Written evidence submitted by Dr Akshi Kumar (SMH0040)

Senior Lecturer in Computer Science, Director-Post Graduate Research (PGR), Department of Computing, Goldsmiths, University of London, United Kingdom

Science, Innovation, and Technology Committee
Inquiry
On

Social media, misinformation and harmful algorithms

Algorithmic Amplification and Public Trust: A Call to Action Against Digital Harm

Key Terminology:

- **Misinformation**: False or inaccurate information spread without intent to deceive, often leading to unintended harm.
- **Disinformation**: Deliberately false or misleading information created to deceive or cause harm, often targeting societal divisions.
- **Algorithmic Transparency**: The practice of making the mechanisms behind content-ranking algorithms visible and auditable to ensure they do not amplify harmful narratives.
- **"Legal but Harmful" Content**: Content that is not illegal but has significant potential to harm individuals or society, such as hate speech or health misinformation.
- **Generative AI**: Advanced AI systems capable of creating content (e.g., text, images, videos), with potential to both combat and amplify misinformation.

Executive Summary:

- **Impact of Algorithms and AI**: Social media content recommendation systems and algorithms amplify both misinformation and disinformation, as seen in the 2024 UK riots, highlighting their role in spreading harmful content and societal unrest.
- **Online Safety Act 2023**: Strengthens regulation by empowering Ofcom, mandating transparency, and penalizing non-compliance but struggles with enforcement, global jurisdiction challenges, and addressing "legal but harmful" content.
- **Challenges and Gaps**: Rapid AI evolution outpaces regulation, enforcement is reactive, and cross-border governance remains weak, limiting the framework's effectiveness.
- **Proposed Solutions**: Broaden the regulatory scope, deploy real-time AI tools, promote media and AI literacy, and establish global collaborations for cohesive online safety measures.
- **Generative AI Vision**: Harness AI for transparency, misinformation detection, and empowering critical thinking to build a safer digital ecosystem.

I. How Effective is the UK's Regulatory and Legislative Framework on Tackling These Issues?

The UK has made considerable efforts to tackle the proliferation of harmful content through legislative and regulatory frameworks, but effectiveness remains mixed due to challenges in implementation, technological complexity, and globalized internet platforms.

1. **Strengths of the Current Framework:**

   1.1. **Online Safety Act 2023[1]:**

   a) Establishes clear duties for online platforms to mitigate illegal and harmful content, with a focus on protecting children and vulnerable groups.

   b) Empowers Ofcom to regulate and penalize platforms for non-compliance.

   c) Creates requirements for transparency reports, helping to expose harmful algorithmic practices.

   1.2. **Data Protection Act 2018 (DPA)[2]:** Ensures responsible handling of user data, indirectly curbing misuse for targeted misinformation campaigns.

2. **Challenges:**

   2.1. **Global Platform Dynamics:** Platforms are subject to UK regulations under the Online Safety Act if they have a significant number of UK users, target UK audiences, or host content posing risks to UK users. However, enforcement can still face practical challenges due to conflicting international legal frameworks, compliance resistance, and the global scale of these platforms[3].

   2.2. **Monitoring Challenges:** Even with Ofcom's expanded role, the sheer volume of online platforms—ranging from global giants to niche services—makes comprehensive oversight complex. Resource limitations, evolving content types, and monitoring smaller platforms further strain enforcement efforts.[4].

   2.3. **Lagging Adaptation:** The rapid pace of technological advancements, particularly in areas such as generative AI, deepfakes, and manipulated media, makes it difficult for legislation to remain current and effective.[5,6]

   a) *Deepfakes:* Increasing realism will make it harder to differentiate authentic content from fabricated material. Applications are expected to diversify, ranging from political impersonations to corporate fraud and personal attacks.

   b) *AI-Generated Content:* A surge in AI-produced text, including fake news and propaganda, will likely flood digital platforms. Such content can also be hyper-targeted to specific demographics, amplifying its potential impact.

   c) *Manipulated Media:* Subtle edits and hybrid content, combining real and fabricated elements, will distort narratives, challenging detection mechanisms and influencing public perception.

   2.4. **Evolution of Dissemination Platforms and Channels**:

[1] https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer

[2] https://www.legislation.gov.uk/ukpga/2018/12/contents

[3] https://library.oapen.org/bitstream/id/4c449edf-6312-428c-94ed-45aa230602cf/978-3-030-95220-4.pdf

[4] Judson, E., Kira, B., & Howard, J. W. (2024). The Bypass Strategy: platforms, the Online Safety Act and future of online speech. *Journal of Media Law*, 1-22.

[5] Bontcheva, K., Papadopoulous, S., Tsalakanidou, F., Gallotti, R., Dutkiewicz, L., Krack, N., ... & Verdoliva, L. (2024). Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities.

[6] Montasari, R. (2024). Responding to Deepfake Challenges in the United Kingdom: Legal and Technical Insights with Recommendations. In *Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution: Threats, Assessment and Responses* (pp. 241-258). Cham: Springer International Publishing.

a) Social media algorithms may unintentionally amplify misleading or sensational content. Niche platforms and private messaging apps, such as WhatsApp and Telegram, present additional challenges due to encrypted communication and the rapid spread of rumours within closed groups[7, 8].

b) Misinformation spreads differently across platforms[9] due to their unique features—such as rapid, real-time sharing on X (formerly Twitter), video-based trends on TikTok, and private messaging on WhatsApp. Effective regulation must account for these differences by adopting platform-specific strategies that address their distinct dissemination patterns and user behaviours.

2.5. Emerging Technologies as Misinformation Vectors: Emerging technologies such as artificial intelligence (AI), generative AI, blockchain, and quantum computing evolve rapidly, often outpacing existing regulatory and legislative frameworks. To ensure timely and effective responses to the challenges posed by these technologies, agile legislative mechanisms are critical. These mechanisms aim to create flexible, adaptable, and proactive governance structures that keep pace with technological advancements while safeguarding public interests.

II. How Effective Will the Online Safety Act Be in Combating Harmful Social Media Content?

The Online Safety Act (OSA) 2023 represents a pivotal step in regulating harmful content on digital platforms, introducing robust measures to protect users and enforce accountability. However, its effectiveness will hinge on its implementation, adaptability to emerging technologies, and international collaboration.

3. Positive Impacts

3.1. Increased Accountability: The Act mandates that platforms actively identify and remove harmful content, with significant penalties for non-compliance. Companies face fines of up to 10% of their global revenue, creating a strong financial deterrent[10].

3.2. Risk Assessments and Transparency[11]:

a) Platforms are required to conduct regular risk assessments of their algorithms and content moderation policies, ensuring that they are aware of and addressing risks proactively.

b) Transparency reports will provide visibility into the moderation practices of platforms, helping regulators and users understand how harmful content is managed.

4. Limitations

4.1. "Legal but Harmful" Content:

[7] Andrey, S., Rand, A., Masoodi, M. J., & Tran, S. (2021, May). *Private Messaging, Public Harms*.

[8] Kalogeropoulos, A., & Rossini, P. (2023). Unraveling WhatsApp group dynamics to understand the threat of misinformation in messaging apps. *New Media & Society*, 14614448231199247.

[9] Bragazzi, N. L., & Garbarino, S. (2024). Understanding and Combating Misinformation: An Evolutionary Perspective. *Available at SSRN*.

[10] https://www.gov.uk/government/news/britain-makes-internet-safer-as-online-safety-bill-finished-and-ready-to-become-law

[11] https://www.gov.uk/government/publications/draft-statement-of-strategic-priorities-for-online-safety/draft-statement-of-strategic-priorities-for-online-safety

a) The Act now shifts its focus entirely to illegal content, removing the previous expectations for platforms to moderate "legal but harmful" material. While this content can significantly harm individuals and society, it no longer falls within the regulatory scope of the Act.[12]

b) The exclusion of this category limits the scope of protection against pervasive harms, particularly in areas like mental health, misinformation, and societal division.

c) Vulnerable groups, including minorities and children, remain at heightened risk from exposure to such content, which can perpetuate discrimination or mental health issues[13].

d) Platforms are no longer obligated to mitigate the spread of content with potential long-term societal impacts, such as extremist rhetoric or conspiracy theories.

e) *Example 1:* During the 2024 riots, xenophobic and anti-immigration content spread widely, intensifying societal tensions. While this content was not explicitly illegal, its harmful impact was undeniable. Social media platforms grappled with the challenge of moderation, often hesitating due to the absence of clear legal guidance—a stark reminder that one person's hate speech can be perceived as another's free speech.

f) *Example 2:* Anti-vaccine misinformation during the COVID-19 pandemic caused public health risks but did not always violate laws, leading to slower moderation responses.

4.2. Reactive Nature, Not Preventive:

a) While the Act includes some proactive duties, its overall emphasis leans more toward reactive enforcement. It aims to regulate harm that has already occurred or been identified, relying on penalties and post-incident reviews to ensure compliance. Platforms may take hours or even days to act on reported content, particularly during crises or high-volume events. For example, during the 2024 UK riots, false narratives circulated widely before being flagged and moderated, contributing to unrest in the interim.

b) To truly become proactive, the legislation would need to emphasize real-time monitoring, early detection systems, and adaptive regulatory mechanisms to address the fast-evolving nature of online harm.

4.3. Burden on Users:

a) The system shifts a significant portion of the responsibility to users, requiring them to recognize, report, and articulate their concerns about harmful content. This can be particularly problematic for vulnerable individuals who may not feel empowered or equipped to engage with reporting mechanisms.

b) Reliance on user reports often leads to inconsistent enforcement. Content flagged by one community may be overlooked in another, depending on cultural sensitivities, platform biases, or the volume of complaints received.

c) In the vast digital ecosystem, relying on user reports becomes impractical, especially for large platforms that host millions of pieces of content daily. Automated moderation systems may lack the contextual understanding needed to complement this reactive approach effectively.

III. What More Should Be Done to Combat Potentially Harmful Social Media and AI Content?

To tackle the pervasive spread of misinformation and harmful content, a multifaceted approach is essential. By focusing on algorithmic transparency, AI literacy, collaborative governance, and real-

[12] https://www.bbc.co.uk/news/technology-63782082

[13] Sullivan-Tibbs, M. A. (2024). *A Systematic Review: Mental Health of Minority Communities' Exposure to Negative Media* (Doctoral dissertation, California Southern University).

time moderation tools, we can address these challenges effectively. This framework integrates innovative technologies, education, and partnerships to foster a safer, more informed digital landscape.

5. The terms misinformation and disinformation[14] are often used interchangeably but have distinct meanings, especially in discussions about the spread of harmful content online:

5.1. **Misinformation** refers to false or inaccurate information that is spread without the intent to deceive. For example, during the 2024 UK riots, some users may have unknowingly shared incorrect claims about the Southport incident.

5.2. **Disinformation** is deliberately false information created with the intention to mislead or cause harm. For instance, generative AI tools could have been used to fabricate malicious narratives to incite unrest during the riots.[15, 16]

5.3. In the case of the 2024 UK riots, both misinformation (e.g., unintentionally shared false claims) and disinformation (e.g., targeted, AI-generated divisive content) likely played a role in fuelling societal tensions. Algorithmic amplification by social media platforms further exacerbated the reach and impact of both types of content.[17]

6. **Algorithmic Transparency and Audits:** Social media platforms must disclose how their algorithms rank content and undergo independent audits to assess their impact on amplifying harmful narratives.

6.1. **Case Study: The 2024 Southport Incident:** During the summer riots in 2024, misinformation surrounding the Southport stabbing reached over 15.7 million users. Erroneous claims that an asylum seeker was responsible fuelled nationwide unrest. While no specific social media platform was explicitly highlighted, platforms with significant user bases and prior criticisms, demonstrate the potential role of algorithms in amplifying divisive narratives and contributing to misinformation dissemination. The Sky documentary *"Doom Scroll: Andrew Tate and The Dark Side of The Internet"* revealed how YouTube's algorithms directed young users toward misogynistic content[18], demonstrating how recommendation systems can inadvertently amplify harmful material[19]. Recent research also reports that the algorithmic processes on TikTok and other social media sites target people's vulnerabilities – such as loneliness or feelings of loss of control – and gamify harmful content[20]

[14] Kumar, S., Kumar, A., Mallik, A., & Singh, R. R. (2023). Optnet-fake: Fake news detection in socio-cyber platforms using grasshopper optimization and deep neural network. *IEEE Transactions on Computational Social Systems*.

[15] Carpenter, P. (2024). *FAIK: A Practical Guide to Living in a World of Deepfakes, Disinformation, and AI-Generated Deceptions*. John Wiley & Sons.

[16] https://www.theguardian.com/politics/article/2024/aug/02/how-tiktok-bots-and-ai-have-powered-a-resurgence-in-uk-far-right-violence

[17] https://www.oversightboard.com/news/de-nouveaux-cas-concernent-des-publications-partagees-en-soutien-aux-emeutes-au-royaume-uni/?lang=fr

[18] Peterson, L. (2023). Manufacturing Misogyny: How The YouTube Recommendation Algorithm Radicalizes Young Men.

[19] https://www.isdglobal.org/isd-in-the-news/isd-study-reveals-how-youtubes-algorithm-pushes-problematic-content-despite-user-interest/#:~:text=ISD%20study%20reveals%20how%20YouTube's%20algorithm%20pushes,content%20despite%20regardless%20of%20interests%20or%20age.

[20] https://www.ucl.ac.uk/news/2024/feb/social-media-algorithms-amplify-misogynistic-content-teens#:~:text=Social%20media%20algorithms%20amplify%20extreme%20content%2C%20such,new%20report%20led%20by%20a%20UCL%20researcher.

6.2. These examples underscore the need for algorithmic transparency to prevent social tensions from being exacerbated. Independent audits can ensure algorithms are not prioritizing harmful content over credible information.

7. **AI Literacy Initiatives:** Governments should also spearhead initiatives to enhance AI and media literacy, as public education campaigns, empowering users to discern credible information and navigate misinformation critically.

7.1. Finland's Media Literacy Curriculum: Finland's systematic media literacy[21] initiative emphasizes high-quality education across all age groups to develop critical media evaluation skills. This approach has successfully reduced misinformation's impact by fostering an informed population.

7.2. Artificial Intelligence Literacy in the United States: The *Artificial Intelligence Literacy Act of 2023 (H.R. 6791)*[22] formalized AI literacy as a vital part of digital literacy, equipping citizens with the knowledge to understand AI's principles, applications, and ethical considerations. By integrating AI literacy into existing programs, the U.S. aims to prepare its population for an AI-driven future while maintaining technological leadership.

7.3. These initiatives highlight the importance of media[23] and AI literacy[24] in enabling users to navigate the complexities of modern information ecosystems with discernment and integrity.

8. **Collaborative Governance:** Developing partnerships among governments, tech companies, and academics is critical to combating misinformation through shared expertise and resources.

8.1. Co-Designing Safer Algorithms: Platforms should collaborate with independent researchers to create algorithms that minimize bias and prevent the amplification of harmful content.

8.2. Example: Collaborative filtering mechanisms, where communities tag and verify misinformation, can be enhanced by AI systems that offer initial assessments, streamlining the validation process.

9. **Proactive AI and Real-Time Detection Tools**

Advanced AI systems are indispensable for real-time detection and mitigation of harmful content. Platforms should adopt the following strategies:

9.1. Influential Node Detection: Identifying influential nodes within social networks can significantly enhance the ability to counteract misinformation at its source. Platforms can utilize models such as Community Structure with Integrated Feature Ranking[25] to pinpoint users or accounts that play a pivotal role in spreading rumours. By targeting these nodes, platforms can disrupt the amplification of harmful narratives and minimize their reach.

[21] https://medialukutaitosuomessa.fi/mediaeducationpolicy.pdf

[22] https://www.congress.gov/bill/118th-congress/house-bill/6791/text

[23] Kumar, A. 2024. Consultation Response: Ofcom's three-year media literacy strategy. https://www.ofcom.org.uk/media-use-and-attitudes/media-literacy/ofcoms-three-year-media-literacy-strategy/

[24] Kumar, A., & Sangwan, S. R. (2024). Conceptualizing AI Literacy: Educational and Policy Initiatives for a Future-Ready Society. *International Journal of All Research Education & Scientific Methods*, *12*(4), 1543-1551.

[25] Kumar, S., Kumar, A. *, Panda B.S. (2022) Identifying Influential Nodes for Smart Enterprises using Community structure with Integrated Feature Ranking *IEEE Transactions on Industrial Informatics*, https://doi.org/10.1109/TII.2022.3203059

9.2. **SIRA (Spreading Immunization and Rumour Control Model)**: The **SIRA model**[26] offers an epidemic-inspired framework for controlling the propagation of misinformation. By simulating the spread of rumours and incorporating immunization techniques, this approach can help platforms proactively mitigate the impact of misinformation during crises, such as public health emergencies or political unrest.

9.3. **Real-Time Moderation with LLMs**[27]: Large Language Models (LLMs) such as GPT, Llama-3, and Gemma-1.1 can analyze content in real-time to detect harmful narratives, conspiracy theories, and fake news.

a) *Automated Fact-Checking*: LLMs compare claims against reliable datasets, flagging inaccuracies early.

b) *Contextual Understanding:* Unlike rule-based systems, LLMs identify nuanced harmful content, such as xenophobic undertones in subtly worded posts.

c) *Scalability:* LLMs can moderate vast amounts of content, detecting patterns faster than human moderators.

d) *Integration with Tools:* Combined with systems like Sensity AI[28], LLMs can analyze text accompanying deepfake videos to detect fabrications.

9.4. **Regular Audits for Ethical Compliance**: To ensure these tools remain effective and unbiased, regular audits are necessary. Ethical guidelines should focus on transparency, fairness, and privacy protection to avoid unintended harm or discriminatory practices.

10. Envisioning a Future Solution: Envisioning a future solution to harmful content, LLMs are seen not merely as **L**arge **L**anguage **M**odels but through two transformative and forward-looking perspectives that unlock new possibilities[29].

10.1. **L**ies, **L**ogic, and **M**edia – Deciphering how misinformation spreads, how logic dismantles it, and how media becomes the frontline in the battle for truth.

10.2. **L**ogical **L**iteracy in **M**edia – Redefining LLMs as tools that empower critical thinking and enable individuals and organizations to navigate AI-generated content with discernment and integrity.

10.3. Together, these perspectives demonstrate how Generative AI can shape a digital ecosystem where trust triumphs over fabrication, and technology serves humanity with ethical integrity. By integrating transparency, education, governance, and real-time tools, we can pioneer a future where AI amplifies truth and empowers society.

10.4. Key findings from the UK and European elections report that Generative AI played less of a role in boosting the virality of disinformation compared to traditional interference methods and human influencers[30].

[26] Kumar, A., Kumar, S.*, Aggarwal, N. (2022). SIRA: A Model for Propagation and Rumor Control with Epidemic Spreading and Immunization for Healthcare 5.0, *Soft Computing*, *A Fusion of Foundations, Methodologies and Applications*, Springer, https://doi.org/10.1007/s00500-022-07397-x

[27] Kumar, A. (2024). Generative AI and Information Fabrication: NLP Techniques for Truth and Trust. https://research.gold.ac.uk/id/eprint/37980/2/Generative%20AI%20and%20Information%20Fabrication.pdf

[28] https://sensity.ai/

[29] Kumar, A. (2024). Generative AI and Information Fabrication: NLP Techniques for Truth and Trust. https://research.gold.ac.uk/id/eprint/37980/2/Generative%20AI%20and%20Information%20Fabrication.pdf

[30] https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-analysis-2024-uk-and-european-elections

10.5. Generative AI can therefore transform the fight against harmful content, shifting from a reactive stance to a proactive one.

11. A smart solution involves implementing AI-driven, platform-specific moderation systems that adapt to each platform's unique features:

11.1. **Real-Time Monitoring for X:** Deploy AI tools to detect and flag viral misinformation through trending topics and rapid content analysis.

11.2. **Video Content Verification for TikTok:** Use AI-powered visual and audio analysis, combined with fact-checking overlays, to address misleading video trends.

11.3. **Community-Driven Detection for WhatsApp:** Integrate encrypted, privacy-preserving tools that enable users to flag misinformation within groups, paired with media literacy prompts.

11.4. **Content Verification and Algorithmic Adjustments for YouTube:** Implement AI-powered analysis of video content, titles, and descriptions, with fact-checking overlays and adjusted recommendation algorithms to limit the spread of misleading information.

12. This multi-layered, platform-specific approach ensures proactive intervention while respecting platform dynamics and user behaviours.

IV. What Role Do Ofcom and the National Security Online Information Team Play in Preventing the Spread of Harmful and False Content Online?

13. **Ofcom's Role:**

13.1. **Regulation and Enforcement:** Oversees compliance with the Online Safety Act, monitoring platforms for harmful content and imposing penalties.

13.2. **Transparency and Guidance:** Publishes periodic reports and guidance to ensure platforms adhere to safe practices.

13.3. **Algorithmic Oversight:** Partners with academia and tech experts to better understand and regulate algorithms.

14. **NSOIT's Role:**

14.1. **Real-Time Threat Monitoring:** Analyzes online spaces for misinformation that poses national security threats or risks public safety.

14.2. **Strategic Interventions:** Works with law enforcement and intelligence agencies to neutralize misinformation campaigns during crises like the summer riots.

V. Which Bodies Should Be Held Accountable for the Spread of Misinformation, Disinformation, and Harmful Content?

15. Accountability for the spread of harmful content must be distributed across multiple stakeholders, each of whom plays a critical role in influencing the digital information ecosystem.

16. **Social Media Platforms:** Platforms must implement robust moderation practices, algorithmic transparency, and independent audits to minimize the spread of harmful narratives.

17. **Search Engines**

17.1. Search engines influence content discovery and must ensure that their algorithms do not amplify disinformation or prioritize unreliable sources.

17.2. Search results should favour credible information, with clear guidelines on ranking systems to reduce the visibility of harmful or misleading content.

17.3. Transparency around how content is indexed and prioritized is critical for accountability.

18. **Regulatory Bodies**

18.1. Ofcom, as the UK's designated regulator under the Online Safety Act, is responsible for enforcing compliance and ensuring consistent oversight of digital platforms.

18.2. Ofcom must be adequately resourced and empowered to monitor and address harmful content in real time.

18.3. The body should establish clear mechanisms for reviewing platforms' compliance and issuing penalties for non-conformance.

19. **Content Creators**

19.1. Individuals or organizations producing harmful content must face legal consequences for their actions.

19.2. This includes those intentionally generating disinformation, such as conspiracy theories or malicious deepfakes, with the aim of inciting harm or unrest.

19.3. Holding creators accountable ensures personal responsibility and deters the production of harmful material.

20. To effectively combat misinformation and harmful content, a multi-faceted approach integrating real-time AI tools, platform-specific strategies, and collaborative governance is essential. By prioritizing algorithmic transparency, AI literacy, and proactive regulation, we can build a safer, more resilient digital ecosystem.

18 December 2024

Credentials and the foundation for the evidence submission concerning Social Media, Misinformation, and Harmful Algorithms

Greetings,
I am Dr. Akshi Kumar, a Senior Lecturer in Computer Science at Goldsmiths, University of London, specializing in Natural Language Processing (NLP), misinformation dynamics, and AI ethics. Over the years, my research has focused on advancing interdisciplinary approaches to understanding and mitigating the impacts of digital misinformation, content propagation, and the ethical adoption of AI in critical domains.

With a strong background in the field, I have had the opportunity to contribute insights and evidence to several high-profile inquiries, providing evidence to the UK Parliament and other organizations on matters central to AI, misinformation, and cyber resilience. My **previous written evidence submissions** include:

1. **UK Parliament Written Evidence (AIG0003):**
   - *“Balancing Act: Risks and Opportunities in AI Adoption within UK Government Services”*
   - Submitted to the **House of Commons Public Accounts Committee** inquiry on AI in Government (2024). Read here.
2. **UK Parliament Written Evidence (FON0002):**
   - *“Media in Transition: Assessing the Impact of Technology and AI on News Integrity and Trust”*
   - Submitted to the **House of Lords Communications and Digital Committee** on the future of news and technology (2024). Read here.
3. **UK Parliament Written Evidence (CYB0001):**
   - *“Guarding the UK's Critical Infrastructure: The Rumour Challenge in Cyber Resilience”*
   - Submitted to the **House of Commons Science, Innovation, and Technology Select Committee** inquiry on cyber resilience of the UK’s critical national infrastructure (2023). Read here.

These submissions reflect my expertise in the interplay between AI, societal trust, and national resilience. In addition to these contributions, my active role as a **member of the Steering Group for the Mayor of London’s Violence Reduction Unit** enables me to address online harms through collaborative research and policy-making initiatives. I have also **presented keynotes** such as *“Generative AI and Information Fabrication: NLP Techniques for Truth and Trust”* and contributed to shaping media literacy strategies in consultation with Ofcom.

Related Publications and Impact

- *Generative AI and Information Fabrication: NLP Techniques for Truth and Trust* (2024). Read here.

- *Consultation Response: Ofcom's Three-Year Media Literacy Strategy* (2024). Read here.
- *Conceptualizing AI Literacy: Educational and Policy Initiatives for a Future-Ready Society* (2024). International Journal of All Research Education & Scientific Methods. https://doi.org/10.56025/IJARESM.2023.1201241543
- *OptNet-Fake: Fake News Detection in Socio-cyber platforms using Grasshopper Optimization and Deep Neural Network* (2023) IEEE Transactions on Computational Social Systems- https://doi.org/ 10.1109/TCSS.2023.3246479
- *SIRA: A Model for Propagation and Rumor Control with Epidemic Spreading and Immunization* (2022). Soft Computing, Springer, https://doi.org/10.1007/s00500-022-07397-x
- *Identifying Influential Nodes for Smart Enterprises using Community structure with Integrated Feature Ranking* (2022) IEEE Transactions on Industrial Informatics, https://doi.org/10.1109/TII.2022.3203059

Driving Force Behind My Submissions

My evidence-based approach is driven by:

a) **Evolving AI Ecosystem:** The rapid adoption of AI brings both transformative opportunities and ethical challenges, particularly in governance, media, and infrastructure resilience.

b) **Interdisciplinary Collaboration:** Bridging gaps between technical innovation, public trust, and policy frameworks through collective expertise.

My overarching aim is to harness advanced AI methods, foster transparent regulatory strategies, and enable ethical technological development to build a safer, more resilient society.

Warm regards,
Akshi

Endorsed by *Royal Academy of Engineering*, UK: Exceptional Talent in the field of AI/Data Science, 2022
"Top 2% highly cited scientist of the world", Stanford University List, USA (2021-2024)
Senior Member, IEEE