Roundtable discussion: How can we manage the risk from AI to spread Information Disorder?

4th & 5th February 2025

In-person workshop. Cranfield University.

Dr Jack Steadman, Cranfield University

Steadman, J. & Riley-Smith, T. (2025). *Summary report – Roundtable Discussion: How can we manage the risk from AI to spread Information Disorder?*

Table of Contents

List of Abbreviations

AI – Artificial Intelligence

ID – Information Disorder

DSIT – Department for Science, Innovation and Technology

DSTL – Defence Science and Technology Laboratory

NSEC – Network for Security, Excellence and Collaboration

IEEE – Institute of Electrical and Electronics Engineers

ETSI – European Telecommunications Standards Institute

ITU – International Telecommunication Union

Preface

Deceivers. Dissemblers. Tricksters.

These protagonists are found in ancient mythologies from around the world, suggesting that "fake news" is as old as humanity itself.

In February 2025, we explored the challenge of "Information Disorder" (ID) from a new angle, addressing the capacity of an emerging technology to turbocharge this threat.

Three linguistic milestones from the recent past highlight the topicality of this report:

- "**infodemic"** appeared in 2013 after an explosion of news relating to the SARS epidemic: it is defined as a proliferation of diverse - often unsubstantiated - information linked to a crisis, controversy, or event, that travels rapidly and uncontrollably through news, online, and social media, and intensifies public speculation or anxiety;
- in 2016, "**post-truth"** became the Oxford English Dictionary's international Word of the Year, describing occasions when objective facts are less influential in shaping public opinion than emotional appeals;
- **"AI"** was Word of the Year for Collins Dictionary in 2023, following 12 months in which use of the term had quadrupled: this reflected an increase, it was said, in conversations about whether AI will be a force for revolutionary good or apocalyptic destruction.

What happens when the power of AI is harnessed by ill-intentioned actors to generate an infodemic in a post-truth world? What, specifically, are the risks to the UK and how can these be managed? These were the questions we addressed in February 2025, when NSEC[1] brought problem-owners and experts together to illuminate the challenge.

The difficulties associated with this challenge (countering the risk from bad actors using AI to spread information disorder) are clearly presented in this report. For instance:

[1] https://nsec.uk/

- AI will bring benefit to many walks-of-life and has a valuable contribution to make to "content generation"; but it can also equip those who are ill-intentioned with new powers to wreak havoc;
- assessing "disordered" information is not easy: we celebrate and promote free speech, and there are countless legitimate reasons for people to spin the facts - a political party advancing its manifesto; a technology start-up pitching for investment; a charity promoting its cause; a stand-up comedian satirizing those in positions of power);
- it is equally challenging to assess the relevant vulnerabilities and harms in/to our open and diverse democracy, not least because damage can be caused accidentally and unintentionally.

But valuable insights are to be found here about the threat, potential targets of attack and both systemic and cultural vulnerabilities. I was struck by a collective concern – at our roundtable – about the lack of critical judgment (across our population) when it comes to the consumption of news; and if one of the more exciting recommendations in this report is for an **Office for Media Literacy** to strengthen our national resilience here.

NSEC is building a reputation for wrestling with "wicked" security problems like this. We create a safe space where stakeholders and researchers can work together to anatomize a major security challenge and propose remedies. The approach taken here - when dealing, as in this case, with a transformational technology – is well-aligned with UKRI's promotion of responsible research and innovation, exemplified by the AREA framework (**A**nticipate, **R**eflect, **E**ngage, **A**ct)[2].

We are grateful to Simon Harwood at Leonardo for sponsoring this roundtable; and to Cranfield University for hosting the event (led by Nick Lindley and Caroline Dawson). Laura Samuels provided an invaluable service in connecting Government and Academia together as the Home Office's appointment as NSEC administrator; and Dr Jack Steadman is to be congratulated for translating notes from our roundtable into this report.

[2] Framework for responsible research and innovation – UKRI

Ultimately, the quality of insights contained here has only been achieved thanks to the enthusiastic contribution of all participants– as can be seen from the list of attendees in Appendix C, they add authority and authenticity to the debate through the range of organizations and depth of experience reflected at the roundtable.

Dr. Tristram Riley-Smith

Network for Security Excellence and Collaboration

17 April 2025.

Building the bulletproof library: Trust, information and security in the post-digital age

Introduction

The Defence and Security team at Cranfield University has partnered with the Network for Security Excellence and Collaboration (NSEC) to host a roundtable discussion on Artificial Intelligence (AI) and information disorder (ID). The work was sponsored by Leonardo UK. The aim of the roundtable was to encourage the exchange of experience, expertise, and insights among participants, drawn from a diverse pool of problem owners, industry representatives, and researchers. A detailed Agenda of the event is provided in Appendix A.

Artificial Intelligence (AI) is judged to represent a pivotal technological advance in the frontiers of human capability. It promises to transform efficiency and effectiveness in many spheres of human endeavour, with multi-modal AI, capable of processing and creating voice, video, and images, as well as text.

This roundtable was held in response to concerns expressed by policymakers and other stakeholders about how this technology could be used by hostile actors to promulgate disordered information. This report, summarising the outcome of discussion over two days, aims to promote awareness of the issues, to focus minds, and to help formulate an effective policy response. In setting out these findings and recommendations, all participants were keen to avoid demonising AI, which promises to deliver so much public good. But the downsides should not be ignored, and need addressing.

'Information Disorder' (ID) is used-to describe various types of information manipulation or misrepresentation, categorised as 'misinformation, malinformation, or disinformation'[3]. Fundamentally, this debate concerns the nature of truth and belief in the post-digital age. With the increasing capability of generative AI and its related technologies, the truthfulness

[3] Princeton Public Library. *Misinformation, Disinformation & Malinformation: A Guide*. viz., '…The deployment of such information to alter social behaviours, drive crime and generate instability'.
Available from: https://princetonlibrary.org/guides/misinformation-disinformation-malinformation-a-guide/.

of events is increasingly open to misrepresentation and deception. This roundtable aimed to understand the risks involved in this emerging space between AI and ID, describe potential vulnerabilities, develop mitigation strategies, and help formulate robust frameworks in the prevention and interception of ID generated by AI.

Trust plays a critical role in shaping our judgment about the veracity of information, especially when consuming "news". Leaps in 'generative' AI inevitably lead to the 'generation' of content. As knowledge grows concerning the capacity of AI to generate realistic synthetic content, indistinguishable from human-generated content, there will be increasing uncertainty about provenance and authenticity, particularly in relation to emotive issues and partisan reporting.

Communities and wider society may become more cynical or sceptical of conventional information sources. Any effort to counter these seeds of disbelief requires directed, potent and authoritative / compelling communication.

The roundtable discussion included a range of experts of policymakers, industry and academia to discuss the threat and risk landscape of AI-ID. Two breakout sessions were part of the roundtable, where delegates were divided into four groups, each discussing a list of questions regarding Understanding and Managing the Risk (see Appendix B).

Key points of discussion to emerge from our deliberations are divided into *Findings*, *Vulnerabilities*, and *Recommendations*. To assist policymakers and readers, a detailed Annex provides further detail regarding *Issues and Definitions*, to help encourage balanced appraisal and consideration.

Findings

The Threat

- There has always been a challenge in knowing what information we can trust, and there have always been individuals and organisations that will distort or invent information to meet their own ends and disadvantage others.

- Changes to the information infrastructure serving us in the last quarter of the 20th Century have made things more difficult:

    - the global nature of the World Wide Web and the fragmented nature of Social Media have shifted power dynamics around "control of the narrative" (from homogeneity to heterogeneity): it is becoming harder to assess the reliability of the information we consume.

- This problem is further exacerbated by the growth of AI, where the widespread availability of machine learning tools – including generative models and deepfake technologies – means anyone anywhere can create or modify data (text, images, voice or video) with minimal effort, low cost, and increased realism.

    - AI capabilities that have been the preserve of sovereign states are becoming democratised (for example, the economic impact and reportedly low cost of Deepseek)[4]; and Organised Crime and other groups can offer "Influence as a Service".

    - AI technology itself is developing at speed: for instance, **agentic AI** introduces the ability for malign actors to improve the efficacy of disinformation campaigns, through iteratively finessing the message to suit target audiences (as legitimate marketing operations might do);

[4] Stanford University Human-Centred Artificial Intelligence (2025). *How Disruptive Is DeepSeek?* Author. Available from: https://hai.stanford.edu/news/how-disruptive-deepseek-stanford-hai-faculty-discuss-chinas-new-model.

- We are aware of malign actors (such as Hostile States, Organised Crime Groups and commercial or ideological adversaries) who attack UK interests through dis- or mal-information operations; but it can be difficult to prove intent;
    - Fake news may be generated to meet the strategic aims of a hostile state, for amusement of a teenager in their bedroom, or even by mistake;
    - The attention economy underpinning commercial models pursued by Social Media and other platforms can prioritise engagement over the accuracy of content; this can leverage factors that stimulate controversary and seek to promote 'infotainment' from conflict.
- But the harms created by Information Disorder can be the same, regardless of intent.
    - Harms can include societal unrest at home, especially at a time of heightened tensions or economic pressures, driving wedges between different communities (e.g. ethnic groups) or political viewpoints (e.g. Scottish Independence or Just Stop Oil), or triggering panic responses (such as a run on the banks);
    - UK interests can also be damaged abroad: we have seen Information Disorder operations conducted against companies overseas (e.g. in Mali and Serbia), causing substantial commercial impact.
    - There can be profound and long-lasting secondary effects: when fake news is suspected or exposed, there can be a damaging decline in public trust in all sources of information.

Targets

The spheres below represent potential vectors that could either be targeted by malign actors, or could suffer disproportionately from information disorder.

- Events: e.g., Elections, Pandemics, disasters
    - Trust, in reporting; response, in policy and action.
    - Responsibilities, of the UK; Consequences, for parties involved.

- Relationships
  - Domestic: politicians, police, media, publics.
  - Overseas: diplomacy; trade, broader policy.
- Industrial Sector
  - 'Infiltrative penetration', e.g. hardware, networks, edge-type communications.
  - Supply chain factors – dependencies, (in)stability, resilience, automation.
  - Regulatory impacts (e.g. driving up costs, stifling innovation).
  - Reputational – 'smear' campaigns, generated by AI and widely shared.
- Economics and finance – e.g. stock market attacks / rumours.
- General public trust and sentiment (e.g. crises – pandemics; Britons abroad).

Vulnerabilities

- All sovereign nations – including the UK – are vulnerable to the global / international nature of Information Disorder.
- In many cases, the sources of disinformation are out of reach of any national Law Enforcement Agencies;
  - there are few – if any – international agreements to address this weakness.
  - Malign actors have the upper hand, here.
- In general, our capacity to apply critical judgment to fake news is low: we lack "ID/Media Literacy";
  - there are examples of good practice – e.g. Finland[5].
- There is no **centralised authority or dedicated policy space** to facilitate both thought leadership and the experimental apparatus to prepare for future threats. This is needed to promote the development and testing of tools to help identify "real" and "reliable" sources of information online; and to develop Trustworthy AI principles and standards.

[5] European Digital Media Observatory. *Mapping the Medial Literacy Sector – Finland*. Author. Available from: https://edmo.eu/resources/repositories/mapping-the-media-literacy-sector/finland/

- Some initiatives from standardisation groups are moving in this space (IEEE, ETSI, ITU, etc.).

- Despite the scale of challenges presented, there is no formal **framework** for quantifying, considering or even measuring possible harms enacted by AI and Information Disorder.
  - we lack the tools to assess the risk effectively: we don't understand the scale and effect of ID operations, nor can we currently map these out;
  - In the absence of any measurement model, targeted action, skills development and effective funding strategies are desirous, but unclear or impossible.
- There is no **dedicated policy or experimental space** to develop, test and consider possible scenarios, strategies and outcomes of the threats posed by AI – ID.
  - For example, it is unclear whether 'counter-narratives' will have their intended effect once disinformation has taken hold, or which strategy would work best under which circumstances.
  - Nor are we clear on the potential consequences of different levels of disruption and distrust sowed by organised and targeted ID campaigns.
- The rapid and unprecedented proliferation of ID requires a concerted **communication strategy and authority** to meet the novel challenges presented.
  - As "noise pollution" increases with the proliferation of information channels, it becomes increasingly difficult for Government to communicate with its people;
  - The provenance of information is vital, as is the intended audience.
  - There is a challenge around how to equip that audience with the relevant capabilities, to consider the provenance of what they are consuming.
  - A robust solution would pre-emptively target increasing narrative multiplicity, drawing on community mindedness, objectivity and mutual verification.

- This would require dedicated leadership (e.g., communication channels, policy, co-ordination) innovative and intuitive design, and possibly technologies, where Government can compellingly stay in touch with its people, even in the worst-case scenario.

- Failure to act threatens multiple consequences:
  - Technological advancement may hit a paradigm shift, where intervention may be very difficult or impossible.
  - Preventative measures and education can help equip us for an increasingly uncertain future.
  - Failure to do so runs the risk where Public opinion and trust may be irrevocably lost, proving difficult or unfeasible to fully reclaim.
  - Expertise, skills, research and resources are best 'front loaded', where possible, to prepare for worst case scenarios.
    - For example: hostile actors misusing malicious AI systems, exploiting prominent (unrelated) AI systems and infrastructure (e.g., supply chains, energy management, communications, finance).

Recommendations

R1 – Set up an 'Office for Media Literacy', with the following remit:

- develop and implement a Risk Management Framework (informed by assessment of harms – R3).
- develop balanced mitigations, such as promotion of greater media literacy in the critical appraisal of news and information; options for action could include amendments to the national curriculum (following with Finnish model) and public messaging campaigns.
- seek to achieve public support and cohesion in this work through shared community participation and peer review.

  Relevant lead(s): Department for Digital, Culture, Media and Sport; Department of Education; Home Office

R2– Plan concept mapping of 'sandbox' space (expertise, technology, policy), potentially hosted at this new Office (R1), to invest in, encourage, and experiment with:

- Blockchain / digital ledger and associated emerging technologies such as AI watermarking, to prepare foundations of '8 tick check' piece, see R5, *below*)
- Lessons Learned: glean insights from marketing sector regarding quantification of *comms effects*; to build a framework for quantifying harms (e.g. for a commercial company to gain half a point of market share, it needs to outspend its competitors on comms by 10%).
- Research diversification – for example,
  - Fundamental / **applied game–theory** type research (experimenting with offensive capability deployments; strategic appearance / action; integration of military / defence); emulating relational approaches.

- Counterfactual 'world ending' **scenario** type research (causal inference informed), with a specific gear towards policy, costing and implementation;
- Procedural / **process driven** evaluation-type research, to explicitly plan for mechanisms, communication lines, contingencies, and preparedness. Examples include:
  - Agent-based modelling and Multi-Agent Systems research, particularly regarding Deception analysis with artificial intelligence[6].
  - Canada's "Fault Lines" report[7], to help inform costing and quantification (costing socioeconomic harms, causal mechanism specification).

- Build in mechanism 'by design', to funnel promising or successful research projects into actual policy experimentation – possibly with challenge-based approaches that incentivise solutions.

  Relevant lead(s): Home Office; DSIT

R3 – Assessment of Harm:

- Develop and test models to measure and assess possible harms, with explicit commitment to help fund related research and bids, built around costed frameworks.
  - Similar to National Risk Register[8].
  - Quantify (possible) harms, for example arising from work following R2.
  - Cross-cutting, interpolated (exploring connections between potential risks following from AI – ID) – R1 may help with co-ordination, R2 with development and incentives.

[6] Sarkadi, 2024. *Deception Analysis with Artificial Intelligence – An Interdisciplinary Perspective*. Computer Science: Multiagent Systems. Available from: https://arxiv.org/pdf/2406.05724.
[7] Council of Canadian Academies (2023). *Fault Lines: Expert panel on the socioeconomic impacts of science and health misinformation*. Author. Available from: https://cca-reports.ca/reports/the-socioeconomic-impacts-of-health-and-science-misinformation/
[8] HM Government (2023). *National Risk Register*. Available from: https://assets.publishing.service.gov.uk/media/67b5f85732b2aab18314bbe4/National_Risk_Register_2025.pdf.

- Commission research from pool of interested academics (NSEC drafting general letter of support).

Relevant lead(s): Treasury; Research policy / central funding.

R4 – Implement a Resilience Programme – proactively work to develop a framework whereby likely targets of AI – ID campaigns are:

- Appropriately conditioned to withstand co-ordinated AI – ID 'testing'. Requires conditional scenario planning, thinking through procedure and process.
- Start thought piece about 'purposeful control of information' – how to communicate 'compellingly'? (watermarking, *8-tick check*; see R5, below).
- Consider strategic and foundational preparation regarding less conventional, deception information capabilities (in responding to possible AI – ID events).
- Begin planning specific mechanisms, outlining the 'killchain' (take down) of responding to AI – ID events, with 'built in' processes for review and improvement.
- Recruit pool of 'vetted' academic / technology 'reservists'; develop embedded expertise / knowledge base, to prepare for potential scenarios (and have mechanisms and plans in place for that event).
  - Example: DSTL Biscuit Book, '*Human-centred Ways of Working with AI in Intelligence Analysis'* as a useful starting point, with the emphasis of more human-AI teaming for deception analysis[9].
  - Example: Implementation guidance research, for piloting optimal regulations and governance in (for example) knowledge sharing. These approaches are useful in both in the absence of historical data, but also when working 'against deception', such as targeted ID campaigns[10].

Relevant lead(s): Home Office; 'new Office'(R1); Cabinet Office; DSIT

[9] DSTL (2023). *Human-centred ways of working with AI in intelligence analysis*. Author. Available from: https://www.gov.uk/government/publications/human-centred-ways-of-working-with-ai-in-intelligence-analysis.
[10] Sarkadi, S (2021). *The evolution of deception*. Royal Society Open Science. Available from: https://royalsocietypublishing.org/doi/full/10.1098/rsos.201032.

R5 – Produce and promulgate guidance for media, news outlets, and the wider public using a confidence framework to assess and test the risk of ID linked to specific News Items.

- Implementation requires careful development and practical planning, involvement of experts and market testing.
- A pilot output as below focuses on trust and confidence, with users directly asked to indicate their (total) level of endorsement.
- Aim to build this into a targeted communications campaign, available for public interaction and third-party verification, to learn more about each step and assess the 'confidence journey' of reported information (and its consumption).

Sample Questions for ~~8~~-Tick-Check Confidence Matrix

	Please tick relevant fields where confident – **if in doubt, do not tick**	*Check*
1	Is this from a trustworthy source?	
2	Do you trust those reporting that information to you?	
3	Is the path from source to reporter clear and trustworthy?	
3	Where there is parallel reporting from trusted outlets, are they likely to fact-check?	
5	Has the information been reviewed and verified by experts?	
6	Are you satisfied that it causes no harm to public interest?	
7	Are you clear that no hostile interest benefits from this information?	
8	Are you sure there is no manipulation/distortion (including by AI)?	
	Total Number of unchecked	

Relevant lead(s): OFCOM, BBC; DSIT

R6 – Implement structural approaches aimed at achieving relevant outcomes

- Design resilient supply chains and diligent procurement; reduce dependencies, practice and develop cautious regulation / legislation.
- Aggressively 'stress test' public services (sandboxing) to disinformation campaigns.
    - Link possible outcomes to scenario planning (linking to R4); design both 'back up' and 'fail safe' options, for optimal positioning against aggressive disinformation campaigns.
    - Outline relevant 'people, systems and procedures' ready to act against coordinated disinformation.
    - … but also have relevant materials (e.g. '8 tick check' awareness and communications) so that the public can be 'self-sufficient' for short to medium term periods (e.g. Network blackouts).
        - Develop robust communication framework, couched in intelligible language (public sentiment), to prepare 'getting ahead' of the threat, in terms of preparedness and messages 'being heard'. ['Public messaging is key': preparedness, and will to respond.]
        - Prepare and plan communication strategy – trustworthy, shared, authoritative, convincing.

Relevant lead(s): Cabinet Office (Resilience Directorate), DCMS, DSIT, Ofcom, NCSC

Annex

Issues and Definitions

Whilst awareness of the range of threats and vulnerabilities is advised for planning, the following issues are important to consider.

Consideration of such can help balance preparedness and responsiveness – avoid being 'too prepared'.

Identification and provenance

- Be aware of, and work to avoid hyper perception and sensitivity to (perceived) threats.
- Reconsideration and robustness framework, assisted through a) specified measurement and comparison, before b) more advanced (counterfactual) modelling ('what if' type scenario planning).
- Provenance certainty – definitively (and correctly) identifying 'an AI threat' – as well as responsive programming (countering).
- Requires 'joined up' strategy and thinking.

Tactical considerations

- The UK is not 'especially' vulnerable to issues of AI and information disorder
- However, effective response (e.g. to other international actors) must consider more robust and variable strategy (conditional deployment).
- Critically, we should retain 'the surprise factor', and remain adaptable to threats of a different play than we may expect or be used to (e.g., authoritarian systems, tactics and strategy).
- At the very least, a more experimental approach, tested in sandboxing / simulations, would help refine and calibrate optimal strategy.

- For example, could consider exploiting attributional issues, namely perceptions from other actors regarding our likely conduct and capabilities.

Mapping the Lifecycle

- In preparing an interventional strategy, it can be useful to map the lifecycle of targeted information disorder operations.
- With the procedural cycle clearly laid out, the entry point, and specific role of AI in that chain, may then be considered, as may the appropriate shape of our response.
- This helps design and target possible interventions and 'gambit' type strategies, for responding to possible threats with confidence and agility.

The future AI landscape

- Concerns may be expressed regarding 'Artificial General Intelligence' (AGI), for example in organisational decision making, collaboration (with human decision makers or other AI systems).
- Extrapolation of such capabilities may raise concerns in the Information Disorder sphere.
- However, it is important not to needlessly inflate the hysteria. Any link between AI and Information Disorder must not be unquestionably extended to general concerns regarding AGI.
- Advances in GenAI, Agentic AI and reasoning, and its combinations, move us towards AGI.
- At present, AGI remains predominantly a **theoretical** construct.
- We may even be in a good position to capitalise on possibilities through focusing on innovation and messaging in the public domain regarding these technologies.

If anything, full advantage should be taken of the 'pre – AGI' era, to ensure the UK remains 'first with the truth', and at the cutting edge of technological capability.

Appendix A: Roundtable Agenda

Agenda

Tuesday 4 February

1200 Registration

1230 Lunch

1345 Plenary: Opening Remarks

- Launch: **Tristram Riley-Smith,** *NSEC*
- Welcome to Cranfield: **David Denyer**, *Cranfield University*
- Intro to NSEC: **Fiona Strens,** *NSEC*
- Introduction to the NSEC/Leonardo Series: **Simon Harwood,** *Leonardo*
- Open Forum: Problem-Owners invited to surface key concerns.

1430 Breakout Groups*: **Understanding the Risk: Unpacking the Issues & Challenges**

1600 Tea Break & Feedback Preparation

1630 Plenary: Feedback from Breakout Groups inc Q&A

1730 Plenary: Lightning Talk + Q&A from **Professor Lorenzo Cavallaro,** *UCL*

1800 Close

1830 Drinks Reception

1915 Working Dinner

Wednesday 5 February

0900 Plenary: Lightning Talk + *Q&A from* **Dan Sexton,** *Chief Technology Officer, IWF*

9030 Breakout Groups:**Managing the Risk: Developing Solutions & Recommended Actions**

1100 Coffee Break & Feedback Preparation

1130 Plenary: Feedback from Breakout Groups

1245 Wash-Up/Close-Down

1300 Lunch

1400 Delegates depart

Appendix B: List of Questions for Breakout Groups

How can we manage the risk from AI used by hostile/malicious actors and proxies to spread Information Disorder[1]?

I. UNDERSTANDING THE RISK

1. **How is AI used to deliver dis-, mis- or mal-information (aka *Information Disorder*[2])?**
    - What difference does AI make here (now and in the future)?
    - Are the AI tools readily accessible to all?
2. **Which hostile actors should we worry about?**
    - States, proxies, and/or criminals?
        - Are proxies always criminal entities?
        - Do state actors have focussed campaigns, or encourage general mischief?
3. **Why/how are we vulnerable in the UK?**
4. **Why does this threat matter?**
    - What does harm look like and what are the consequences? How serious is this for UK national interest?
        - For example, can bots be used to add names to an online petition?
    - Do we understand the scale / effect of AI-enabled Information Disorder operations?
        - Is there a consensus on this; if not, how can this be achieved?

II. MANAGING THE RISK

5. **Who owns the risk, with responsibility for countering the threat?**

    - Which key stakeholders need to be brought together?
        - Who does the UK Government need to partner with (in the UK) to manage this threat?
        - How can we establish strong relationships with like-minded states to tackle this collectively?
    - How do we position ourselves to manage short-term and longer-term risks from these threats?

6. **What tools / processes / systems can help us to detect and understand this threat?**

    - Can you use AI to detect AI (e.g. on social media)?
    - How easy is it to distinguish the intended vs unintended use of AI to cause harm?
    - At what stage does an event become trackable / identifiable; can this be improved?
    - What are the cascade events you need to control from an incident?

7. **What can be done to mitigate the risk?**

    - Strategy: will we get best return on investment through threat & risk understanding, defence or proactive action; or a balanced approach?
        - How can we optimise defence and reduce vulnerabilities?
        - Should we develop an offensive capability?
    - Communications: how can we provide authoritative communication that is trusted to combat dis-, mis- or mal-information?
    - National/International Law: what "take-down" powers exist (or could be established) to counter the threat?

Appendix C: List of Attendees

Tony A	National Cyber Security Centre (NCSC)
Professor Oli Buckley	Loughborough University
Tash Buckley	Royal Holloway University of London / RUSI
Dr Joe Burton	Lancaster University
Dr Marie Cahillane	Cranfield University
Professor Lorenzo Cavallaro	UCL
Dr Giulio Corsi	University of Cambridge
Professor David Denyer	Cranfield University
Dr Matthew Edwards	University of Bristol
Owen F	NCSC
Tom Garnett	Refute
John Guelke	Department of Science, Innovation & Technology
Professor Weisi Guo	Cranfield University
Dr Andreas Haggman	OFCOM
Dr Simon Harwood	Leonardo
Caitlin Healy	Cabinet Office
James Hill	Cranfield University
Dr Duncan Hodges	Leonardo
Professor Martin Innes	Cardiff University
Dr Hamoon Khelghat-Doost	University of Lincoln
Professor Stephan Lewandowsky	University of Bristol
Nick Lindley	Cranfield University
Professor Jesus Martinez del Rincon	Queen's University Belfast
Dr Tristram Riley-Smith	NSEC
Laura Samuels	NSEC
Dr Stefan Sarkadi	King's College London
Dan Sexton	Internet Watch Foundation
Alishah Shariff	Nominet
Dr Rupert Small	Egregious

Dr Jack Steadman	Queen Mary, University of London
Dr Ryan Stendall	Central AI Risk Function - DSIT
Fiona Strens	NSEC (University of Lincoln)
Alasdair Stuart	BBC Media Action
Graham Tooley	Homeland Security Group - Home Office
Dr Adam Zagorecki	Cranfield University
Nina Smith	Home office