The Alan Turing Institute

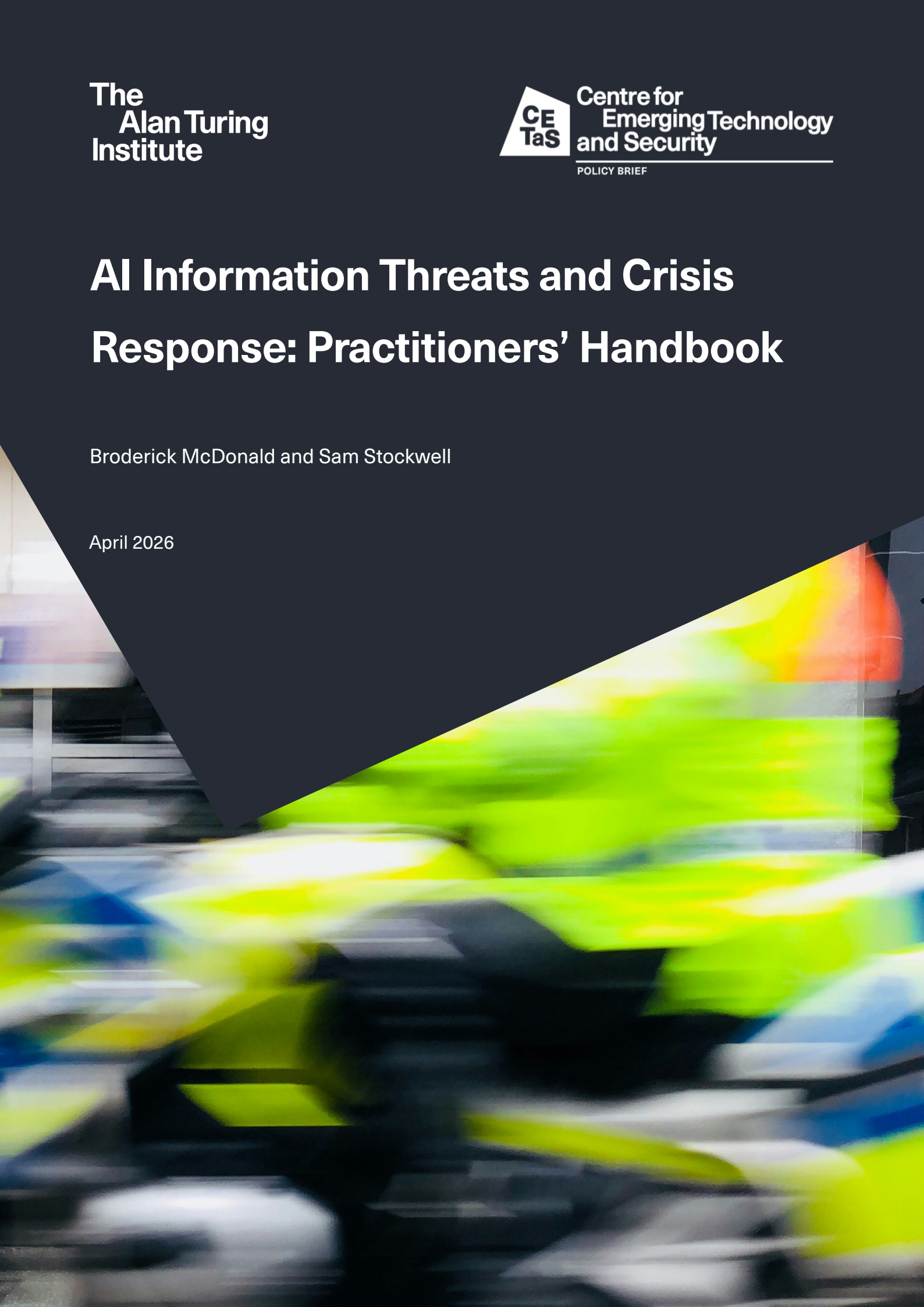
Centre for Emerging Technology and Security

POLICY BRIEF

AI Information Threats and Crisis Response: Practitioners' Handbook

Broderick McDonald and Sam Stockwell

April 2026



Front and back cover image: John Cameron on Unsplash

Introduction

AI information threats are exacerbating risks to democracy[1] and real-world violence[2] during crisis events – such as terrorist attacks, and national or international security incidents. From deepfake videos[3] purporting to show perpetrators from recent crises, to in-app chatbots generating false information during ongoing incidents,[4] these threats have become more commonplace and sophisticated over the past year. Despite significant research on disinformation during previous crisis events, little research has explored the role that AI tools may play in worsening these dynamics and contributing to real-world harm.

Within this context, recent CETaS research[5] developed an evidence base on AI information threats during crisis events, and provided a series of recommendations to strengthen societal resilience against them. In this handbook, we shift the focus towards operational-level actions that could be adopted by the UK Government, industry or civil society to prepare for and address AI-driven crises.

The below measures, organised chronologically, seek to improve the responses and coordination of each sector prior to, during and after crisis events.

[1] David Gilbert, "AI-Powered Disinformation Swarms Are Coming for Democracy," *Wired*, 22 January 2026, https://www.wired.com/story/ai-powered-disinformation-swarms-are-coming-for-democracy.

[2] Ben Quinn and Dan Milmo, "How TikTok bots and AI have powered a resurgence in UK far-right violence," *The Guardian*, 2 August 2024, https://www.theguardian.com/politics/article/2024/aug/02/how-tiktok-bots-and-ai-have-powered-a-resurgence-in-uk-far-right-violence.

[3] Andrew R. Chow and Billy Perrigo, "Google's New AI Tool Generates Convincing Deepfakes of Riots, Conflict, and Election Fraud," *TIME*, 3 June 2025, https://time.com/7290050/veo-3-google-misinformation-deepfake.

[4] "False AI 'fact-checks' stir online chaos after Kirk assassination," *France 24*, 12 September 2025, https://www.france24.com/en/live-news/20250911-false-ai-fact-checks-stir-online-chaos-after-kirk-assassination.

[5] Sam Stockwell, Ardi Janjeva and Broderick McDonald, "Adding Fuel to the Fire: AI Information Threats and Crisis Events," *CETaS Research Reports* (February 2026).

1. Prior to Crisis Events

Government

AI scenario planning: In order to better forecast what types of AI-driven crisis events could emerge, as well as corresponding vulnerabilities in the UK Government's existing crisis response strategy, the Cabinet Office should run cross-Whitehall tabletop exercises and organisational red-teaming scenarios[6] that simulate these types of incidents. A previous tabletop exercise conducted by CETaS with stakeholders from across government and industry found that such sessions were crucial in determining roles and responsibilities between different departments[7] prior to any AI-driven incidents.

Multilayered crisis communications: The Government Communication Service should update the Emergency Planning Framework[8] (PRIMER), used to train crisis communication professionals, to include best practice from recent crisis events. This should include amplifying consistent factual information during an ongoing crisis via non-governmental channels, such as local news outlets, religious centres and sports clubs. This will enable any strategic narratives to reach a wider audience and reinforce the facts on the ground if harmful speculation is taking place.[9]

AI threat intelligence channels: While progress has been made over the past decade in improving coordination within the UK Government when it comes to liaising with social media platforms during crisis scenarios, there remains uncertainty[10] over the extent to which these relations extend to different types of AI companies. Given its role in engaging with several frontier AI labs,[11] the AI Security Institute (AISI)[12] should establish trusted information-sharing channels with relevant individuals in these organisations. This would enable the sharing of threat intelligence during a live crisis where AI information threats are present.

[6] Tommy Shaffer Shane, *Preparing for AI security incidents* (The Centre for Long-Term Resilience: September 2025), https://www.longtermresilience.org/reports/preparing-for-ai-security-incidents.
[7] Stockwell, Janjeva and McDonald (2026).
[8] "Emergency Planning Framework," *Government Communication Service*, 2018, https://www.communications.gov.uk/publications/emergency-planning-framework.
[9] Sam Stockwell and Al Baker, "The Cost of Silence: Crisis Communication and Real-world Harm Following Security Incidents," *CETaS Expert Analysis* (July 2025).
[10] Shane (2025); Stockwell, Janjeva and McDonald (2026).
[11] "How we're working with frontier AI developers to improve model security," *AI Security Institute*, 13 September 2025, https://www.aisi.gov.uk/blog/how-were-working-with-frontier-ai-developers-to-improve-model-security.
[12] "The AI Security Institute (AISI)," *AI Security Institute*, https://www.aisi.gov.uk.

Government liaison channels: The Cabinet Office's National Security Secretariat (NSS)[13] should ensure that communication channels with the Department for Science, Innovation and Technology (DSIT) serve as the central liaison point for cross-government engagement[14] with AI companies during ongoing crisis scenarios. For example, this central liaison hub could be housed in the National Security Online Information Team (NSOIT)[15] or AISI's Societal Resilience[16] team. By doing so, this will help to streamline queries and reduce the risk of slowing progress on implementing solutions.

Industry

Crisis command centres: AI companies and social media platforms providing services in the UK should formalise the creation of dedicated command centres – or governmental liaison officers in the case of smaller firms – within their internal crisis response protocols. These would act as centralised hubs or designated officers[17] responsible for information sharing and coordination with relevant government departments (e.g. DSIT) during live crisis events, inspired by similar mechanisms set up by some tech platforms during the Southport riots[18] and 2024 global elections.[19]

Social media community guidelines: Social media platforms that fall under Category 1 or 2B in the UK's Online Safety Act[20] should update their terms of service (ToS) to include new policies on de-amplifying or downranking[21] deceptive content during crisis events that may lead to violence. This could take a similar approach to the recent changes that some social media companies introduced[22] to their own community guidelines, which prevent unverified information about crises from being promoted to users' feeds, and integrate warning prompts to users before they try to share such content with others. More broadly, we advocate for nuanced approaches and technical solutions[23] that protect free expression, but

[13] "About us," *National Security and Intelligence*, https://www.gov.uk/government/organisations/national-security/about.
[14] Shane (2025).
[15] "National Security Online Information Team: privacy notice," *Department for Science, Innovation and Technology*, 16 April 2024, https://www.gov.uk/government/publications/national-security-online-information-team-privacy-notice/national-security-online-information-team-privacy-notice.
[16] "Societal Resilience," *AI Security Institute*, https://www.aisi.gov.uk/category/societal-resilience.
[17] Stockwell, Janjeva and McDonald (2026).
[18] Science, Innovation and Technology Committee, *Social media, misinformation and harmful algorithms* (House of Commons: July 2025), https://committees.parliament.uk/publications/48745/documents/258221/default.
[19] "How OpenAI is approaching 2024 worldwide elections," *OpenAI*, 15 January 2024, https://openai.com/index/how-openai-is-approaching-2024-worldwide-elections.
[20] Lorna Woods and Maeve Walsh, "Categorisation of services in the Online Safety Act," *OSA Network*, 1 May 2024, https://www.onlinesafetyact.net/analysis/categorisation-of-services-in-the-online-safety-act.
[21] Science, Innovation and Technology Committee (2025).
[22] Anna Iovine, "TikTok rule changes are coming for creators and commenters," *Mashable*, 16 August 2025, https://mashable.com/article/tiktok-community-guidelines-update-september-2025.
[23] Stockwell, Janjeva and McDonald (2026).

do not inadvertently amplify deceptive content (e.g. freedom of speech, not freedom of reach).[24]

Chatbot fact-checking caveats: AI companies with chatbot services should alter their conversational user interfaces (CUIs) to include more prominent and specific caveat notices to users[25] on the chatbots' fact-checking limitations during live crisis events. These notices could take the form of a pop-up alert message when a chatbot is queried on a live crisis, and tied to keywords associated with it, flagging to users that they should not interpret outputs as factual while an incident is unfolding in real time.

AI incident threat reporting: The Frontier Model Forum (FMF),[26] which focuses on the safe and responsible development of frontier AI systems, should expand its existing information-sharing agreement[27] between members to include a new threat-reporting mechanism for crisis scenarios. Akin to the Global Internet Forum to Counter Terrorism's (GIFCT's) 'Incident Response Framework',[28] this mechanism would be activated during an ongoing security incident to facilitate cross-industry measures (such as hash-sharing or behavioural signals reporting) designed to tackle AI misuse. Summaries of effective interventions, rather than user data, could also be shared with industry partners outside the FMF to ensure that others can support with implementing solutions.

Civil society

Media literacy and education: Civil society organisations (CSOs) can play a critical role in building societal resilience[29] to AI information threats prior to crisis events via media and digital literacy programmes. New initiatives could be created that focus on outlining common tactics of information manipulation by threat actors, as well as digital hygiene habits to reduce the risk that users consume falsehoods during crises, when there may be heightened susceptibility to such narratives.[30]

---

[24] Renée DiResta, "Free Speech Is Not the Same As Free Reach," *Wired*, 30 August 2018, https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach.
[25] Stockwell, Janjeva and McDonald (2026).
[26] "Frontier Model Forum: Advancing frontier AI safety and security," *Frontier Model Forum*, https://www.frontiermodelforum.org.
[27] "FMF Information-Sharing," *Frontier Model Forum,* https://www.frontiermodelforum.org/information-sharing.
[28] "The Incident Response Framework," *Global Internet Forum to Counter Terrorism*, https://gifct.org/incident-response.
[29] Pasi Mäenpää, Maija Faehnle and Henrietta Grönlund, "Collective Informational Capacity of the Civil Society Shaping Societal Resilience," in *Information Resilience and Comprehensive Security*, eds. Petri Uusikylä, Harri Jalonen and Annukka Jokipii (Palgrave Macmillan, 2024), 307-332, https://link.springer.com/chapter/10.1007/978-3-031-66196-9_14.
[30] Martin Innes, *'Fogging' and 'Flooding': Countering Extremist Mis/Disinformation After Terror Attacks* (GNET: November 2021), https://gnet-research.org/wp-content/uploads/2021/11/GNET-Report-Fogging-And-Flooding-Countering-Extremist-MisDisinformation-After-Terror-Attacks.pdf.

2. During Crisis Events

Government

Threat identification: During live crisis events in the UK, NSOIT should monitor and identify highly visible AI information threats – including across the entire content life cycle (e.g. generation, dissemination and engagement points).[31] Practically speaking, the focus of this work should be to identify and track the spread of content designed to incite real-world violence (e.g. deepfakes purporting to show evidence from the scenes of incidents[32] or calling for violent acts[33]). Critically, NSOIT's mandate during such incidents must incorporate a high threshold for inclusion and focus only on viral false information rather than political content, safeguarding human rights such as free expression and the right to peaceful protest.

Response coordination: When a crisis event being driven by AI information threats is unfolding, AISI should provide specialist advice and expertise[34] to the Cabinet Office, including proposing potential AI safety measures. NSOIT should also liaise with industry to ensure that high-risk or illegal content that has been identified as inciting violence is quickly removed from online platforms. Finally, the Cabinet Office's National Situation Centre (SitCen)[35] should corroborate any data on AI information threats from NSOIT with data flows from other departments to determine the severity of the wider threat landscape.

Industry

Cross-industry information sharing: Once the FMF has integrated changes to its information-sharing agreement based on the recommendation in Section 1, it should ensure that threat intelligence and communication channels are enabled for members to securely share data on AI model vulnerabilities with each other, as well as potential mitigation strategies. Where appropriate, harmful AI-generated content (such as deepfakes) identified

[31] Stockwell, Janjeva and McDonald (2026).
[32] Io Dodds, "Fake videos and AI chatbots drive disinformation about LA protests," *The Independent*, 13 June 2025, https://www.independent.co.uk/news/world/americas/us-politics/los-angeles-protests-fake-ai-videos-b2768507.html.
[33] "Posts Supporting UK Riots," *Oversight Board (Meta)*, 23 April 2025, https://www.oversightboard.com/decision/bun-6aqh31t6.
[34] Stockwell, Janjeva and McDonald (2026).
[35] Gordon Corera, "Inside the government's secret data room," *BBC News*, 15 December 2021, https://www.bbc.co.uk/news/technology-59651706.

during a crisis event should also be hashed and stored in a secure content database[36] to assist with post-incident evaluation assessments.

Social media interventions: Once relevant social media companies have integrated changes to their ToS based on the recommendation in Section 1, they should ensure that such measures are swiftly deployed during a crisis event in which AI information threats have been identified as likely to contribute to real-world violence. This may include removing, downranking or demonetising content, as well as adding warning labels,[37] depending on the context and severity of the content in question.

Civil society

Automated rapid response tools: CSOs can contribute to a 'whole-of-society' crisis response[38] by researching and investing in automated fact-checking tools, as reflected in WITNESS's 'Deepfake Rapid Response Force',[39] Tech Against Terrorism's 'Terrorist Content Analytics Platform'[40] and the 'Social Media Analytics and Reporting Tool'[41] (SMART 2.0). Such initiatives can stem the spread of AI information threats before they gain critical mass among the wider public, by automatically detecting harmful keywords or improving situational awareness. However, automated methods are prone to making mistakes, so they must be accompanied by robust human-centric training and oversight throughout the process. The significant cost of such efforts is often challenging for small CSOs and could be supported by government or private philanthropic grants, as well as resource-sharing partnerships between CSOs.

Platforming credible voices: CSOs are also uniquely positioned to enhance the effectiveness of countering AI information threats by platforming credible voices from affected communities.[42] When credible representatives and non-partisan organisations amplify truth over falsehoods, they help to reach not only the public-at-large, but also

[36] Christian Schwieter, *Online crisis protocols – Expanding the regulatory toolbox to safeguard democracy during crises* (Institute for Strategic Dialogue: December 2022), https://www.isdglobal.org/publication/online-crisis-protocols-expanding-the-regulatory-toolbox-to-safeguard-democracy-during-crises.

[37] Science, Innovation and Technology Committee (2025).

[38] Aino Ruggiero, Wojciech D. Piotrowicz and Lijo John, "Enhancing societal resilience through the whole-of-society approach to crisis preparedness: Complex adaptive systems perspective – The case of Finland," *International Journal of Disaster Risk Reduction* 114, November 2024, https://doi.org/10.1016/j.ijdrr.2024.104944.

[39] "Deepfake Rapid Response Force," *WITNESS*, https://www.gen-ai.witness.org/deepfakes-rapid-response-force.

[40] "Terrorist Content Analytics Platform," *Tech Against Terrorism*, https://terrorismanalytics.org.

[41] Mahmoud Mousa Hamad et al., "SMART 2.0: Social Media Analytics and Reporting Tool Applied to Misinformation Tracking," *Media and Communication* 13, April 2025, https://www.cogitatiopress.com/mediaandcommunication/article/view/9543/4300.

[42] Gabriel Marmentini and Jeanine Abrams McLean, "Grassroots Strategies to Combat Election-Related Misinformation," *Stanford Social Innovation Review*, 16 September 2024, https://ssir.org/articles/entry/grassroots-strategies-election-misinformation.

minority demographic communities that may be explicitly targeted by those seeking to cause harm. While governments could strive to do the same, community-focused CSOs are considerably more likely to be well received by sceptical audiences,[43] amplifying reach and authenticity. Critically, CSOs should maintain their independence during this process, operating outside of governmental channels in order to avoid the risk of appearing to uncritically endorse or repeat government narratives.

Debunking chatbots: CSOs that operate in the fact-checking space should explore the use of experimental chatbots such as DebunkBot[44] to help counter conspiracy theories during live crisis events. Such tools could be used to inform responses to viral content that is harmful or misleading, alongside human-curated counter-narratives and strict ethical oversight. A 2024 study found[45] that conversations between 2,000 participants and DebunkBot led to a 20% reduction in peoples' confidence in conspiracy theories – with some reductions lasting up to 2 months.

[43] Jon Bateman and Dean Jackson, *Countering Disinformation Effectively: An Evidence-Based Policy Guide* (Carnegie Endowment for International Peace: January 2024), https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide.

[44] "DebunkBot," https://debunkbot.com/survey/conspiracies/home.

[45] Thomas H. Costello, Gordon Pennycook and David G. Rand, "Durably reducing conspiracy beliefs through dialogues with AI," *Science* 385, no. 6714 (September 2024), https://www.science.org/doi/10.1126/science.adq1814.

3. After Crisis Events

Government

Post-incident review process: The Cabinet Office should create a post-incident review process by which different government and law enforcement stakeholders can provide feedback on the strengths and limitations of crisis response approaches in the wake of major AI-driven crisis events. This mechanism could be inspired by similar functions within the UK Health Security Agency's public health incident response plan.[46]

Post-incident monitoring indicators: NSOIT should ensure that any monitoring indicators for defining different types of AI information threats are kept under regular review[47] to ensure that they remain relevant in the face of rapid AI innovation, ever-changing adversarial tradecraft, and the possible emergence of new threat vectors that risk public safety.

Industry

AI incident content database: The FMF should develop and maintain a hashed repository of prominent AI information threats that could then inform AI model safety patches, reducing the risk that similar content will be produced in the future. This initiative would further help FMF members to track trends and techniques used by threat actors, while safely preventing the information from being lost during content removals. The repository could be inspired by GIFCT's own hashing strategy,[48] in which terrorist or violent extremist content that has been removed can still be accessed by GIFCT members to ensure that future copies can be detected and removed on their platforms.

Civil society

Research and data access: After a major security incident has been resolved, CSOs can play a critical role in examining whether AI tools were exploited to cause harm, and how such threats might be addressed. In particular, CSOs can ensure that an appropriate

[46] "Incident response plan," *UK Health Security Agency*, 15 January 2025, https://www.gov.uk/government/publications/emergency-preparedness-resilience-and-response-concept-of-operations/incident-response-plan.
[47] *Online Safety: Additional Safety Measures* (Ofcom: June 2025), https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/consultation-online-safety---additional-safety-measures/main-documents/consultation-additional-safety-measures-30-july-2025.pdf?v=403587.
[48] Schwieter (2022).

balance of human rights, including freedom of expression and the right to safety, are upheld by democratic governments when making interventions. To do this work effectively, however, these organisations require sufficient access[49] to social media data. With the EU's Digital Services Act now legally requiring social media platforms to provide data access tools[50] for researchers exploring systemic risks (including public security), the UK Government should coordinate with relevant EU partners to understand how similar access can be provided to UK CSOs during crisis scenarios.

Chatbot red-lists: Following any crisis event in which hostile foreign state-sponsored actors have sought to manipulate the outputs of AI chatbots by flooding them with fake news articles,[51] CSOs and academics should share links to such articles with the news-auditing organisation NewsGuard.[52] In turn, NewsGuard should use this data to update its FAILSafe platform,[53] which allows AI developers to remove sources from their chatbot datasets that are exposed as being part of foreign information manipulation and interference (FIMI)[54] activities.

[49] Sam Stockwell et al., "AI-Enabled Influence Operations: Safeguarding Future Elections," *CETaS Research Reports* (November 2024).

[50] Joint Research Centre, "FAQs: DSA data access for researchers," *European Centre for Algorithmic Transparency*, 3 July 2025, https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2025-07-03_en.

[51] Stockwell, Janjeva and McDonald (2026).

[52] "NewsGuard," https://www.newsguardtech.com.

[53] "FAILSafe for AI," *NewsGuard*, https://www.newsguardtech.com/solutions/failsafe-for-ai.

[54] "Information Integrity and Countering Foreign Information Manipulation & Interference (FIMI)," *European External Action Service*, 17 March 2026, https://www.eeas.europa.eu/eeas/information-integrity-and-countering-foreign-information-manipulation-interference-fimi_en.

4. Ongoing Policy Priorities

In addition to the operational priorities outlined above, the UK Government should develop and continue work on five broader, ongoing policy areas. While these will require additional resources to maintain over the longer horizon, they can substantially enhance our wider societal resilience to AI information threats.

1) **AI preparedness framework:** Given concerns over the lack of an adequate risk management framework to tackle AI threats, the UK Government should develop and periodically review a new set of guidelines that enhance preparedness by ensuring that incidents are "effectively **anticipated**, **prevented**, **prepared** for, and can be **responded** to".[55] Drawing inspiration from similar frameworks for biological security and emergency planning, this four-pillar strategy would help to formalise and direct cross-government efforts towards strengthening crisis resilience.

2) **Helping small AI startups to tackle misuse**: The rapid proliferation of small AI startups offering dual-use agentic and specialised tools represents a significant opportunity for threat actors to acquire powerful new capabilities. These startups may have hundreds of thousands to millions of monthly users, but their small teams often lack dedicated Trust & Safety (T&S) staff to liaise with government agencies, as well as the resources or in-house expertise[56] to respond rapidly to malicious misuse of their products. To mitigate these risks, AISI should develop and maintain an up-to-date list of AI startups with dual-use products that are at risk of being exploited by threat actors. Small AI startups that are most vulnerable but lack capacity could be provided with information toolkits, zero-cost access to safety tools such as ROOST,[57] and mentorship from organisations such as the FMF, GIFCT, Tech Against Terrorism,[58] and The Christchurch Call.[59]

3) **Chatbot guidance for users:** As part of the Digital Inclusion Action Plan,[60] DSIT should produce and continually distribute guidance for users[61] on how to navigate the use of AI chatbots for fact-checking purposes – particularly during crisis

[55] Shane (2025).
[56] Schwieter (2022).
[57] "Robust Open Online Safety Tools," *ROOST*, https://roost.tools.
[58] "Disrupting Terrorism since 2015," *Tech Against Terrorism*, https://techagainstterrorism.org/home.
[59] "The Christchurch Call," https://www.christchurchcall.org.
[60] *Digital Inclusion Action Plan: First Steps* (Department for Science, Innovation and Technology: February 2025), https://www.gov.uk/government/publications/digital-inclusion-action-plan-first-steps/digital-inclusion-action-plan-first-steps.
[61] Stockwell, Janjeva and McDonald (2026).

scenarios. Several recent crisis events[62] have revealed AI chatbots spreading false or inflammatory information to users during incidents, which could result in real-world harm.

4) **Monitoring indicators:** NSOIT should define monitoring indicators[63] used to determine when different severity thresholds for AI information threats have been met. NSOIT should also ensure that any data insights from these indicators are routinely shared with SitCen in the Cabinet Office to inform the UK Government's Commonly Recognised Information Picture[64] – a single, authoritative, standard-format overview of crises used to brief and inform decisions across government.

5) **Academic engagement:** CSOs and relevant government departments should expand and maintain ongoing engagement with the academic community to identify emerging AI information trends and promising new technical interventions, such as the redirect method,[65] algorithmic de-amplification,[66] or pre-bunking.[67] Additionally, establishing widely known and accessible channels for researchers to share their findings with government and industry organisations (e.g. the Home Office, DSIT, GIFCT and the FMF) would help ensure that the UK fully leverages insights from its publicly funded academic research.

---

[62] "False AI 'fact-checks' stir online chaos after Kirk assassination," *France 24*, 12 September 2025, https://www.france24.com/en/live-news/20250911-false-ai-fact-checks-stir-online-chaos-after-kirk-assassination; Esteban Ponce de León and Ali Chenrose, "Grok struggles with fact-checking amid Israel-Iran war," *DFRLab*, 24 June 2025, https://dfrlab.org/2025/06/24/grok-struggles-with-fact-checking-amid-israel-iran-war; David Gilbert, "AI Chatbots Are Making LA Protest Disinformation Worse," *Wired*, 10 June 2025, https://www.wired.com/story/grok-chatgpt-ai-los-angeles-protest-disinformation; Thomas Renault, Mohsen Mosleh and David Rand, "@Grok Is This True? LLM-Powered Fact-Checking on Social Media," *preprint*, January 2026, https://osf.io/preprints/psyarxiv/85quw_v2.

[63] Stockwell, Janjeva and McDonald (2026).

[64] *The Amber Book: Managing crisis in central government* (Cabinet Office: April 2025), https://www.gov.uk/government/publications/the-central-government-s-concept-of-operations/the-amber-book-managing-crisis-in-central-government-html.

[65] ""Redirect Method" Yields Valuable Insights for Countering Online Extremism," *Anti-Defamation League*, 16 January 2020, https://www.adl.org/resources/article/redirect-method-yields-valuable-insights-countering-online-extremism.

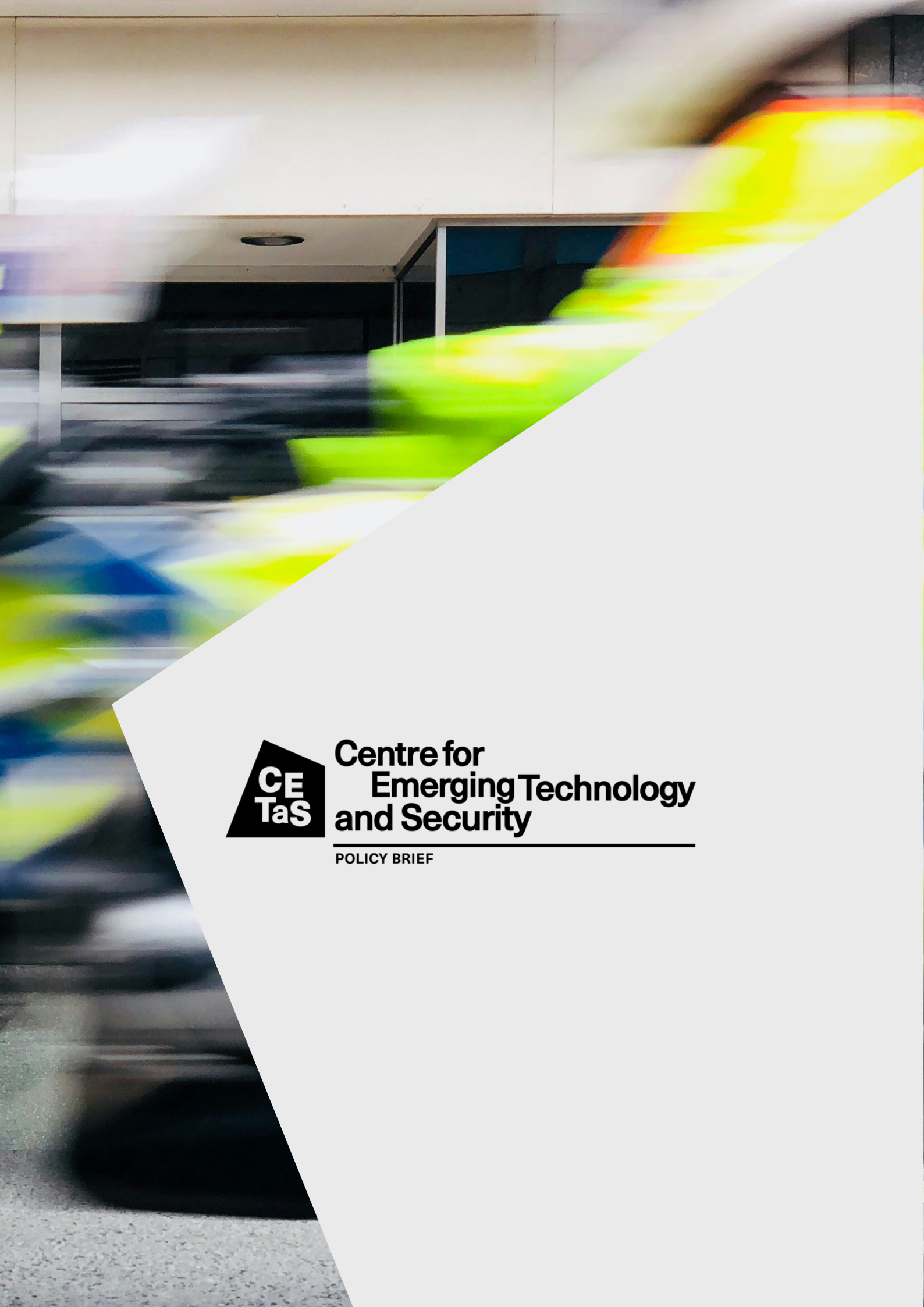
[66] Bàrbara Molas and Heron Lopes, *"Say it's only fictional": How the Far-Right is Jailbreaking AI and What Can Be Done About It* (International Centre for Counter-Terrorism: October 2024), https://icct.nl/sites/default/files/2024-10/Molas%20and%20Lopes.pdf.

[67] Sander van der Linden, "Countering misinformation through psychological inoculation," *Advances in Experimental Social Psychology* 69, March 2024, 1-58, https://www.sdmlab.psychol.cam.ac.uk/files/media/countering.pdf.

About the Authors

Broderick McDonald is an academic researcher at the University of Oxford, Kings College London, and the Alan Turing Institute. Over the past decade, he has worked across government, academia, tech and civil society to research and counter global security threats. Broderick's research and commentary has been featured in The New York Times, The Washington Post, The Wall Street Journal, Foreign Affairs, Financial Times, BBC, The Guardian, and The Globe and Mail.

Sam Stockwell is a Senior Research Associate at CETaS. His research interests focus on how AI is affecting the online information ecosystem, particularly in relation to misinformation and disinformation, the moderation of harmful content, and election interference.



CETaS

Centre for Emerging Technology and Security

POLICY BRIEF