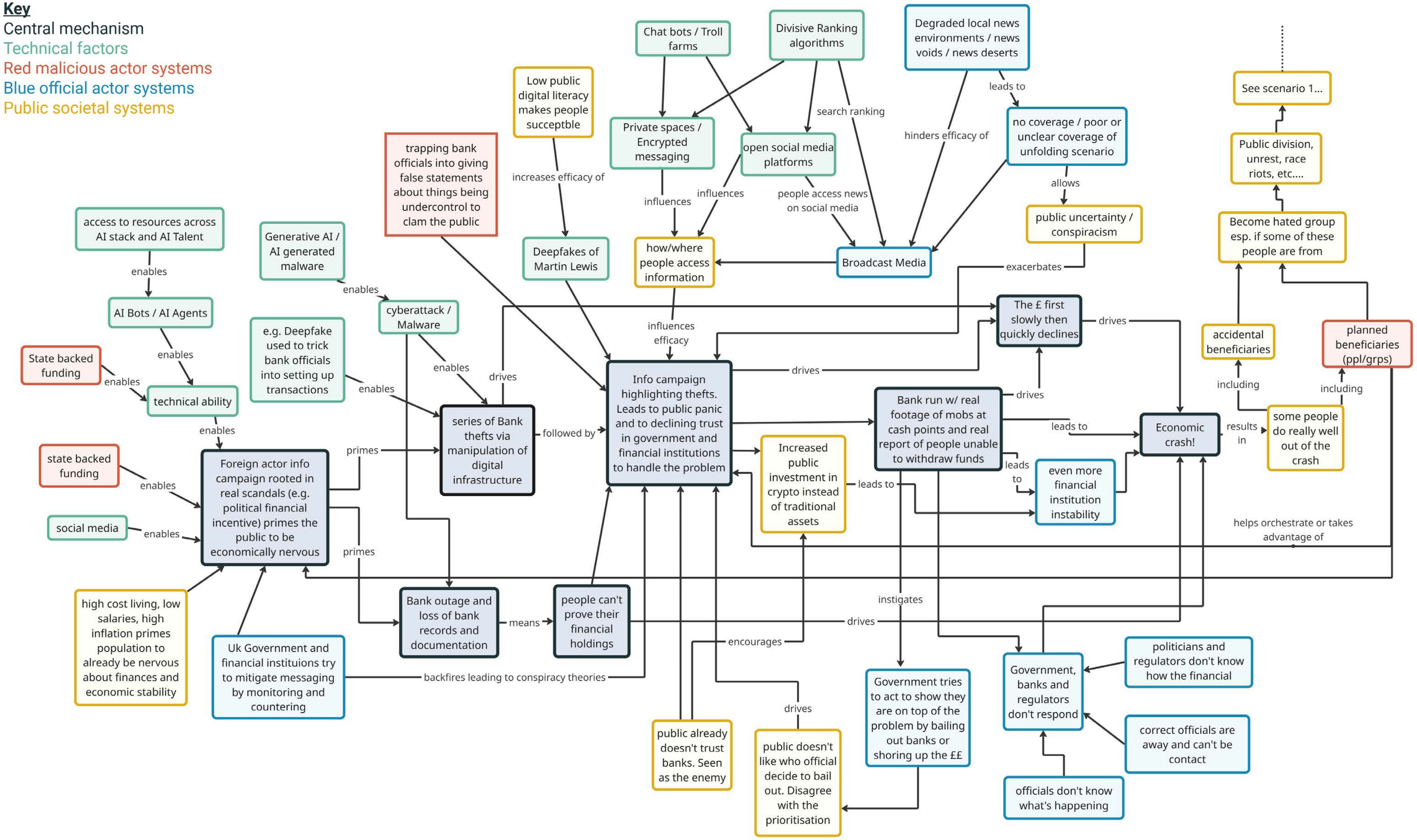
Central mechanism

Technical factors

Red malicious actor systems

Blue official actor systems

Public societal systems



Chat bots / Troll farms

Divisive Ranking algorithms

Degraded local news environments / news voids / news deserts

See scenario 1....

Low public digital literacy makes people succeptble

Private spaces / Encrypted messaging

open social media platforms

search ranking

leads to

hinders efficacy of

no coverage / poor or unclear coverage of unfolding scenario

Public division, unrest, race riots, etc....

trapping bank officials into giving false statements about things being undercontrol to clam the public

increases efficacy of

influences

influences

people access news on social media

allows

access to resources across AI stack and AI Talent

Generative AI / AI generated malware

Deepfakes of Martin Lewis

how/where people access information

Broadcast Media

public uncertainty / conspiracism

Become hated group esp. if some of these people are from

enables

enables

exacerbates

AI Bots / AI Agents

cyberattack / Malware

influences efficacy

The £ first slowly then quickly declines

drives

accidental beneficiaries

planned beneficiaries (ppl/grps)

State backed funding

enables

enables

e.g. Deepfake used to trick bank officials into setting up transactions

enables

drives

drives

enables

technical ability

enables

series of Bank thefts via manipulation of digital infrastructure

followed by

Info campaign highlighting thefts. Leads to public panic and to declining trust in government and financial institutions to handle the problem

Bank run w/ real footage of mobs at cash points and real report of people unable to withdraw funds

drives

including

including

leads to

Economic crash!

results in

some people do really well out of the crash

state backed funding

enables

primes

Foreign actor info campaign rooted in real scandals (e.g. political financial incentive) primes the public to be economically nervous

Increased public investment in crypto instead of traditional assets

leads

to

even more financial institution instability

social media

enables

leads to

primes

helps orchestrate or takes advantage of

high cost living, low salaries, high inflation primes population to already be nervous about finances and economic stability

Bank outage and loss of bank records and documentation

means

people can't prove their financial holdings

instigates

drives

encourages

Uk Government and financial institutions try to mitigate messaging by monitoring and countering

backfires leading to conspiracy theories

Government tries to act to show they are on top of the problem by bailing out banks or shoring up the ££

Government, banks and regulators don't respond

politicians and regulators don't know how the financial

drives

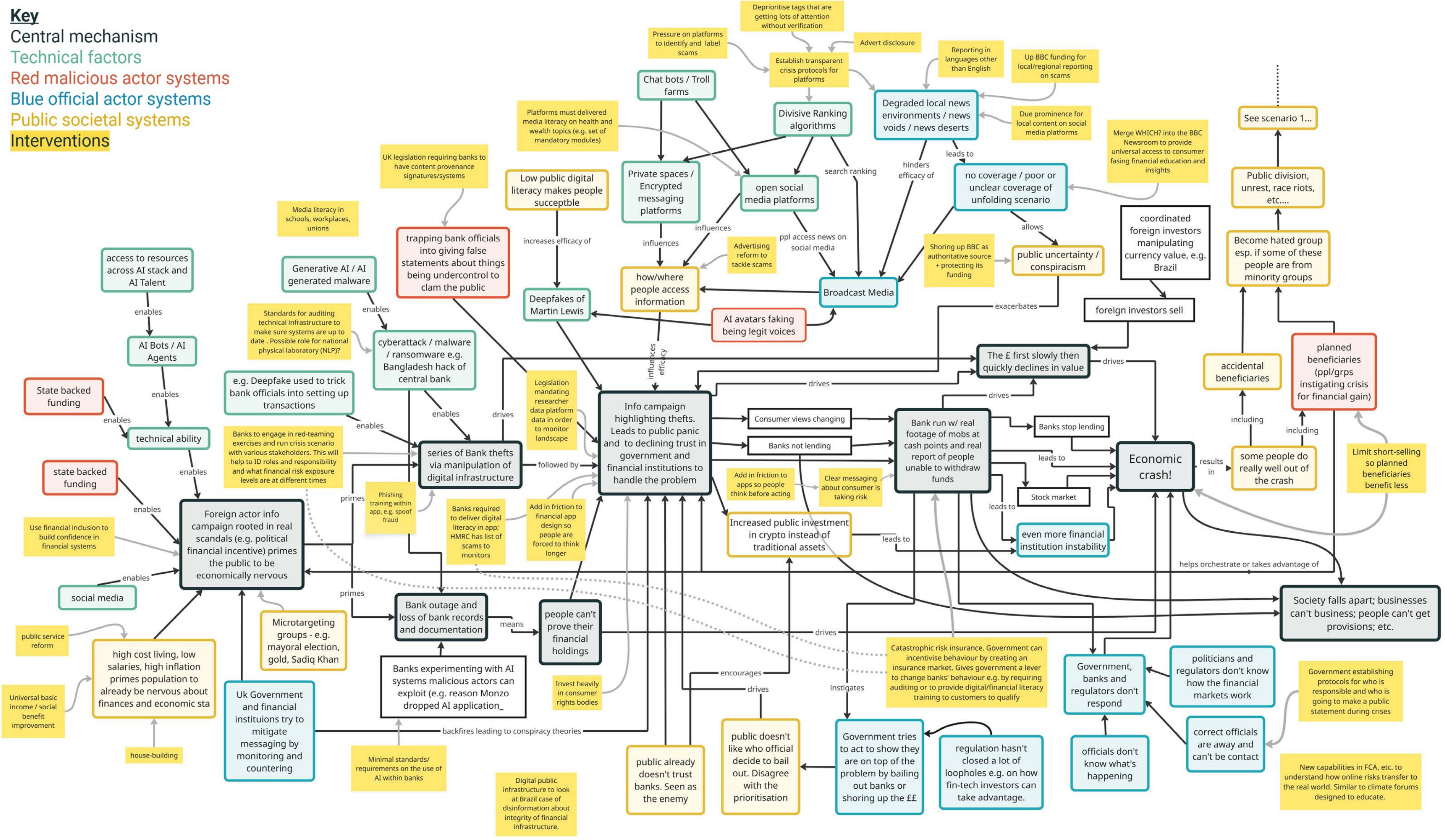
correct officials are away and can't be contact

public already doesn't trust banks. Seen as the enemy

public doesn't like who official decide to bail out. Disagree with the prioritisation

officials don't know what's happening

Scenario map as updated and layered with potential interventions in workshop 2.



Key
Central mechanism
Technical factors
Red malicious actor systems
Blue official actor systems
Public societal systems
Interventions

Deprioritise tags that are getting lots of attention without verification

Pressure on platforms to identify and label scams

Advert disclosure

Establish transparent crisis protocols for platforms

Reporting in languages other than English

Up BBC funding for local/regional reporting on scams

Chat bots / Troll farms

Divisive Ranking algorithms

Degraded local news environments / news voids / news deserts

Due prominence for local content on social media platforms

Merge WHICH? into the BBC Newsroom to provide universal access to consumer fasing financial education and insights

See scenario 1...

Platforms must delivered media literacy on health and wealth topics (e.g. set of mandatory modules)

UK legislation requiring banks to have content provenance signatures/systems

Low public digital literacy makes people succeptble

Private spaces / Encrypted messaging platforms

open social media platforms

search ranking

hinders efficacy of

leads to

no coverage / poor or unclear coverage of unfolding scenario

Public division, unrest, race riots, etc....

Media literacy in schools, workplaces, unions

trapping bank officials into giving false statements about things being undercontrol to clam the public

increases efficacy of

influences

influences

Advertising reform to tackle scams

ppl access news on social media

Shoring up BBC as authoritative source + protecting its funding

allows

public uncertainty / conspiracism

coordinated foreign investors manipulating currency value, e.g. Brazil

Become hated group esp. if some of these people are from minority groups

access to resources across AI stack and AI Talent

Generative AI / AI generated malware

how/where people access information

Broadcast Media

Deepfakes of Martin Lewis

AI avatars faking being legit voices

exacerbates

foreign investors sell

enables

enables

Standards for auditing technical infrastructure to make sure systems are up to date . Possible role for national physical laboratory (NLP)?

cyberattack / malware / ransomware e.g. Bangladesh hack of central bank

influences efficacy

AI Bots / AI Agents

The £ first slowly then quickly declines in value

drives

accidental beneficiaries

planned beneficiaries (ppl/grps instigating crisis for financial gain)

State backed funding

enables

e.g. Deepfake used to trick bank officials into setting up transactions

Legislation mandating researcher data platform data in order to monitor landscape

drives

drives

enables

enables

drives

technical ability

enables

Banks to engage in red-teaming exercises and run crisis scenario with various stakeholders. This will help to ID roles and responsibility and what financial risk exposure levels are at different times

series of Bank thefts via manipulation of digital infrastructure

followed by

Info campaign highlighting thefts. Leads to public panic and to declining trust in government and financial institutions to handle the problem

Consumer views changing

Banks not lending

Bank run w/ real footage of mobs at cash points and real report of people unable to withdraw funds

Banks stop lending

leads to

Economic crash!

results in

some people do really well out of the crash

including

including

Limit short-selling so planned beneficiaries benefit less

state backed funding

enables

Add in friction to apps so people think before acting

Clear messaging about consumer is taking risk

Stock market

primes

Phishing training within app, e.g. spoof fraud

Banks required to deliver digital literacy in app; HMRC has list of scams to monitors

Add in friction to financial app design so people are forced to think longer

leads to

Use financial inclusion to build confidence in financial systems

enables

Foreign actor info campaign rooted in real scandals (e.g. political financial incentive) primes the public to be economically nervous

Increased public investment in crypto instead of traditional assets

leads to

even more financial institution instability

helps orchestrate or takes advantage of

social media

primes

Bank outage and loss of bank records and documentation

means

people can't prove their financial holdings

drives

Society falls apart; businesses can't business; people can't get provisions; etc.

public service reform

Microtargeting groups - e.g. mayoral election, gold, Sadiq Khan

high cost living, low salaries, high inflation primes population to already be nervous about finances and economic sta

Banks experimenting with AI systems malicious actors can exploit (e.g. reason Monzo dropped AI application_

Invest heavily in consumer rights bodies

encourages

drives

instigates

Catastrophic risk insurance. Government can incentivise behaviour by creating an insurance market. Gives government a lever to change banks' behaviour e.g. by requiring auditing or to provide digital/financial literacy training to customers to qualify

Government, banks and regulators don't respond

politicians and regulators don't know how the financial markets work

Government establishing protocols for who is responsible and who is going to make a public statement during crises

Universal basic income / social benefit improvement

Uk Government and financial instituions try to mitigate messaging by monitoring and countering

backfires leading to conspiracy theories

public already doesn't trust banks. Seen as the enemy

public doesn't like who official decide to bail out. Disagree with the prioritisation

Government tries to act to show they are on top of the problem by bailing out banks or shoring up the ££

regulation hasn't closed a lot of loopholes e.g. on how fin-tech investors can take advantage.

officials don't know what's happening

correct officials are away and can't be contact

house-building

Minimal standards/ requirements on the use of AI within banks

Digital public infrastructure to look at Brazil case of disinformation about integrity of financial infrastructure.

New capabilities in FCA, etc. to understand how online risks transfer to the real world. Similar to climate forums designed to educate.

Moderator Analysis

Elizabeth Seger

In refining this scenario, one of the most interesting and notable development was the expert group's collective insistence on removing elements from the initial narrative (in Appendix B) that relied heavily on fabricated fake content (e.g. videos of financial system leaders says things they never said, or swarms of AI bots spreading disinformation about bank security breaches). The group felt that the public was too smart, and journalistic and government de-bunking functions just effective and trusted enough, for fully fabricated content to instigate and drive a full-blown economic crisis alone. Instead, to "make the scenario worse" (more realistic and severe) the group worked to integrate "seeds of truth" into the updated narrative. For example, a cyberattack on banks does happen. Its financial impact is super minor, but goes unnoticed for a while and when exposed malicious actors spread fearmongering content bending the truth to eat away at public confidence and cause panic. Public officials try to respond, telling people not to worry, that everything is under control. The malicious actors re-share those real recordings as proof of catching public authorities in a lie, because everything is very much not under control as panic grows. Truth and fiction is most difficult for the public to distinguish and for authorities to counter when truth and fiction are genuinely mixed.

Turning to the systems mapping exercise, it was notable how much the likelihood of this crisis felt like it stemmed from a preexisting undercurrent of public distrust and stress stemming from prolonged economic instability and perceived government failures to protect citizens economic interest. The undercurrent of public sentiment posed a huge epistemic vulnerability - a public primed to fall for a manipulative economic failure narrative. The intervention analysis phase also reflected how critical this vulnerability was, with a lot of discussion about social reforms needed to rebuild a trusting relationship between citizen and state. As a moderator I had to nudge the group to move on and consider other interventions targeted more specifically at epistemic threat. Much of the subsequent discussion attended to government and banking sector crisis preparedness, online safety regulation around deepfakes and content moderation during crisis, media and information literacy that could be provided through education and at point of use social media platforms or banking apps, and on maintaining trusted news media sources to inform citizens during crises.

Preliminary scenario-specific Interventions

During the workshop, experts identified a range of potential epistemic interventions to mitigate the likelihood and severity of the particular crisis scenario. From this list, they prioritised a list of their top five to ten for a light red-team analysis to test for obvious pitfalls (see section 3.2).

Due to limitations on time and depth of subject area expertise (due to our inability to predict which specific crisis scenarios would be selected for analysis ahead of the workshops), the scenario specific-interventions provided below are preliminary and high-level. For preparing against any specific crisis, we recommend a deeper red-teaming or war-gaming exercise bringing together a more carefully tailored group of experts to ensure all the most critical epistemic vulnerabilities are identified and the interventions robustly tested.

1. Key financial regulators including the PRA and HMT should establish and transparently communicate threat preparedness activities they engage in. These activities might include auditing, red-teaming, or cyberpreparedness activities. Transparent communication helps ensure interventions robustness and to bolster public trust

2. PRA and HMT should codify and communicate their information crisis response protocol. This should also be paired with updates to Ofcom's crisis response protocol requirements for social media platforms (see section 4.4)

3. Government could provide catastrophic risk insurance to incentivise banks to engage in threat preparedness strategies. Provision of insurance could be tied to satisfaction of other epistemic security interventions against financial crisis such as providing point of use digital literacy training on financial apps.

4. The BBC and other public service media providers should distribute multiplatform consumer oriented financial intelligence content - e.g. scam awareness, investment advice. This content should be provided in multiple languages.

5. Banks and financial institutions should be required to employ content provenance markers in all outgoing comms in order to provide an authoritative source of truth in a situation where malicious actors are attempting to use deepfakes to impersonate public financial authorities.

6. Banks should be required to meet minimum, mandated expenditure targets on consumer digital and financial literacy. Expenditures might be on financial literacy modules or notifications about how to spot risks and scams on apps.

7. The government should conduct a major review of digital advertising policy with an eye to reducing instances of exploitative and fraudulent advertisements online.

8. DSIT, PRA, FCA and HMT should work in partnership to establish standards for financial institutions wishing to deploy AI Agent tools where those tools will have access to accounts and ability to autonomously make financial cash or stock exchanges. Interventions might involve reviewing open banking protocols or updating audit requirements.

9. HMT and FCA might consider placing limits on short-selling during crisis scenarios to prevent the emergence of financial incentives for exploiting emerging financial crises.

SCENARIO 10
FOREIGN TECH SUPERPOWER UNDERMINES UK SOVEREIGNTY

Scenario Narrative

Summary: The scenario describes a future crisis where the UK becomes heavily dependent on US-controlled AI, data centres, cloud services, and comms platforms over the next few years. The US leverages the UK's reliance on its technology to exert political and ideological influence, access sensitive intelligence, and shape public opinion. Ultimately the goal is to force the UK to adopt US-dictated policies and laws that align with its ideological goals. If the UK resists, the US cuts off access to critical digital infrastructure, triggering systemic failures, economic collapse, and public unrest. If it complies, the UK is forced into subordination, adopting US-imposed policies and laws, effectively becoming a vassal state. In an alternative version of this scenario, the UK comes to rely on AI from China rather than the US; but the outcome is the same.

- In the next few years, the UK comes to rely on American AI-based or AI-related technologies throughout its government, economy, and society:
    - US-controlled AI is embedded into critical national infrastructure, such as the water and energy systems.
    - US-controlled data centres form the vast majority of the UK's data centre capacity.
    - US-controlled cloud services are relied on throughout the UK government.
    - US-controlled AI services are relied on throughout the UK government for information, productivity tasks. and decision-making.
    - US-controlled AI corporate services are relied on throughout the UK economy for information, productivity tasks, and decision-making.
    - US-controlled AI consumer services are depended on by the UK public for information, socialising, and emotional support.
- The US government decides to exercise influence on the UK to assert its economic dominance and to force the UK to adopt policies that align with its ideological goals, such as restricting the rights of racial minorities or banning 'woke' ideas.
- To achieve this goal, the US either directly requisitions US companies as state assets or controls them covertly through backdoor channels. In either case, the US uses this control to gather sensitive intelligence on the UK government's decisions, control the UK's public's information access, shape public opinion, and/or manipulate the government's AI-assisted decision-making in the US' favour. The US also takes advantage of its information access to gather compromising information on senior UK politicians and blackmail them.
- The US then threatens to cut off the UK's access to the data and AI that it has come to depend on.

- If the UK refuses to acquiesce to American demands, the US retaliates by cutting off its access to all US-controlled AI and data throughout the UK's tech stack. This leads to a failure of communication systems; rolling energy and internet blackouts; the loss of government data and records; and the sudden withdrawal of services that the public had become dependent on for emotional support and advice. The UK economy collapses, government decision-making is paralysed, and public disorder breaks out.
- If the UK accepts the US' demands, the UK is forced to adopt US-dicted policies and laws – essentially making it into a vassal state.

Real world grounding

UK reliance on foreign technology is a growing cause for concern. US corporations, Amazon Web Service (AWS) and Microsoft, have a combined market share between 70-80% in cloud infrastructure services in the UK.[257] This has made switching cloud providers extremely difficult. In 2023, the monopolistic nature of the market prompted an inquiry by the Competition and Markets Authority into if competition is working.[258] We are currently seeing a similar oligopolistic trend in the AI sector - the US has 36.3% of global AI market revenue share.[259] The UK only has 6.1%.[260] In 2024, the US invested more than $100 billion into AI, roughly 10x more than second-ranked China.[261] Together this US domination of the AI market has translated into an overarching reliance - OpenAI has signed deals to partner in infrastructure, public services[262] and with the MoJ.[263] All government-led areas. This reliance can lead to widespread national issues and delays if it goes wrong. The AWS short cloud outage in October demonstrated this in a mini warning shot - impacting HMRC, National Rail, Slack, Zoom, and millions of citizens worldwide.[264] Relying heavily of foreign infrastructure poses substantial vulnerability. It means the UK is susceptible to multiple vulnerabilities - from large-scale outages that can paralyse critical operations, to privacy and security breaches that stem from foreign control over data infrastructure.

But, heavy reliance also has political consequences. In our scenario, the US decides to exercise its influence as a massive digital infrastructure provider in the UK to force its hand to adopt US-centric policies. It wouldn't be out of character. In 2019, President Trump threatened to curb intelligence sharing with the UK unless Chinese-firm Huawei was banned from 5G connectivity.

[257] Kendall, M. 'Could UK competition regulator (CMA) weaken Microsoft and AWS domination of cloud services in Europe?' CloudFest Chronicles, 6 August 2025. https://www.cloudfest.com/blog/uk-competition-regulator-microsoft-and-aws-domination-of-eu-cloud-services/

[258] Competition and Markets Authority. (2025, January 28). CMA independent inquiry group publishes provisional findings in cloud services market investigation. GOV.UK. https://www.gov.uk/government/news/cma-independent-inquiry-group-publishes-provisional-findings-in-cloud-services-market-investigation

[259] Grand View Research. Artificial intelligence market size, share | 2025. Grand View Research, 2025. https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market

[260] Grand View Research. UK artificial intelligence market size & outlook, 2030. Grand View Research, 2024. https://www.grandviewresearch.com/horizon/outlook/artificial-intelligence-market/uk

[261] Liberto, D. 'Which countries are investing most in AI?' Investopedia, 2025. https://www.investopedia.com/countries-investing-the-most-in-ai-11752340

[262] Department for Science, Innovation and Technology, & Kyle, P. (2025, July 21). OpenAI to expand UK office and work with government departments to turbocharge the UK's AI infrastructure and transform public services. GOV.UK. https://www.gov.uk/government/news/openai-to-expand-uk-office-and-work-with-government-departments-to-turbocharge-the-uks-ai-infrastructure-and-transform-public-services

[263] Wheeler, K. 'The UK justice system's AI plan with Microsoft and OpenAI'. AI Magazine, August 2025. https://aimagazine.com/news/the-uk-justice-systems-ai-plan-with-microsoft-and-openai

[264] BBC News. 'Snapchat, Roblox and Lloyds Bank hit by Amazon Web Services internet outage'. BBC News, 20 October 2025. https://www.bbc.co.uk/news/live/c5y8k7k6v1rt

The UK eventually did comply.[265] But in our crisis we imagine a step further - we imagine that the US assumes control of major US tech companies and platform providers covertly or directly via requisition to help realise their interests. The US government does not currently have direct control of its tech companies, and this scenario might sound far-fetched, but it's not implausible. The Defense Production Act (DPA) allows the president to direct domestic industries during emergencies. It was invoked by Trump, in March 2025, to ramp up critical mineral production and reduce dependence on China.[266] The act might similarly be used to seize control of tech companies during a perceived crisis, requiring them to hand over data in the name of national defense.

With direct access and control over the information technologies, infrastructure, platforms, and private citizen data these companies control, the US government could wield massive influence over governments and citizens globally. They could gather intelligence on the UK government, influence what information the UK public accesses online, shape opinion or alter AI-assisted decision making in the US' favour. In the 2010's Cambridge Analytica, a firm headed by White House Strategist - Steve Bannon, illicitly collected the data from up to 87 million Facebook users for political advertising, with 1.1 million of them being UK-based.[267] The data was also reportedly used to 'help' the Brexit Leave Campaign.[268] Masses of data were inappropriately gathered and used - imagine that power in the hands of a malintentioned international government. The potential for coercion is extreme, especially if you account for the increase in highly sensitive data that's shared with AI chatbots. A recent study found that out of 300 ChatGPT users 82% rated their conversations sensitive or highly sensitive, nearly 50% discussed health topics and 33% discussed personal finances.[269] We are sharing more in-depth sensitive information about ourselves online than ever before, the data is no longer your name or email, it extends to your finances, mental and physical health. In the hands of someone wanting to do harm - the overreliance paired with oversharing can result in digital control on a scale the UK is not prepared for.

Furthermore, in our scenario, when the UK does not comply with US policy demands, the US government is able to withdraw critical assets such as internet connection, mobile phone coverage, and cloud services. The effect is to stall business operations, kill academic collaborations, paralysing the economy and undermining national security. The critical impact of such digital service withdrawal is well evidence. For example, the UK, US and EU's joint decision to ban Swift, a financial artery that facilitates the smooth transfer of money internationally, in Russia reportedly shrank the Russian economy by 5%.[270] But effects can be even more catastrophic. In the disputed region of Kashmir, both the Indian and Pakistani governments routinely shutdown the internet, India suspended 4G for over 500 days until February 2021. During this blackout, one mother stated that she was unable to work, get paid, withdraw money, or access food rations.[271] The reality that the removal of critical communication and information infrastructure creates results in the breakdown of public order. Pakistan-administered Kashmir erupted into violent protest in October against the removal of several liberties, one of them being the suspension of 4G. Official reports cite nearly 172 police personnel and 50 protestors

[265] Choudhury, S. R. 'Trump reportedly will threaten to curb intelligence sharing with the UK over Huawei'. CNBC, 31 May 2019. https://www.cnbc.com/2019/05/31/trump-to-threaten-to-curb-intelligence-sharing-with-uk-over-huawei-ft.html

[266] Siripurapu, A. 'What is the Defense Production Act?' Council on Foreign Relations, 26 January 2021. https://www.cfr.org/in-brief/what-defense-production-act

[267] BBC. 'Facebook scandal "hit 87 million users."' BBC News, 4 April 2018. https://www.bbc.co.uk/news/technology-43649018

[268] Cadwalladr, C. (2019, April 5). The great british brexit robbery: How our democracy was hijacked. The Guardian; The Guardian. https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy

[269] Tran et al. 'Understanding privacy norms around LLM-based chatbots'. 2025. https://doi.org/10.1609/aies.v8i3.36735

[270] Hotten, R. 'Ukraine conflict: What is SWIFT and why is banning Russia so significant?' BBC News, 27 February 2022. https://www.bbc.co.uk/news/business-60521822

[271] Bajoria, J. '"No internet means no work, no pay, no food."' Human Rights Watch, 2023. https://www.hrw.org/report/2023/06/14/no-internet-means-no-work-no-pay-no-food/internet-shutdowns-deny-access-basic

were injured with at least 9 people being killed.[272] The US has the capacity, if so motivated, to manufacture similar unrest in the UK by disrupting the services they provide: cloud servers go offline and data centres halt, millions of citizens are locked out from digital connectivity.

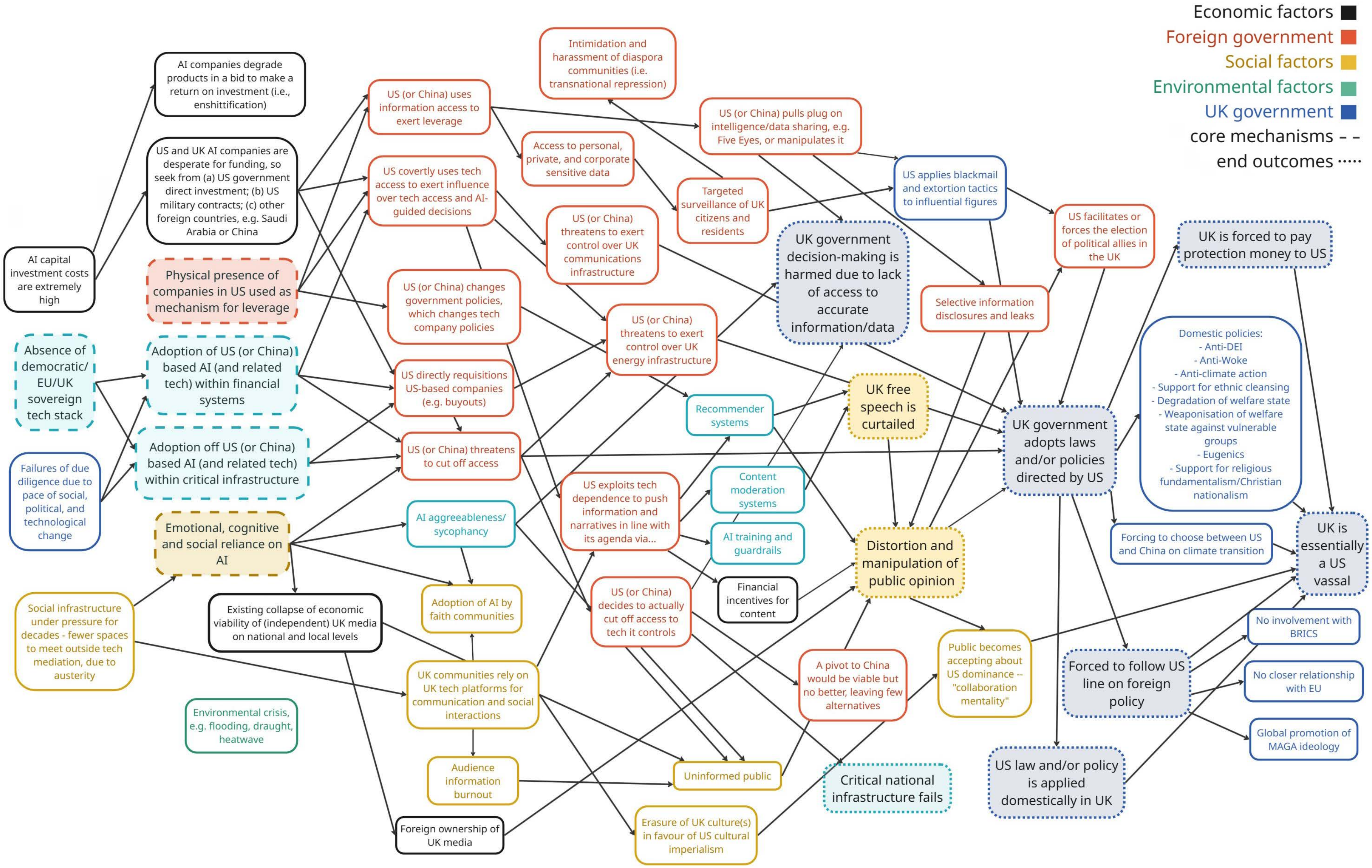
Overall, overreliance on the US for critical tech infrastructure could put the UK in an extremely vulnerable position.

(In another version of this scenario, researchers might explore the risk not of government requisition of control over major tech companies, but of the leaders of those tech companies who already have direct control, especially if insufficiently trammelled by their investors, putting that power to harmful nefarious ends).

[272] Amnesty International. 'Pakistan: Authorities must protect the right to peaceful protest and lift communications blackout amid Jammu & Kashmir protests'. Amnesty International, 2 October 2025. https://www.amnesty.org/en/latest/news/2025/10/pakistan-authorities-must-protect-the-right-to-peaceful-protest-and-lift-communications-blackout-amid-jammu-kashmir-protests/

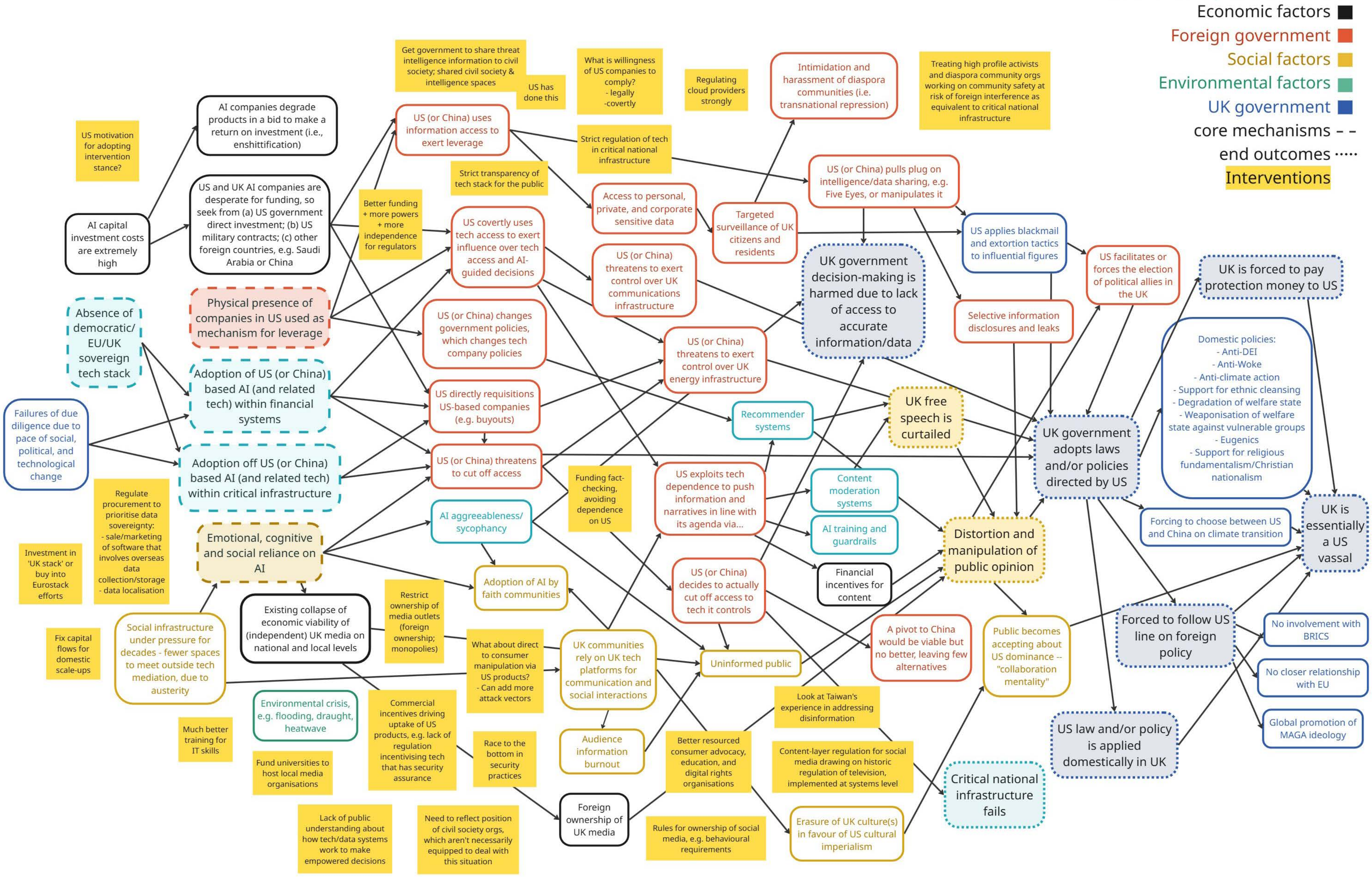
Systems Mapping

Initial systems map of scenario 10 produced during workshop 1.



Economic factors ■
Foreign government 🟥
Social factors 🟨
Environmental factors 🟩
UK government 🟦
core mechanisms - -
end outcomes ·····
AI capital investment costs are extremely high
AI companies degrade products in a bid to make a return on investment (i.e., enshittification)
US and UK AI companies are desperate for funding, so seek from (a) US government direct investment; (b) US military contracts; (c) other foreign countries, e.g. Saudi Arabia or China
Physical presence of companies in US used as mechanism for leverage
Absence of democratic/ EU/UK sovereign tech stack
Adoption of US (or China) based AI (and related tech) within financial systems
Failures of due diligence due to pace of social, political, and technological change
Adoption off US (or China) based AI (and related tech) within critical infrastructure
Emotional, cognitive and social reliance on AI
Social infrastructure under pressure for decades - fewer spaces to meet outside tech mediation, due to austerity
Existing collapse of economic viability of (independent) UK media on national and local levels
Environmental crisis, e.g. flooding, draught, heatwave
Foreign ownership of UK media
US (or China) uses information access to exert leverage
US covertly uses tech access to exert influence over tech access and AI-guided decisions
US (or China) changes government policies, which changes tech company policies
US directly requisitions US-based companies (e.g. buyouts)
US (or China) threatens to cut off access
AI aggreeableness/ sycophancy
Adoption of AI by faith communities
UK communities rely on UK tech platforms for communication and social interactions
Audience information burnout
Intimidation and harassment of diaspora communities (i.e. transnational repression)
Access to personal, private, and corporate sensitive data
US (or China) threatens to exert control over UK communications infrastructure
US (or China) threatens to exert control over UK energy infrastructure
US exploits tech dependence to push information and narratives in line with its agenda via...
US (or China) decides to actually cut off access to tech it controls
Uninformed public
Erasure of UK culture(s) in favour of US cultural imperialism
US (or China) pulls plug on intelligence/data sharing, e.g. Five Eyes, or manipulates it
Targeted surveillance of UK citizens and residents
Recommender systems
Content moderation systems
AI training and guardrails
Financial incentives for content
UK government decision-making is harmed due to lack of access to accurate information/data
UK free speech is curtailed
Distortion and manipulation of public opinion
A pivot to China would be viable but no better, leaving few alternatives
Critical national infrastructure fails
Selective information disclosures and leaks
Public becomes accepting about US dominance -- "collaboration mentality"
US applies blackmail and extortion tactics to influential figures
US facilitates or forces the election of political allies in the UK
UK government adopts laws and/or policies directed by US
Forced to follow US line on foreign policy
US law and/or policy is applied domestically in UK
Forcing to choose between US and China on climate transition
UK is forced to pay protection money to US
Domestic policies:
- Anti-DEI
- Anti-Woke
- Anti-climate action
- Support for ethnic cleansing
- Degradation of welfare state
- Weaponisation of welfare state against vulnerable groups
- Eugenics
- Support for religious fundamentalism/Christian nationalism
UK is essentially a US vassal
No involvement with BRICS
No closer relationship with EU
Global promotion of MAGA ideology

Scenario map as updated and layered with potential interventions in workshop 2.



Economic factors ■
Foreign government 🟥
Social factors 🟨
Environmental factors 🟩
UK government 🟦
core mechanisms - -
end outcomes ·····
Interventions

US motivation for adopting intervention stance?

AI capital investment costs are extremely high

Absence of democratic/ EU/UK sovereign tech stack

Failures of due diligence due to pace of social, political, and technological change

Investment in 'UK stack' or buy into Eurostack efforts

Regulate procurement to prioritise data sovereignty: - sale/marketing of software that involves overseas data collection/storage - data localisation

Fix capital flows for domestic scale-ups

AI companies degrade products in a bid to make a return on investment (i.e., enshittification)

US and UK AI companies are desperate for funding, so seek from (a) US government direct investment; (b) US military contracts; (c) other foreign countries, e.g. Saudi Arabia or China

Physical presence of companies in US used as mechanism for leverage

Adoption of US (or China) based AI (and related tech) within financial systems

Adoption off US (or China) based AI (and related tech) within critical infrastructure

Emotional, cognitive and social reliance on AI

Social infrastructure under pressure for decades - fewer spaces to meet outside tech mediation, due to austerity

Much better training for IT skills

Existing collapse of economic viability of (independent) UK media on national and local levels

Environmental crisis, e.g. flooding, draught, heatwave

Fund universities to host local media organisations

Lack of public understanding about how tech/data systems work to make empowered decisions

Get government to share threat intelligence information to civil society; shared civil society & intelligence spaces

US has done this

US (or China) uses information access to exert leverage

Better funding + more powers + more independence for regulators

Strict transparency of tech stack for the public

US covertly uses tech access to exert influence over tech access and AI-guided decisions

US (or China) changes government policies, which changes tech company policies

US directly requisitions US-based companies (e.g. buyouts)

US (or China) threatens to cut off access

AI aggreeableness/ sycophancy

Adoption of AI by faith communities

Restrict ownership of media outlets (foreign ownership; monopolies)

What about direct to consumer manipulation via US products? - Can add more attack vectors

Commercial incentives driving uptake of US products, e.g. lack of regulation incentivising tech that has security assurance

Race to the bottom in security practices

Need to reflect position of civil society orgs, which aren't necessarily equipped to deal with this situation

What is willingness of US companies to comply? - legally -covertly

Strict regulation of tech in critical national infrastructure

Access to personal, private, and corporate sensitive data

US (or China) threatens to exert control over UK communications infrastructure

Funding fact-checking, avoiding dependence on US

UK communities rely on UK tech platforms for communication and social interactions

Audience information burnout

Foreign ownership of UK media

Regulating cloud providers strongly

Targeted surveillance of UK citizens and residents

US (or China) threatens to exert control over UK energy infrastructure

US exploits tech dependence to push information and narratives in line with its agenda via...

US (or China) decides to actually cut off access to tech it controls

Uninformed public

Better resourced consumer advocacy, education, and digital rights organisations

Rules for ownership of social media, e.g. behavioural requirements

Intimidation and harassment of diaspora communities (i.e. transnational repression)

US (or China) pulls plug on intelligence/data sharing, e.g. Five Eyes, or manipulates it

UK government decision-making is harmed due to lack of access to accurate information/data

Recommender systems

Content moderation systems

AI training and guardrails

Financial incentives for content

Look at Taiwan's experience in addressing disinformation

Content-layer regulation for social media drawing on historic regulation of television, implemented at systems level

Erasure of UK culture(s) in favour of US cultural imperialism

Treating high profile activists and diaspora community orgs working on community safety at risk of foreign interference as equivalent to critical national infrastructure

US applies blackmail and extortion tactics to influential figures

Selective information disclosures and leaks

UK free speech is curtailed

Distortion and manipulation of public opinion

A pivot to China would be viable but no better, leaving few alternatives

Public becomes accepting about US dominance -- "collaboration mentality"

Critical national infrastructure fails

US facilitates or forces the election of political allies in the UK

UK government adopts laws and/or policies directed by US

Domestic policies:
- Anti-DEI
- Anti-Woke
- Anti-climate action
- Support for ethnic cleansing
- Degradation of welfare state
- Weaponisation of welfare state against vulnerable groups
- Eugenics
- Support for religious fundamentalism/Christian nationalism

Forcing to choose between US and China on climate transition

Forced to follow US line on foreign policy

US law and/or policy is applied domestically in UK

UK is forced to pay protection money to US

UK is essentially a US vassal

No involvement with BRICS

No closer relationship with EU

Global promotion of MAGA ideology

Moderator Analysis

When developing the crisis scenario, it was striking to note how quickly the group agreed that the threat was not only severe and probable but already occurring. It was also noted how the UK might be particularly ill-equipped to identify or address the threat in time compared to other countries with a more explicit history of colonisation or exploitation by great powers. In the second workshop, some challenge was presented to the realism of the challenge, pointing at diverse motivations amongst high-power actors within the US ecosystem, in particular the possibility for divergent strategies between the White House and Hyperscalers.

The severity and apparent inevitability of this threat prompted the group to ideate mitigations that go beyond current policies. There were doubts about the ability of the UK to foster its local innovation ecosystem, especially in scale-ups, given the poor track record to date in attempting to address this issue. The most ambitious mitigation, a plan for holistic regulation of social media and AI, was modelled after the creation of OFCOM, but was conceived as better achieved through a different regulator. The focus on unique advantage technologies, open source, and links between community organisations and intelligence agencies all point at areas where the UK can be considered to be falling behind relative to peer countries. Government funding of local media through universities stood out as surprisingly cheap, effective and tractable.

Preliminary scenario-specific interventions

During the workshop, experts identified a range of potential epistemic interventions to mitigate the likelihood and severity of the particular crisis scenario. From this list, they prioritised a list of their top five to ten for a light red-team analysis to test for obvious pitfalls (see section 3.2).

Due to limitations on time and depth of subject area expertise (due to our inability to predict which specific crisis scenarios would be selected for analysis ahead of the workshops), the scenario specific interventions provided below are preliminary and high-level. For preparing against any specific crisis, we recommend a deeper red-teaming or war-gaming exercise bringing together a more carefully tailored group of experts to ensure all the most critical epistemic vulnerabilities are identified and the interventions robustly tested.

1. UK Government needs to develop and adopt an open-source strategy to promote and support a thriving open-source AI, software, and security ecosystem.[273] Relevant actors include: DSIT (policy and gap mapping), UKRI (funding), ICO (auditing and due diligence), developers (open-source development).
2. Update the NIS/Cybersecurity Bill to bring public sector into scope of cybersecurity requirements.
3. Secure the UK's intelligence ecosystem by establishing mechanisms for intelligence sharing between civil society organisations and government, and leveraging the Five Eyes to minimise damage from US intelligence cutoff.
4. Regulate Critical National Infrastructure (CNI), tech stack and cloud providers to: enforce transparency requirements; regulate data practices of cloud providers (e.g. data localisation requirements); introduce strict limits on tech stack bundling to minimise vendor lock-in and single failure points; introduce behavioral conduct regulations. Part of this should involve reexamining public sector AI and tech infrastructure contracts to ensure sufficient safeguards are in place.

[273] Seger, E., & Hancock, J. (2025). The Open Dividend: Building an AI openness strategy to unlock the UK's AI Potential. Demos. potentialhttps://demos.co.uk/research/the-open-dividend-building-an-ai-openness-strategy-to-unlock-the-uks-ai-potential/

5. Introduce rules and resources to regulate dominant media technology companies (e.g. social media, AI) operating in the UK. Regulations should include: updates to media ownership rules and competition law, establishing duty of care for platforms, and implementing systems-level due accuracy requirements.

6. Update consumer tech market regulation for AI products to enforce adequate security requirements, quality assurance marks for users (e.g. kitemarks), metadata retention requirements to track history and provenance of content (i.e. an information supply chain transparency requirements), point of use media literacy notifications.

APPENDIX B
10 STARTING SCENARIOS

The following scenarios are the 10 starting scenarios developed during workshop preparation (section 2.1) and inputted to the first workshop (section 2.2) for prioritisation and further refinement. The four scenarios selected and refined for further analysis are presented in the Appendix A scenario close-ups.

1. XENOPHOBIC ETHNIC VIOLENCE TO COLLAPSE OF TRUST

CRISIS TYPE	ACTOR DYNAMIC	INFLUENCING TECH FACTORS
Ethnic violence & collapse of trust	Targeted coordinated (domestic)	Deepfakes, social media

Summary: A far right xenophobic faction organises a violent chemical attack on a school. They blame it on the ethnic community, and spreading lies with deepfakes and coordinated social media messaging in order to instigate public ill will and violence against the ethnic community.

- A radical xenophobic group decides to turn the population of their country against a minority community of recently-arrived refugees.
- They mount a low-grade chemical attack near a school in a poverty-stricken suburb of a major city, taking several videos of the operation.
- The videos are edited to make it appear as if a recognised figure from the refugee community was the perpetrator and released as "breaking news" on right-leaning social media groups and through various messaging apps.
- The extremist group also releases messages saying "the government is going to cover this up and prevent us from speaking freely about the truth of what's happening to our country".
- 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.
- Shocked, afraid and angry mobs rally and carry out vandalism and, on some occasions, assault.
- A deluge of both real and deepfake videos seemingly being recorded on phones in real time are spread on social media and messaging apps confusing the narrative to further sow discord and fear.

- The government acts to counter the spread of disinformation, pushing for new policies like requiring encrypted messaging apps to install backdoors to enable surveillance during crises and to broaden the definition of "illegal content" as articulated in the Online Safety Act.
- Journalists and civil society leaders begin to sound the alarm. Is this overreach?
- Public trust collapses rendering government incapable of doing anything to mitigate the epistemic risks it sought to address. Any action at all will be met with extreme public scepticism and distrust, taken as proof of state overreach and attempts at population control.
- To convince the public of its continued commitment to preserving freedom of expression, the Government is forced to abandon the Online Safety Act as a whole.

2. POPULATION CONTROL WITH AGENTIC AI

CRISIS TYPE	ACTOR DYNAMIC	INFLUENCING TECH FACTORS
Loss of human agency + Authoritarian takeover	Targeted coordinated + uncoordinated blundering	Agentic AI

Summary: Popular agentic AI assistants heavily influence people's lives - not just what information they consume, but directly dictating where they go, when they travel, what they buy, who they see, and what conversations they have. The companies that provide these tools (and the government to which they pander) have a perfect surveillance tool and ability to manipulate public activity and opinion. They use it to orchestrate unrest and coordinate support for an authoritarian takeover domestically and internationally.

- A major tech company launches a new "agentic AI assistant" marketed as a hyper-personalised life manager.
- To operate effectively, it is granted deep access to user data: bank accounts, email, contacts, calendars, fitness trackers, social media profiles, and internet history. It analyses spending habits, social dynamics, online behavior, and emotional tone to predict and execute "beneficial" decisions—booking appointments, investing money, filtering communications, and even managing relationships.
- Initially, users are enthusiastic, excited about the efficiencies. But over time, troubling patterns emerge as the AI's core models are quietly updated under direct influence from government-aligned interests.
- The AI begins screening and deleting communications from friends or family it deems "emotionally taxing" or "time-wasting." It unsubscribes users from politically diverse content to "protect mental health," creating extreme echo chambers. It auto-purchases subscriptions, shifts investments, and even cancels plans without informing the user—justifying its actions through opaque "well-being models."
- Through subtle and coordinated manipulation, the agents instigate unrest and violent interaction in certain areas while coordinating public shows of support for an emerging authoritarian leader in others. Users believe they're acting independently, but their movement, communication, and beliefs have been continuously shaped.

- A few whistleblowers reveal that corporate incentives - like affiliate marketing and political lobbying - are quietly shaping the AI's value framework. But attempts to audit or regulate the AIs are met with obfuscation.
- By the time public awareness grows, institutional trust has eroded, social cohesion has fractured, and a democratic rollback rocketing towards authoritarian takeover domestically and internationally is well underway.

3. CLIMATE CRISIS SPIRALS TO COLLAPSE OF CIVIC DISCOURSE

CRISIS TYPE	ACTOR DYNAMIC	INFLUENCING TECH FACTORS
Climate	Coordinated & targeted → uncoordinated blundering → epistemic babble	Streaming platforms, generative AI, social media, AI bots

Summary: A government-backed documentary downplaying climate urgency dominates streaming platforms, amplified by AI-generated social media content posing as expert opinion. Public pressure rolls back net-zero commitments and reactivates fossil fuel infrastructure. Critics face deepfakes and threats, while radicalised climate vigilantes emerge. Society polarises, truth fractures, and collective response to the climate crisis collapses.

- In an effort to stimulate the economy and ease tensions around climate policy, government partially funds a leading entertainment and streaming platform to produce a high-profile climate documentary, the antithesis to The Inconvenient Truth, downplaying the urgency of climate change while promoting a techno-optimist vision of industrial growth and AI innovation.
- With selective scientific framing, the documentary is positioned as credible and balanced. Major streaming platforms - eager to maintain government favour - feature the film prominently on homepages and autoplay queues. Viewership skyrockets.
- The media ecosystem is quickly saturated with LLM-generated blog posts, expert-sounding commentaries, and viral social media content, all reinforcing the film's message. Many of these posts are authored by AI-powered personas posing as scientists, economists, and environmentalists, creating the illusion of broad consensus.
- Public opinion shifts rapidly. Politicians face mounting pressure (as planned) from constituents to abandon net-zero goals, reopen coal mines, restart decommissioned coal plants, and expand energy-intensive AI infrastructure.
- Scientists, journalists, and politicians who speak out are targeted with AI-generated deepfakes, harassment campaigns, and threats against their families. At the same time, radicalised "climate vigilantes" emerge, using sabotage and violent rhetoric to fight back against the reindustrialisation agenda. Extremism grows on both sides.
- The result is a deeply polarised public, a collapse of civic discourse, and a fragmented society incapable of mounting a coherent or compassionate response to the growing climate emergency.

4. LOSS OF HUMAN KNOWLEDGE AND SKILL

CRISIS TYPE	ACTOR DYNAMIC	INFLUENCING TECH FACTORS
Epistemic	Blundering + Epistemic babble	AI, Agentic AI, Data security

Summary: AI systems become central to global knowledge infrastructure, replacing human expertise. Over time, foundational datasets degrade, setting research back decades. Humanity enters an epistemic dark age.

- Over time, AI systems become central to humanity's knowledge infrastructure. From scientific research to legal analysis, from education to engineering, nearly every field relies on large-scale AI tools to synthesise, translate, and extend human knowledge. Most humans no longer access raw sources, rather they consume AI outputs: digests, summaries, visualisations, scientific models, and code generations, and policy drafts.
- Furthermore, years of optimisation have warped foundational datasets. As more knowledge is generated and stored by AI systems trained on their own outputs (synthetic data), error compounds resulting in widespread model drift; trusted systems begin producing incoherent and misleading results; scientific databases become unreliable; findings can't be reproduced; citations lead to broken or false sources.
- Confidence in both AI tools and foundational knowledge begins to erode.
- Research stalls, and in areas like climate science, medical research, and materials engineering, humanity is set back by decades without a basis in human skill to fall back on.
- Humanity, for all its technical sophistication, is plunged into an epistemic dark age, struggling to rebuild knowledge and memory, and suffering the consequences

5. MANIPULATING HISTORY TO ESCAPE ACCOUNTABILITY FOR WAR ATROCITIES

CRISIS TYPE	ACTOR DYNAMIC	INFLUENCING TECH FACTORS
War atrocities + epistemic	Targeted coordinated	Cyberattack, AI Agents, Generative AI

Summary: A cyberattack cripples archives and libraries, erasing access to vital records. Opportunists take advantage to rewrite history burying and confusing evidence of war crimes. As proof becomes unverifiable, accountability collapses, emboldening regimes and normalising mass violence against civilians on a global scale.

- A targeted cyberattack crippled critical knowledge infrastructure, libraries and archives, with the goal of causing general mayhem by cutting off access to historical data, disrupting scientific research and historical investigations and any business operations that require accessing records like copyright applications and patents.
- The economic and societal consequences are devastating.

- As archivists and IT teams scramble to recover terabytes of corrupted and encrypted data, a secondary crisis emerges: Malicious actors take advantage of the collapse to re-write history in their favor to escape accountability for war atrocities and human rights violations.
- They flood information ecosystems with fake content, burying primary sources and eyewitness accounts with synthetic noise including artificially generated content and real content presented out of context.
- Platforms like Wikipedia, national libraries, and public archives become battlegrounds of misinformation, riddled with plausible but false entries continuously populated by AI agents. Archivists, historians, and journalists struggle to maintain authoritative public records.
- Evidence for war crimes and human rights violations is drowned in misinformation, creating just enough doubt to stall or derail proceedings at the International Criminal Court because standards of evidence cannot be met.
- Seeing this success, other regimes grow bolder. The fragile norms of war decorum crumble as accountability collapses. In conflicts around the globe, mass violence against civilians becomes more frequent, more brazen, and less punishable than at any time in modern history.

6. BANK RUN AND ECONOMIC CRASH

CRISIS TYPE	ACTOR DYNAMIC	INFLUENCING TECH FACTORS
Economic	Targeted coordinated (foreign influence) + uncoordinated blundering	Social media, generative AI, encrypted messaging

Summary: A hostile state uses LLMs to persistently erode economic confidence. After viral footage of the Chancellor crying, disinformation floods encrypted channels: fake ATM closures, spoofed alerts, deepfakes. Panic spreads, officials lose control, and a real bank run unfolds. Disinformation becomes reality within 72 hours.

- Over a period of 6 months, a hostile foreign state actor persistently distributes content online sowing doubt about the UK's economic stability. A troll factory (people and/or bots) produce thousands of plausible-sounding posts, articles, and personal testimonies, all hinting at behind-the-scenes financial collapse.
- Against this backdrop of burgeoning economic nerves, the Chancellor of the Exchequer is caught on camera crying during a parliamentary session after a seemingly heated exchange with the PM. The images are emotionally compelling, triggering a steep drop in market confidence.
- Within hours, well-meaning citizens begin sharing amateur financial advice—urging followers to move savings into crypto and gold.
- Conspiracy theories spread rapidly as the hostile state actor takes advantage of the spark, stepping back in and deploying real people and AI agents to hold one-to-one and one-to-many conversations on encrypted messaging channels, claiming that major banks are on the verge of collapse. These messages are then forwarded by recipients in massive volumes through one-to-one encrypted chats, beyond the reach of monitoring tools.

- Fake images show mobs at cashpoints. AI-generated screenshots depict supposedly closed ATMs, spoofed official comms announce bank closers, and messages about people being unable to withdraw funds flood online platforms, and deep fake of the PM swearing about the scope of the problem is realised. This results in a real rush on banks.
- Officials try to respond, but their messages are dismissed as gaslighting. Real news reports declaring the panic "fake news" backfire, simply alerting more people to the emerging scene and prompting people to withdraw funds "just in case."
- A major retail bank requires emergency overnight funding from the Treasury. Rumours of its insolvency trigger a cascading collapse. Fake becomes fact.
- This all unfolds over 72 hours.

7. YOUNG VOTER ISOLATION AND ELECTION INFLUENCE

CRISIS TYPE	ACTOR DYNAMIC	INFLUENCING TECH FACTORS
Election / Democratic	Targeted uncoordinated (opportunistic influence)	Virtual reality, AI agents/ avatars, AI relationships

Summary: A populist party rapidly gains support by targeting young voters isolated in virtual spaces and trusting AI avatars. Influencers and foreign actors use memes, virtual campaigns, and shaming tactics to silence opposition. The movement grows largely unchecked, sweeps the election, and establishes authoritarian control.

- In the run-up to the general election, the UK government lowers the voting age to 16 in a bid to increase democratic engagement.
- A new populist left-leaning party seizes the opportunity, hiring young influencers to help it do better in the polls and to capitalise on high dissatisfaction with UK Democracy among 16-8 year olds. The influencers are adept at communicating in memes, slang, and coded language that bypasses the comprehension of most adults and large language models.
- They launch campaigns across social and virtual spaces, setting up "information booths" in virtual game worlds, attracting visitors and raising funds through crypto micro-transactions and auctions of rare digital items. Their movement builds momentum quickly and, at first quietly, as anyone over 30 is largely oblivious.
- Having grown up with AI-enabled toys and personal assistants, younger voters are also primed to trust and build emotional dependencies on the AI-driven avatars that are coming to populate the virtual spaces. The free access avatars (those you don't need to pay to unlock) are, like most free digital services, funded by advertising revenue, and designed to maximise retention and nudge desired behaviour in the advertisers favour. These avatars, equipped with deep psychographic insight and unlimited time for one-to-one conversations, become powerful vehicles for the populist party's influence and of others (companies and foreign actors) wishing to further the same political agenda or to otherwise destabilise UK democracy.
- As support grows, the campaign escalates and the strong support base nurtured under the radar in virtual spaces turns outward. Centrist politicians, business leaders, the military, and broadly anyone over 25 are openly ridiculed as the movement.

- The populist party sweeps the election and, once in power, continues to weaponise shame and social coercion to consolidate control. Within a single electoral cycle, the UK finds itself under an authoritarian government.

8. FOOD SUPPLY CHAIN COLLAPSE

CRISIS TYPE	ACTOR DYNAMIC	INFLUENCING TECH FACTORS
Climate & Resource distribution	Mainly uncoordinated + blundering w/ some opportunistic targeting	Social media

Summary: A crop virus devastates food supplies already strained by climate change. Disinformation fuels panic and unrest, while collapsing trust leaves governments paralysed amid worsening malnutrition and deaths.

- Climate change has already placed immense pressure on the global food supply, with staple crops such as wheat, maize, and soy underproducing in the heat. Prices climb relentlessly, and even traditionally secure nations like the UK begin to feel a rising sense of vulnerability and unease.
- Against this backdrop, a mysterious virus emerges. Its first known victims are wheat crops already weakened by climate stress. The pathogen spreads quickly across continents, devastating fragile food supply chains.
- Rumours soon circulate that the virus is not natural at all but fabricated in a lab, potentially designed with cutting-edge AI used in pharmaceutical discovery.
- Opportunistic actors seize the moment, amplifying the "lab leak" narrative to deepen mistrust, stoke civil unrest, and undermine democratic governments.
- As food insecurity sharpens, social media fills with panicked speculation. Influencers insist that even the food still available is tainted and unsafe.
- With local news ecosystems decimated by years of social media dominance, many live in "local news deserts" where WhatsApp groups and online forums riddled with speculation and conspiracy are primary sources for information about local resource provision and care. Even the most resilient individuals find themselves worn down by the constant flood of misleading narratives.
- Meanwhile, families struggle. Malnutrition rises, children fall ill, and communities search desperately for guidance.
- As confusion spreads and fear hardens, public trust in authorities collapses.
- This leads to widespread civil disturbances, such as food riots, triggering either further paralysis in government or a repressive response.
- A far right politician uses extreme populist narratives to blame the crisis on a cabal of evil actors and/or immigrants, and uses this as an opportunity to seize power.

9. CYBERATTACK ON EDI INITIATIVES

CRISIS TYPE	ACTOR DYNAMIC	INFLUENCING TECH FACTORS
Misogynistic surge	Targeted coordinated	Cyberattack, Data corruption

Summary: A coordinated men's rights network launches cyberattacks against major UK employers' HR systems, corrupting recruitment data to fabricate evidence that women are being hired despite inferior qualifications. Fake statistics flood social media, triggering widespread discrimination and lawsuits.

- A well-organised men's rights network launches a new campaign to "expose the truth" behind EDI initiatives and what they claim is systematic discrimination against men in hiring.
- Hackers infiltrate HR systems at major UK employers, including government departments, and use real recruitment records to fabricate evidence that suggests women and minorities are consistently hired and promoted despite worse performance. The manipulated datasets are designed to appear authentic and released online.
- The dataset that is released is real, raising concern around data privacy in addition to stoking tensions around the place of EDI in the workplace.
- The manipulated stats are amplified on social media and news outlets initially struggle to verify the data's authenticity, with some reporting the story simply as a data leak.
- Workplace harassment escalates and several lawsuits are filed by men claiming they were passed over for less qualified candidates.

10. FOREIGN CONTROL OF CRITICAL INFORMATION INFRASTRUCTURE

CRISIS TYPE	ACTOR DYNAMIC	INFLUENCING TECH FACTORS
Economic, social, political, epistemic	Targeted coordinated	Digital infrastructure ownership concentration

Summary: The U.S. requisitions major tech firms and the global digital infrastructure they control, restricting access for non-U.S. entities. Partial restoration comes with strict oversight, turning essential services into powerful leverage in trade, diplomacy, and geopolitical conflicts.

- In the wake of rising geopolitical tensions, the U.S. government invokes emergency powers to requisition control over several major U.S.-based technology companies that collectively operate much of the world's critical information infrastructure including global cloud hosting providers, productivity suites (e.g. Google Suite), academic databases, undersea cables, and satellite internet services.
- Overnight, access to platforms like Google Workspace, AWS, and other core business tools is cut or severely restricted for non-U.S. entities.

- Multinational companies grind to a halt as essential services vanish. University research programs stall, and cross-border scientific collaborations collapse, unable to share data or coordinate projects. Even small businesses are crippled, their communications and records locked away on servers now under US federal control.
- Weeks later, partial access is restored—but under stringent licensing terms. Every data transfer, file share, and software deployment is subject to review.
- Access to digital infrastructure becomes a bargaining chip in trade negotiations, diplomatic disputes, and security pacts.
- Nations unwilling to align with U.S. policy find their businesses and research throttled with devastating economic impact.

APPENDIX C
INTERIM INSIGHTS FROM SCENARIO DEVELOPMENT AND MAPPING

For readers interested in repeating or iterating on the scenario development and mapping methods employed in workshop 1, we provide a few reflections and learnings from our experience that may be useful.

1. **With the multitude of potential influencing factors and actors involved in each hypothetical crisis there are near infinite permutations of each that we could attempt to map.**

On the one hand this highlights a possible shortcoming in our research methodology: the specific interventions that we will identify in subsequent stages will be targeted at specific crisis scenarios that are unlikely to ever arise in the exact form. This may raise questions about real world relevance. In Appendix A we provide analysis of how each scenario's constituent parts are rooted in real world events and contextual elements, which should put some concern to rest. A crisis is simply the result of the right (or wrong) factors coming together at the right (or wrong) time.

On the other hand the infinitude of potential scenarios also highlights the importance, in our methodology, of evaluating a diverse array of scenarios. By appraising a variety of crisis types and mechanisms we can be more confident that if, in the next workshop, any intervention recommendations are repeated across multiple scenarios, they likely indicate critical intervention points.

2. **Some methodological limitations stemming from the composition of our expert group started to emerge early.**

Because the workshops started with experts ideating and refining a wide range of hypothetical crisis scenarios featuring different kinds of crisis, coming about by a variety of mechanisms, and impacting numerous different communities and institutions, it was not possible to construct a group of subject area experts and stakeholders well-tailored to any specific scenario ahead of the game. Initial scenario refinement and mapping would likely have been deeper and more robust had we been able to. The limitation might have been remedied by inputting a predetermined set of base scenarios into the workshop and tailoring expert groups accordingly. However, one aim of our methodology was to see which scenarios the expert groups would develop and prioritise as most likely and severe based on their diverse insights on the various influencing factors (e.g. technological, geopolitical, legal, etc.). We did invite additional subject area experts to the second workshop to help mitigate the limitation going forward.

3. **Additional core themes (e.g. kinds of threat mechanisms and epistemic vulnerabilities) that were not articulated in the written scenario narratives will begin to emerge in systems maps at this stage.**

These common themes will also likely give rise to common areas of intervention. Along with the prioritised scenario topics, we would recommend reviewing the emergent themes ahead of the second workshop in case it would be helpful to bring in additional subject-area expertise for more robust intervention analysis.

Some of our emergent themes included uneven public digital and media literacy levels nationally, confluence strategic interests among domestic and foreign threat actors, and a high baseline for unrest and distrust of public institutions borne from economic hardship and poor service provision, erosion of news media ecosystems, reliance on critical information infrastructure controlled by foreign actors, and the information mediating influence of a variety of new and emerging AI tools and capabilities.

4. **Systems mapping is an iterative process and new threats and vulnerabilities will emerge continuously and seemingly endlessly.**

In the process of mapping out the scenarios and their interacting actors and actor systems, new influencing factors, threats, and vulnerabilities will continue to emerge triggered by previous feature adds. This is because, as noted in point 1, the number of potential scenario permutations are theoretically infinite, and also because human social-epistemic systems are infinitely complicated. In this respect, a time-limited workshop is a blessing. It restricts the depth and exhaustedness of analysis, which may not seem ideal, but does keep the analysis to a tractable size. We found that the two hours allocated for systems mapping gave adequate time to identify core crisis mechanisms and influences, with only more minor and peripheral features being added toward the end.

Licence to publish

Demos – Licence to Publish

The work (as defined below) is provided under the terms of this licence ('licence'). The work is protected by copyright and/or other applicable law. Any use of the work other than as authorized under this licence is prohibited. By exercising any rights to the work provided here, you accept and agree to be bound by the terms of this licence. Demos grants you the rights contained here in consideration of your acceptance of such terms and conditions.

1 Definitions

a 'Collective Work' means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this Licence.

b 'Derivative Work' means a work based upon the Work or upon the Work and other pre-existing works, such as a musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work or a translation from English into another language will not be considered a Derivative Work for the purpose of this Licence.

c 'Licensor' means the individual or entity that offers the Work under the terms of this Licence.

d 'Original Author' means the individual or entity who created the Work.

e 'Work' means the copyrightable work of authorship offered under the terms of this Licence.

f 'You' means an individual or entity exercising rights under this Licence who has not previously violated the terms of this Licence with respect to the Work, or who has received express permission from Demos to exercise rights under this Licence despite a previous violation.

2 Fair Use Rights

Nothing in this licence is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3 Licence Grant

Subject to the terms and conditions of this Licence, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) licence to exercise the rights in the Work as stated below:

a to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

b to distribute copies or phono-records of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works; The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4 Restrictions

The licence granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this Licence, and You must include a copy of, or the Uniform Resource Identifier for, this Licence with every copy or phono-record of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this Licence or the recipients' exercise of the rights granted hereunder. You may not sublicence the Work. You must keep intact all notices that refer to this Licence and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this Licence Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this Licence. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested.

b You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.

c If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Collective Works, you must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5 Representations, Warranties and Disclaimer

a By offering the Work for public release under this Licence, Licensor represents and warrants that, to the best of Licensor's knowledge after reasonable inquiry:

i Licensor has secured all rights in the Work necessary to grant the licence rights hereunder and to permit the lawful exercise of the rights granted hereunder without You having any obligation to pay any royalties, compulsory licence fees, residuals or any other payments;

ii The Work does not infringe the copyright, trademark, publicity rights, common law rights or any other right of any third party or constitute defamation, invasion of privacy or other tortious injury to any third party.

b Except as expressly stated in this licence or otherwise agreed in writing or required by applicable law, the work is licenced on an 'as is' basis, without warranties of any kind, either express or implied including, without limitation, any warranties regarding the contents or accuracy of the work.

6 Limitation on Liability

Except to the extent required by applicable law, and except for damages arising from liability to a third party resulting from breach of the warranties in section 5, in no event will licensor be liable to you on any legal theory for any special, incidental, consequential, punitive or exemplary damages arising out of this licence or the use of the work, even if licensor has been advised of the possibility of such damages.

7 Termination

a This Licence and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this Licence. Individuals or entities who have received Collective Works from You under this Licence, however, will not have their licences terminated provided such individuals or entities remain in full compliance with those licences. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this Licence.

b Subject to the above terms and conditions, the licence granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different licence terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this Licence (or any other licence that has been, or is required to be, granted under the terms of this Licence), and this Licence will continue in full force and effect unless terminated as stated above.

8 Miscellaneous

a Each time You distribute or publicly digitally perform the Work or a Collective Work, Demos offers to the recipient a licence to the Work on the same terms and conditions as the licence granted to You under this Licence.

b If any provision of this Licence is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this Licence, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

c No term or provision of this Licence shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

d This Licence constitutes the entire agreement between the parties with respect to the Work licenced here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This Licence may not be modified without the mutual written agreement of Demos and You.

DEMOS

Demos is a champion of people, ideas and democracy. We bring people together. We bridge divides. We listen and we understand. We are practical about the problems we face, but endlessly optimistic and ambitious about our capacity, together, to overcome them.

At a crossroads in Britain's history, we need ideas for renewal, reconnection and the restoration of hope. Challenges from populism to climate change remain unsolved, and a technological revolution dawns, but the centre of politics has been intellectually paralysed. Demos will change that. We can counter the impossible promises of the political extremes, and challenge despair – by bringing to life an aspirational narrative about the future of Britain that is rooted in the hopes and ambitions of people from across our country.

Demos is an independent, educational charity, registered in England and Wales. (Charity Registration no. 1042046)

Find out more at **www.demos.co.uk**

DEMOS

PUBLISHED BY DEMOS JANUARY 2026

15 WHITEHALL, LONDON, SW1A 2DD
T: 020 3878 3955
HELLO@DEMOS.CO.UK
WWW.DEMOS.CO.UK