Logically—written evidence (FON0068)

House of Lords Communications and Digital Select Committee inquiry: The future of news: impartiality, trust and technology

About Logically

Logically is a British-based tech company that tackles online misinformation and influence operations. We combine cutting-edge artificial intelligence (AI) with one of the world's largest integrated teams of open-source investigators and data scientists. We work with Governments around the world, including the US, UK, and New Zealand to help them identify and manage state-level influence operations, with a focus on electoral integrity. Our support of social media platforms such as Facebook, Instagram and TikTok is delivered through our independent fact-checking unit, Logically Facts.

Executive Summary

- Generative AI has significantly reduced the barriers for entry into executing online disinformation operations, which aim to deliberately mislead people. The scope designing and delivering these was previously restricted by high costs and organisational resource requirements. That is no longer the case. Alongside this, such operations can either 'industrialise' production of fake content, such that people lose faith in their wider information ecosystem or deliver highly personalised disinformation tailored for specific individuals, and with messaging adaptable in real time.

- Moreover, the broader perception of what constitutes disinformation appears for many to have evolved, moving away from its traditional definition focused on deliberate deceit for political, economic, or ideological motives. Organisations involved in managing this problem- including Logically- have faced criticism for purportedly suppressing free speech and being parties to the 'chilling' of legitimate political debate.

- Logically is acutely aware of such risks. We believe the basis of our work is provided for by the principles codified in Article 19 of the International Covenant on Civil and Political Rights. When highlighting likely 'misinformation' on social media platforms through 'Logically Facts' we adhere to International Fact Checking Network (IFCN) standards, ensuring accuracy, non-partisanship, and transparency when assessing the credibility of the claims we assess. Our goal is to help people understand whether the claims within the content they consume are credible. But we are not absolute arbiters of the truth. No fact-checker should be. It is ultimately for users to make that choice.

- Similarly, Logically does not seek to reductively apply the term 'disinformation' to narratives solely on the basis of their veracity. Instead, we see disinformation as a set of behaviours on the part of an actor who is deliberately trying to deceive people. Our technology and services are designed to manage that problem. We look to determine whether there is

clear intent behind the circulation of a potentially harmful narrative, and from there, how any organised effort to deceive is actually being delivered.

- In this context, we believe addressing disinformation requires a layered approach beyond mere content verification and credibility assessments- important as those are. Disinformation campaigns aim to create ambiguity, oftentimes utilising social media platforms for broader reach. Logically's HAMLET framework embodies a response model that integrates AI with human analysis to address the spread of disinformation campaigns, so that an intervention aimed at curbing dissemination can be delivered quickly and efficiently.

- The Online Safety Act (2023), while addressing foreign interference effectively, underrepresents the broader societal impacts of mis- and disinformation. It lacks provisions for combating misinformation that, despite not harming individuals directly, erodes civic discourse and trust. That said, Ofcom does have powers under that Act to require that a proactive approach to foreign state backed efforts to actively manipulate UK public opinion through disinformation be taken by online platforms.

- In using them, Logically believes that Ofcom should develop 'generic behavioural profiles' or defined 'tactics, techniques, and procedures' (TTPs), by Ofcom that are synonymous with disinformation attacks and for social media platforms to demonstrate a capability to identify and mitigate these as part of their overall duty of care under the Act. This approach mirrors successful strategies in the EU Code of Practice on Disinformation, which require signatories to identify and report on the prevalence of specific behaviours synonymous with disinformation campaigns on their services.

What is Generative AI doing to the online information ecosystem?

1. Widespread generative AI availability means that one of the fundamental barriers to entry for launching online influence operations- or efforts to actively manipulate public opinion through deception-has lowered. There is no longer a need to build large teams to create the content that such operations require and share it online – a requirement which previously limited the scale of such operations and made them primarily the preserve of states.

2. The capacity to execute a large-scale influence operation will no longer be constrained by access to appropriately skilled and well-organised human expertise. AI can largely automate content production, reduce the overhead in persona creation, and generate culturally appropriate outputs- be those images, text, or audio that are less prone to exhibit conspicuous signs of inauthenticity.

3. A substantial growth in social media-based influence operations will have an impact, and not just in limited online forums. There is credible evidence that such operations can cause noticeable shifts in political beliefs and behaviour and increase xenophobic or discriminatory sentiment. Short-term shifts in social media activity by extremely prominent actors can also have modest but statistically detectable impacts on racially motivated violence.[1]

This is particularly pertinent given that many people now consume news through social media rather than traditional news media. Any discussion on the future of news must therefore address the integrity of these platforms against such operations.

4. Generative AI also delivers a step change in another essential element of a typical online influence operation: the ability to tailor content by audience in a far more effective way. This can be done through techniques such as entity/target-based sentiment analysis and stance detection, which facilitate the automated review of the tone of what a person is posting (for example, sarcasm, confusion or suspicion) and also interpret their view as negative, positive, or neutral in terms of what it is describing or the perspective it is expressing towards the concerned entities or targets.[2]

5. Today, this kind of assessment depends largely on the inference of information, such as political leanings, from other contextual data such as groups and people that someone follows on a social media platform by the actor behind the influence operation. AI-powered sentiment analysis bypasses or at least eases this task. An assessment of a user's leanings through their actual posts online becomes much simpler and faster using AI-powered tools. This allows for a more accurate assessment of what they will be susceptible to.[3]

6. Generative AI also provides a much more efficient way for those launching influence operations to determine what type of content they should develop to achieve their outcome among their intended audience. An online influence operation could copy the social media profiles of people with interests compatible with the narrative they want to promote and use that to prompt a generative AI model to generate the content associated with it, delivering content precisely targeted to that audience.[4]

7. Automated content can be tailored to audiences to ensure the narrative is more closely calibrated to the values and beliefs of those audiences, without having to compromise on the strategic aims of the narrative. Rather than producing just a handful of articles a day, one article can be produced and tailored to 12 different audiences taking five minutes in each case.[5]

8. Generative AI also provides the capability to generate and disseminate huge volumes of content that is not tailored to its recipients at all. We refer to this as 'flooding the zone'. This is already a strategic approach that China is known to use.[6] It is essentially the fully automated generation of vast amounts of content quickly and efficiently with the goal of flooding online platforms with misleading information, disinformation, and spam, leading to decreased trust, confusion, and potential harm to individuals and societies.

---

1 Bateman et al. *Measuring the Effects of Influence Operations: Key Findings and Gaps From Empirical Research (June 2021)*

2 Sedova et al. *AI & The Future of Disinformation Campaigns (December, 2021)*

3 Ibid

4 Goldstein et al. *Generative Language Models & Automated Influence Operations: Emerging Threats and Potential Mitigations (January 2023)*

5 Associate Prof. Kate Starbird in *Thor Benson, This Disinformation is Just for You (*Wired Magazine, August 2023)

6 The Guardian*, Meta closes nearly 9,000 Facebook and Instagram accounts linked to Chinese 'Spamouflage' foreign influence campaign, (29 August 2023)*

9. The content may, on an individual level, be of low quality or not persuasive. But that is not always necessary. If the goal of the actor behind an influence operation is simply to generate generic mistrust and doubt, it can often be the volume of content that is generated and the polluted information environment this creates that leads to mistrust. Its quality is not what achieves this, as the persistence of "cheapfakes" even today demonstrates.

10. The scope for generative AI to do exactly this, in a fully automated way was recently demonstrated through the creation of a 'proof of concept' system called CounterCloud. CounterCloud's AI identifies articles by specific publications and journalists according to its operator's brief. It then automatically prompts an LLM to generate content based on the same themes, with specific perspectives.

11. The system generates fake comments, images, and sound clips to go with the article, as well as generating Twitter (now "X") posts to promote the website hosting it, counter opposing tweets via trolling, or promote positive narratives about what is being said. The total overall cost to deliver the whole system was US$ 400.[7]

12. We have already started to see this kind of approach used by hostile states. A recent report identified "Election Watch" as an inauthentic English-language political news outlet targeting US audiences with content specific to US election cycles, political campaigning, polling, and more. Advertising itself as the "go-to source for everything election-related", the website attempted to present itself as a balanced and non-partisan source of perspectives and issues in US politics.[8]

13. Much of the content was wholly AI generated. Traffic to Election Watch was in turn being driven by the Russian-linked online influence operation known as 'Doppelgänger' which comprises at least 2,000 inauthentic social media accounts.

Avoiding politicisation of 'disinformation'

14. The Committee is rightly concerned that the concept of disinformation has increasingly become associated not with the identification of deliberate attempts to deceive people for either a political, economic or ideological reason but some other abstract notion of what it is legitimate for a person to say or a media outlet to cover. Logically is acutely aware of the distinction and has made the maintenance of it a central element of the work we do.

15. Until September 2023, Logically was a contractor for DS IT's Counter Disinformation Unit (CDU), now known as National Security Online Information Team (NSOIT), aiding in the detection of foreign interference

---

7 Wired Magazine, *It costs just $400 to build a disinformation machine,* (August 2023)

8 Recorded Future, *Obfuscation and AI Content in the Russian Influence Network "Doppelgänger" Signals Evolving Tactics* (December 2023)

and supporting the integrity of civic discourse. We continue to support other UK Government Departments in preserving electoral integrity globally.

16. Since 2019, Logically has also been actively managing disinformation threats to the US federal and state elections, including the midterms, as well as elections in India. We are supporting three US State Governments in preparing for the upcoming US presidential election, as well as the Election Commission of India during the ongoing general election.

17. In addition to becoming a member of the US EI-ISAC in 2022, where we contribute to strategies against election-targeted mis- and disinformation, Logically collaborated with diverse partners during the 2022 US midterm elections to safeguard electoral processes, maintaining close contact with law enforcement. Moreover, we have extended our expertise to the EU, recently partnering with the Slovak Government to navigate potential generative AI disinformation impacts on voter trust in their presidential election.

18. Logically, and other entities that do similar work, have been the subject of persistent criticism for the work we have done in this context. There has been a suggestion that counter-disinformation efforts by public authorities amount to surveillance and government censorship of critics and dissenters, and that contractors delivering them are therefore party to such concerns. Campaigners have claimed that a right to freedom of speech is violated because even where the public authority does not order content to be blocked or taken down by social media platforms, the simple fact that they issue reports highlighting certain content has a "chilling" effect.

19. In that vein, there have been suggestions that reports created by entities like Logically for counter-disinformation efforts, and the fact-checks that Logically Facts does for platforms in relation to online misinformation, lead to the 'flagging' of online posts and content that are not obvious examples of misinformation and disinformation, but merely expressions of opinion on subjective matters. It has been suggested that inclusion of such posts and content in 'disinformation' reports shows that the objective of these efforts is to stifle "legitimate political debate".

20. Finally, the more extreme critics of any attempts to manage mis- and disinformation efforts online suggest that content moderation by social media platforms in itself amounts to censorship as it prevents different viewpoints from being heard. Some political actors have suggested that mainstream social media platforms and the fact checking organisations they partner with have a "liberal" or "woke" bias and their assessments are prejudiced against those that do not share such views.

21. In our view, there is a clear need to ensure that any organisation involved in assessing misinformation- or potentially inaccurate information which is not circulated with malicious intent and which a person may genuinely believe- focuses firstly on equipping ordinary users of social and other media with the tools they need to make their own decisions. They should be empowered to determine the accuracy or veracity of the information they consume.

22. This goal is at the centre of Logically's mission as an organisation and, in our view, it is what organisations who work to counter misinformation should aim to achieve. Our commitment is to assess the credibility of a claim, and to do so according to defined criteria without any political agenda. All of the fact checks that Logically Facts does are published to ensure such efforts are scrutinised.

23. Moreover, when we speak about 'disinformation', we are firmly focused on clear efforts- either by foreign states or domestic actors- to actively manipulate public opinion in a disingenuous way for their own ends, or to actively cause physical harm to people in the real world. The key element for us is 'intent', or more specifically the online behaviours that signify this intent, and not just the nature of whatever the narrative at the centre of the content concerned is. It is very important to keep this context in mind when trying to understand the problem facing us, and how to ensure we don't let the solutions become worse than the problem.

Who fact checks the fact checkers? The risk of overreach

24. We welcome the scrutiny our work receives, and we are happy to discuss it. As a starting point, we believe that the work we do is supportive of free speech. We consider Article 19 of the International Covenant on Civil and Political Rights (ICCPR) to be our guiding principle for understanding free speech, and its facets are central to any public explanation of our work. Crucially, the right to freedom of speech and expression under this international law standard includes the freedom to "seek, receive and impart information and ideas of all kinds". It therefore not only protects freedom of speech but also recognises the need to protect people's right to receive information.

25. While we cannot comment on the policies, approaches and definitions used by other organisations that do the same kind of work we do, Logically does believe that the safeguards that we have adopted and continue to use minimise any risk of overreach. We are an organisation that empowers people to receive accurate information and ensures that efforts to deceive people through disinformation are identified early and dealt with.

26. First and foremost, we have adopted an Ethics Charter, which prohibits, among other things: Working with government clients in countries with unacceptably low levels of democracy and the concomitant risks of human rights violations that arise in such countries. We also do not work with private clients in countries with unacceptably low levels of democracy and insufficient protections for the rule of law. We have, and will continue, to decline work that does not meet this strict criterion. Equally importantly, we do not work with political parties, or any religious organisations in *any* country.

27. Logically and Logically Facts do not recommend actions to be taken pursuant to the products and services we deliver. So, if we issue a fact-check, or we conduct an investigation, the decision of how to respond is left to the sole discretion of the relevant social media platform or public

authority. While we do not make the decision, it is supposed to fit within the framework of a predetermined scope of work, which has to comply with the prohibitions in our Ethics Charter.

28. The reports provided by Logically to public authorities adhere to contemporary best practices and standards, ensuring we can be reasonably satisfied that they are devoid of inaccuracies or bias. Our analysts have been trained on and follow the Obsint Guidelines for Public Interest Open-Source Intelligence (OSINT) Investigations,[9] which includes commitments to principles of accuracy and accountability and encompasses the first effort that has been made to apply an ethical framework for public interest investigations using Open-Source Intelligence.

29. These Guidelines – which Logically played a key part in drafting along with other OSINT experts in Europe – require analysts to ensure their research and subsequent outputs are clear about their sources and their limitations and are presented in a way that allows readers to replicate the work as far as practicable. All reports go through multiple levels of review for quality assurance.

30. Our reports will sometimes flag posts and content which do not themselves amount to misinformation or disinformation. We believe these need to be included to ensure our clients have an accurate understanding of the context and the broader information environment in which narratives are being spread. However, Logically does not label such posts or content as false or misleading and does not recommend that action should be taken against the creators of such posts or content.

31. In its work for social media platforms, Logically Facts adheres to the International Fact Checking Network (IFCN) Code of Principles, to which it is a verified signatory.[10] This includes commitments to non-partisanship, fairness and transparency that apply to every stage of its work, from selection of claims to check to the fact checks it publishes. It also includes a commitment to transparency of funding, with Logically Facts stating publicly where it gets its funding from. All these measures ensure that Logically Facts can justify its work in the field of fact checking and build trust in its work among the general public.

32. The IFCN Code of Principles are the current gold standard to ensure that a fact checker makes all reasonable efforts to avoid inaccuracy or bias, as well as ensure its work can be verified by readers and assessors. Since 2020, the company's adherence to the Code of Principles has been affirmed by the IFCN following its standard review process. Logically Facts has also sought to ensure it complies with regional fact checking standards, including those of the European Fact Checking Standards Network (EFCSN) and the Misinformation Combat Alliance (MCA) in India.

---

9 https://obsint.eu/guidelines-for-public-interest-osint-investigations/

10 These can be viewed at https://www.ifcncodeofprinciples.poynter.org/the-commitments. Logically obtained verified signatory status from the IFCN in 2020, at a time when there was no separation of entities performing OSINT and fact checking work. In 2023, following its formation as a separate entity, Logically Facts filed the application for continuation of IFCN accreditation.

33. Further to the standards in the IFCN Code of Principles, Logically Facts has a robust complaints and corrections policy, which ensures that it can correct any errors or mistakes made in its fact checks. There is no restriction on who can file a complaint or suggest a correction to a fact check, and the fact checking team responds to all complaints received from a variety of different channels. Following the review of a complaint, if it is found to have merit, not only is an update or correction made to a fact check, but this is also explicitly acknowledged by Logically Facts in the fact check itself and on social media.

Online disinformation: What problem do we need to solve? And what is the role of AI in that?

34. Often, the online disinformation threat that we face focuses on whether content behind the campaign is human-generated or not, or whether it is 'true or false'. But it is important to understand that these kinds of questions miss the fundamental objective that has driven – and continues to drive – most active disinformation operations.

35. The explicit aim of most online influence operations is to shape narratives and perceptions, often involving the spread of ambiguity and confusion through mis- and disinformation. Social media is a prevalent means to launch such influence operations, often subsequently amplified in the mainstream media.[11]

36. Because of this, the primary harm of such campaigns is more complex than people simply viewing 'fake' material, and their management has a range of different facets to it. We must move beyond an approach that attempts to authenticate as much material as possible. As the amount of AI-generated content grows, this will become increasingly difficult to do with any certainty.

37. At a basic cognitive level, we are all susceptible to disinformation. Social psychology tells us that regardless of whether a piece of content is true or false, a person's belief in the truth of that claim increases the more that it is repeated and seen by them. In essence, the more viral content is seen, the more the audience believes it is true. This is known as the 'illusory truth' effect.[12]

38. Combined with this, when people are repeatedly exposed to mis- or disinformation, those claims enter memory and gain credibility. This effect appears to hold true regardless of an individual's capability or need for cognition.[13]
39. One landmark study looked at 126,000 rumours spread by ~3 million people on Twitter over a nine-year period. It showed definitively that false news (as verified by six independent fact checkers) reached more people than the truth did. This was primarily because the false news was more

---

11 Dowse & Bachman, *Information Warfare: methods to counter disinformation*, (Edith Cowan University, 2022)

12 Hasher et al. *Frequency and the conference of referential validity*, Journal of Verbal Learning and Verbal Behaviour (1977)

13 De Keersmaeker et al. *Investigating the robustness of the illusory truth effect across individual differences in cognitive ability, need for cognitive closure, and cognitive style* (2020)

novel than that which was true. The authors were also able to show that it takes about six times as long for a true statement to reach 1,500 people over the platform as it does for a false one, and that falsehoods are about 70% more likely to be retweeted than true claims.[14]

40. Moreover, 'bot' accounts accelerated the spread of true and false news at the same rate, implying that false news spreads more than the truth because humans, not robots, are more likely to spread It. When people are repeatedly exposed to disinformation, it tends to alter their memories, which becomes especially pronounced during emotive political events.[15]

41. Recent research shows that the illusory truth effect may be even more powerful than was previously believed, and that efforts to base our response to increasingly widespread forms of disinformation like "deepfakes" on simply labelling their contents as fake are not likely to be sufficient on their own. One forthcoming study from Dr. Simon Clark and Professor Stephan Lewandowsky have looked at the effectiveness of warning people that they were about to watch deepfake videos before they did so.

42. Participants in their study were shown a bespoke deepfake video of a supposed local Government official appearing to admit to accepting a bribe. The viewers were asked questions about the official's guilt to test how they had been influenced by the video's content.

43. Clark and Lewandowsky found that, in assessing the man's guilt, people relied on the content of the deepfake video they had just watched, even when they had been explicitly warned beforehand that it was fake. This result was observed even with participants who indicated that they believed the warning they had been given. In other words, they knew the video to be fake. But they still used it as a reference point. The study found this specific warning to be no more effective than a more generic warning about the existence of deepfake videos that a separate group of participants was given.[16]

44. These findings suggest that identifying and flagging deepfake videos will not entirely negate their influence, or the impact they might have in terms of the promotion of overarching societal polarisation. Labelling and provenance systems of the kind that have been announced by a number of social media platforms and AI companies are helpful. But they will not deal with the potential societal harms that circulation of such content can cause. The innate susceptibility that we have towards disinformation may be more powerful than many people believe.
45. When we combine these predispositions with the features that social media platforms are able to deploy to make accurate assessments of what you would like to see on their platforms, the solution that must be adopted becomes a lot clearer. Through a cross-section of a person's online search history, unique click behaviour and other digital footprints, social media

14 Vosoughi et al. *The spread of true and false news online,* in Science (March 2018)
15 Ibid
16 Clark & Lewandowsky, *Seeing is believing: the continued influence of a known AI Generated deepfake video* (Bristol, 2024). Note: This article is still undergoing peer review and will be publicly available shortly.

platform algorithms can make a fairly accurate assessment of what will engage their users and tailor content accordingly.

46. This is helpful for shopping and research, but evidence shows that 'filtering' also happens with search and news results, thereby impacting the spread of disinformation. Twitter, now known as 'X', has previously published research which found that its users' personalised feed amplified biased content because the algorithms that created it were optimised for user engagement, not accuracy.[17]

47. To control for all these phenomena, we need a portfolio of tools that can act alongside an effort to try to verify whether AI-generated content is real or not. We are inherently vulnerable to disinformation being 'served' to us over and over again, whether that is through the 'filter bubble' or 'echo chamber' phenomenon, or because of a coordinated effort by a foreign state or non-foreign state actor. The problem we need to solve is one of virality.

48. Logically's approach is therefore centred on interventions that seek to both identify harmful content and ensure that disinformation is curbed. Put simply, evidence suggests that the harm that disinformation causes is correlated with how widespread its dissemination becomes - not simply its existence. The earlier a proactive intervention comes, the more likely the harm is going to be controlled.

49. There are a wide range of differing responses that could be taken, including a social media platform removing the content, reducing its algorithmic amplification, or disabling the accounts circulating it. What matters is a process by which disinformation can be identified quickly, so that a response can be deployed and implemented quickly, and ideally be specifically designed to address the disinformation in question.

50. The best way to tackle disinformation risks, including those linked to generative AI, lies in leveraging AI not just to detect 'fake' content, but by using AI models to identify the types of coordination and the tactics, techniques and procedures (TTPs) that disinformation actors (and the models they train) use to actually disseminate that content in front of audiences via social media platforms.

51. Logically has developed a human-in-the-loop AI framework called Human and Machine in the Loop Evaluation and Training, or HAMLET, that allows AI models and human expertise to be combined in order to tackle this problem through continuous input and feedback cycles. Logically's research on the HAMLET framework and its applications to solve the problems in contemporary information environments recently won a 'Best Paper' award at the NATO IST-195 Research Symposium in October 2022.[18]

52. As that paper sets out, we do not believe that AI can – or should – be used to determine the truth. However, AI can flag and prioritise patterns of

---

17 https://blog.twitter.com/en_us/topics/ company/2021/rml-politicalcontent

18 This paper is available at https://www.logically.ai/download-tackling-misinformation-with-hamlet-an-expert-in-the-loop-ai-framework

behaviour that an expert human analyst needs to investigate further to see if they can see evidence of an influence operation occurring.

53. In a fact checking context AI can also be used to 'triage' what a fact checker looks at first, by shortlisting audio-visual content which contain contentious claims that are likely to be fact check worthy. This is very important in a world where the scale of content that can be produced is likely to grow exponentially. Again, AI cannot be used to actually assess credibility, but can help make the fact check process more efficient.

54. NATO's Strategic Communications Centre of Excellence has now published a follow-up guide to how disinformation experts should use AI effectively to combat online influence operations, which is largely based on the capabilities that Logically has developed to identify and map disinformation using the HAMLET model. Logically conducted much of the research underpinning this work.[19]

55. Through HAMLET, Logically is applying AI to the 'dissemination' and 'response' problem. This is achieved through applying a portfolio of different AI and machine learning capabilities, many of which are unique to Logically. Crucially, HAMLET is content agnostic, in line with our view on the importance of tracking the means and characteristics of dissemination rather than focusing on the content itself. This also means that its principles can be applied across geographies and languages, with suitable adjustments where required.

What is the role of Government and regulators in addressing online mis- and disinformation?

56. The primary statute that currently seeks to address the harms of online mis- and disinformation is the Online Safety Act (2023). However, it is notable that the new regulatory regime it establishes only looks at a subset of these harms that applies in very specific circumstances. Comments of one former Secretary of State, Sir Jeremy Wright MP, as to why this may have been the case are particularly illustrative.

57. He is quoted as saying that its omission was partly due to confusion between the remits of different departments. The Department of Culture, Media and Sport "was told that the Cabinet Office would be taking care of all of this. "Don't you worry your pretty little heads about it, it'll be done elsewhere in something called the Defending Democracy agenda,'" he says. "And then I think, subsequently, it wasn't really".[20]

58. In Logically's view, this was an important omission from the Act, and one that a future Government should revisit. Fundamentally, a future Government must address the fact that mis- and disinformation are societal harms. The Online Safety Act is based around the idea that a duty of care is applied solely to promoting safety and preventing harm at an individual level. It explicitly does not cover misinformation that causes societal harm,

19 This paper is available at https://stratcomcoe.org/publications/ai-in-support-of-stratcom-capabilities/296

20 Wired Magazine, *The UK's Controversial Online Safety Act Is Now Law* (October 2023)

even though evidence suggests that such content does just that. The Act's focus on mitigating the direct harm caused by individual pieces of content to individual people risks doing nothing to address the more pervasive indirect harms caused to individuals through the wider degradation of the civic information environment.

59. However, one element of the Online Safety Act that can effectively address a major contemporary concern is the designation of 'foreign interference' as a priority offence under it. In effect, the Act creates an obligation for social media services to proactively monitor their platforms and remove content associated with the offence (which is outlined in the National Security Act, 2023) before users encounter it. It is the task of Ofcom to assess whether they are doing so effectively, ensuring that there will be accountability.

60. The main challenge for Ofcom is to determine a suitable process or methodology for this task. To accomplish this, they must establish a shared understanding of what constitutes "reasonable inference" of foreign interference - which is the threshold under which content is deemed to be ''illegal content" under the Act - and thus subject to the aforementioned obligations for platforms to proactively identify such campaigns and take them down.

61. Logically has consistently advocated that Ofcom can do this through the development of 'generic behavioural profiles' or defined "tactics, techniques and procedures" (TTPs) for foreign interference. In essence, these are a set of online behaviours that are synonymous with foreign influence operations and apply regardless of the social media platform in question.

62. We believe these could act as the foundation for a minimum set of requirements that all social media platforms must be able to identify in order to satisfy their duty of care to manage the risk of foreign interference via disinformation and determine whether they need to take action to remove it in accordance with that duty under the Online Safety Act.

63. There is a clear precedent for such profiles in the EU Code of Practice on Disinformation. It requires signatories, which include all major social media platforms apart from 'X', to reach a mutual understanding of, and implement policies against, manipulative TTPs used in the context of foreign interference and report on their proactive efforts in detecting and mitigating these. Logically is a signatory to this Code and has played an active role in its development.

64. Under this Code, an initial list of TTPs synonymous with disinformation was published in June 2022. This was adapted from DISARM, an open-source framework aimed at combating disinformation by facilitating data and analysis sharing, as well as the coordination of effective action. Logically is part of the subgroup responsible for regularly reviewing the list to incorporate the latest evidence on the TTPs employed by malicious actors.[21]

[21] More details of the specific TTPs that platforms have agreed to track and report on are at: https://disinfocode.eu/reports-archive/reports-july-2023/?chapter=integrity&commitment=commitment-14

65. This subgroup has also developed metrics that are applied in the EU to measure the performance of platforms in tackling these TTPs, called Service-Level Indicators. These are essentially common metrics to assess how adequately platforms are identifying and effectively mitigating disinformation risk.

66. This work is important for Very Large Online Platforms (VLOPs) - defined as platforms with over 45 million users in the EU under the bloc's Digital Services Act (DSA) - as they are subject to more stringent risk assessment and mitigation rules. The European Commission plans to convert the Code of Practice into a Code of Conduct under the DSA, whereby it will serve as a co-regulatory tool under the law. In practice, this means that VLOPs will be audited on their compliance with the Code should they sign up to it.

67. In practical terms, the TTPs set out in the EU context can serve as a baseline example from which Ofcom and the Government can work together on generic behavioural profiles. Broadly, this would facilitate the identification of hallmarks of potential foreign interference, upon which a qualified OSINT analyst can build via contextualisation and attribution.

68. In our view, it is for Ofcom to set out its expectation of social media platforms to demonstrate their capability to identify these generic behaviour profiles, and for them to do so through the risk assessment and mitigation process prescribed by the Act. Certainly, this appears to be what Ministers intended when they implemented the relevant legislation. They stated that the intention was to "compel companies to take action against a broader range of state-sponsored disinformation and state-linked platform manipulation."[22]

69. We are concerned that Ofcom's current proposals stop short of this. Their current draft proposals suggest that "while we know more about how these operations are carried out on certain services, this is not necessarily an accurate reflection of the presence of the harm" and that "evidence available does not allow [us] to draw robust conclusions about the impact and harm associated with foreign influence operations".[23]

70. This appears to run counter to both the Government's view, and the National Cyber Security Centre (NCSC) assessment in its annual report that democratic events, such as elections, almost certainly represent attractive targets for malicious actors and so organisations and individuals need to be prepared for threats, old and new.[24]

71. Instead, Ofcom outline a "series of questions that platforms should ask themselves about foreign interference", without providing an overarching legal threshold that would trigger services to take proactive measures against foreign influence campaigns. We do not believe that this is a sufficiently robust approach.

---

22 HL Deb, 6 July 2022, cWS

23 Ofcom, *Protecting people from illegal harms online Vol 5*, (November 2023)

24 NCSC Annual Report 2023

72. Importantly, as platforms only need to make a "reasonable inference" there is no requirement to prove *mens rea* to a criminal standard in the context of their Online Safety Act obligation. This also means that Ofcom should not look to base its regulation of foreign interference on granular assessments of the content in question. They should instead consider regulating on the basis of a defined set of behaviours that, when combined, allow a "reasonable inference" that foreign interference is taking place.

73. The other key area that Ofcom now has powers in relation to is the False Communications Offence. This, according to the Government, makes it illegal for a person to post content on a social media platform that they know is not true, and that is intended to cause "non trivial psychological or physical harm to a likely audience". As with foreign interference, the onus is now on Ofcom to offer effective practical guidance on what kinds of online content are actually illegal under this offence. In their draft regulatory guidance on illegal content, Ofcom clearly states that "misinformation – that is, misleading or untrue information which is shared by a user who genuinely believes it to be true – is not captured by this offence".

74. So the question is: what does it cover? Ministers have repeatedly been questioned on this point. At various junctures during the Act's passage through Parliament, they described scenarios where a "hoax bomb threat" would constitute illegal content under the offence. Other examples are that the offence made it illegal to knowingly circulate the idea that "drinking bleach cures COVID-19" when the person knows that not to be the case, and intends to cause harm.

75. Ofcom rightly points out in its draft guidance on this topic that "it will be challenging for service providers to make these judgements on their own''. This will no doubt be particularly true for determinations of the intent element. But given that this offence now exists, it would appear appropriate for Ofcom to set out further details on the kind of content that it would consider to cause "nontrivial" psychological harm, and the means that services might need to have to assess whether content meets the 'false communications' threshold.

76. This is a very subjective term. What will cause psychological or physical physiological harm for one person may not cause it for another. As we have noted, international standards on fact checking offer a framework for determining whether a statement is credible or not credible, but Ofcom will need to consider if this can be adapted for the purposes of the offence.

77. Even for the difficult question of intent, we believe that the principles outlined by us can prove useful to Ofcom's approach. Whether the person used well known TTPs to circulate the content is quite often an indicator of whether they were spreading that content in a malicious manner. Therefore, a framework for determining intent could consider whether well-defined TTPs associated with disinformation have been used. However, this needs to be considered further.