CHASE INDIA

AI and Content Moderation

An Analysis of their Interplay in the Era of Misinformation and Deepfake

January 2024

Executive Summary

The growth of smart phones and fast internet infrastructure in India has led to the rise of on-demand content which has brought a paradigm shift in the way online content is consumed. Online platforms including social media, e-market place, over the top (OTT) platforms, among others, have enabled consumers to access and create content from anywhere and at any time. As large amount of data pertaining to online content is created every second, impacting the population at large in a significant way, there has been growing concerns on the creation and spread of misinformation/ disinformation, fake news, etc. as they pose serious threat to the physical and psychological health of the population.

Growing reliance on online platforms for accessing and sharing of information have made socio-economic and political spheres vulnerable to misinformation thus creating new challenges in terms of maintaining public order, consumer behavior manipulation, influencing political ideology, etc. To tackle this problem of misinformation and fake news, social media organizations and technology companies have been developing and implementing various types of content moderation practices and with the advancement of technologies like Artificial Intelligence, such technologies are increasingly being deployed in content moderation activities.

The ability of artificial intelligence to sift through enormous amounts of data with great speed for fact checking and content filtration has made this technology one of the prominent tools to tackle misinformation. However, the extent to which AI tools can be used for content moderation, independent of human content moderators, is a question that needs to be carefully studied and examined because information in online content needs to be checked with nuanced understanding and different perspectives, which emerging technologies like AI may lack in compared to human moderators. Also, there has been a growing concern on rise of disinformation such as deepfakes that can be generated by emerging technologies like generative AI which can be disastrous, given the ability of these technologies to create new content almost like original content which can lead to the violation of intellectual property rights, privacy rights, etc.

Therefore, to realize the potential of AI technologies for content moderation practices, principles for responsible use of AI needs to be built, taking into consideration aspects like protection of human rights, building of trust on AI and accountability of AI applications.

Introduction

The rise of online platforms, ranging from social media to e-market places and web content providers referred to as over the top (OTT) platforms, has brought about a paradigm shift in the internet ecosystem in terms of consuming online content by users. It has enabled consumers to access content anytime and anywhere at the click of a button. This proliferation of digital platforms and ease of sharing information and content on them has also led to concerns about the growing misinformation and disinformation over the internet. These include activities like online harassment, hate speech, fake news, doxing and deep fakes.

The spread of misinformation through online platforms has become a primary concern not just for the platforms themselves but also governments, as it has the potential to manipulate consumer behaviour which can have far reaching implications for society, public order, government functioning and economies. This is especially so when disinformation, which is misinformation with the intention to deceive and inflict harm, is on the rise. Online platforms have been using several content management practices to tackle misinformation and disinformation. Among these are removal and prioritisation of content or suspension of user account in case user content/activities are in violation with the terms and condition of the platforms.[1] One such technological tool used by online platforms is Artificial Intelligence (AI), as AI systems have the capability to sift with great speed through enormous amounts of user-generated data and help in filtration of potentially harmful content.

In 2021, Government of India has released the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 which provide for due diligence to be followed by online social media intermediaries to ensure an open, safe, and trusted internet. The Information Technology Rules originally obligated Social Media Intermediaries (SMIs) to inform its users about the privacy policy, rules and regulations and user agreement that govern its platforms. These rules have put more obligations on intermediaries to deploy technology-based measures[2] to identify harmful information or content such as child sexual abuse material, content related to terrorism, etc. and also remove or disable access to such information with prior intimation to the user.

Though AI systems are considered as prominent technologies in fact-checking or identifying harmful content through analysing and classification of online content or information, a debate is on the technology and regulatory ecosystem as to what extent fact-checking or content moderation, whether by humans or assisted by algorithms, is agile and adaptable to the rapidly evolving misinformation ecosystem.[3] There are also growing concerns related to the potentiality of AI in generating misinformation by itself such as deepfakes.

This paper seeks to understand how AI systems can complement content moderation efforts of online platforms and identifies principle-based frameworks and models that can be adopted by different stakeholders across government and policy, industry, civil society and academia for ethical use of AI in order to tackle misinformation as well as address concerns related to AI generated disinformation.

Potential use of AI for Content Moderation

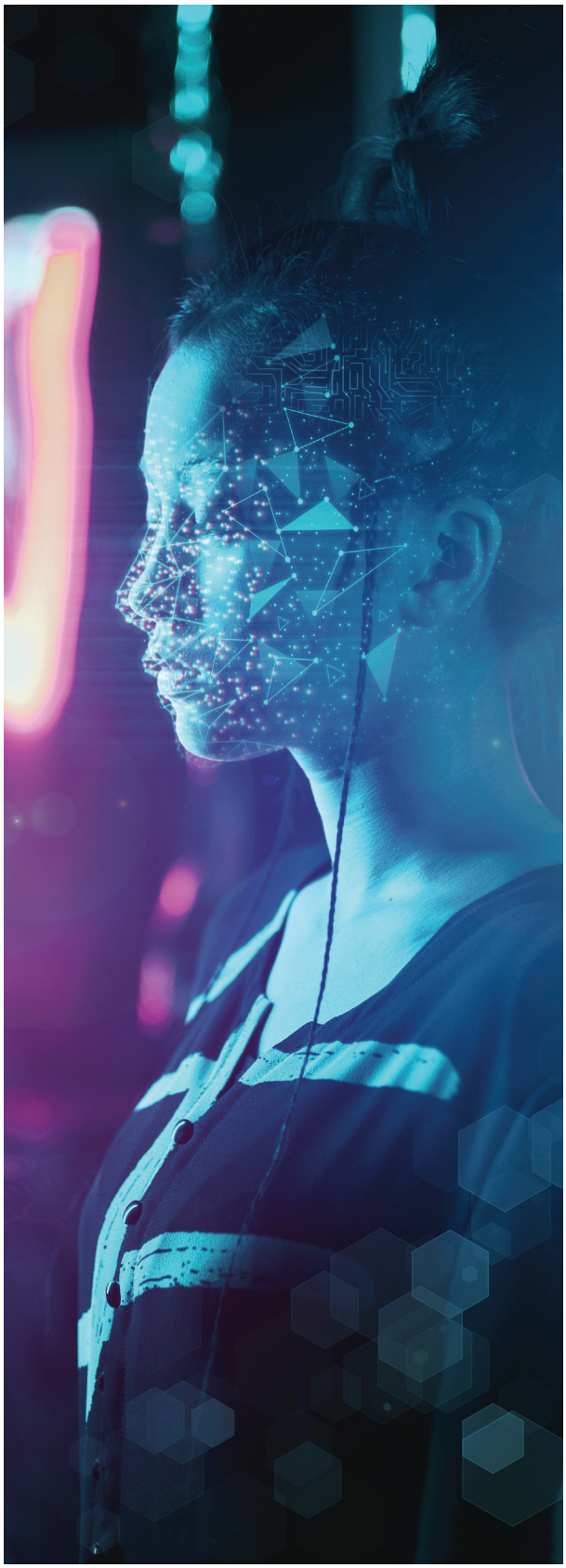
With easy access to information, enormous amounts of data are being created every second. According to the World Economic Forum estimations, humans will create about 463 exabytes[4] (one exabyte is equal to one billion gigabytes) of data every day which would create tremendous challenges for those engaged with content moderation processes to keep up with the pace. AI systems are capable of handling high volumes of data across multiple channels and in real time, including automatic analysis and classification of harmful content, thus increasing the speed and effectiveness of content moderation processes. Additionally, processing of data through AI systems has higher level of accuracy and precision than manual processing.

Consumers often consume user generated content via multiple mediums or sources. Therefore, content moderation is a multimodal challenge rather than a one-size-fits-all one. Due to diversity in content types generated online including texts, photographs and videos, natural language processing (NLP) can be trained to analyse and identify harmful content in different forms which would otherwise need different forms of expertise, had content moderation been done by humans. Technology companies also rely on third-party content moderation services deploying AI models for their high scalability and speed, customization ability and reduction in costs. For example, AI technology is central to the content review process of social media intermediaries like Facebook because AI can detect and remove content that goes against its Community Standards before anyone reports it.[5] AI technologies can also refer doubtful content to human content moderation teams, especially when such content requires further review. Therefore, several online platforms involved in content moderation processes use a combination of human and AI moderators to moderate content that violates their policies such as misinformation, fake news, violent content, hate speech and pornography. The use of AI for content moderation also lessens the impact of harmful content on human moderators who have to go through extensive amount of information in multiple formats, a task which can compromise their mental health[6].

AI-based content moderation processes are deployed through AI systems that are designed based on pre-defined parameters and algorithms to identify and flag harmful content online. These algorithms guide technological tools in decisions like taking down of harmful content or misinformation from online platforms. There are concerns that the data and algorithm used to train AI systems to identify and flag potential harmful content may be underrepresented or biased which might lead these tools to take down content which may be harmless. This is because a lot of information and content generated online are highly contextual and need nuanced understanding. If AI systems are not trained to spot such nuances like the intent behind a particular piece of information alleged to be misinformation or disinformation, it cannot filter harmful online content without affecting human rights such as free speech. Therefore, it is crucial to understand and reflect on the extent that AI systems or tools can be relied on for moderating online content or information. If automation and AI-based content moderation is considered the future, then there will be questions that need to be addressed such as the robustness and ability of AI systems to react to online content with empathy, rationality and emotional intelligence that human content moderators possess.

AI-Generated Disinformation

One of the pertinent questions in the context of content moderation is that while AI can be potential game-changer for content moderation through identification and flagging of harmful content, what about disinformation that can be created by AI itself? This concern is growing because generative models of AI such as Generative Adversarial Networks (GAN) can produce original content, which may benefit different sectors of the economy like health care, education and law, but these models can fall into the hands of propagandists who can create information and content designed to shape perceptions of people in a particular direction which can be misleading and harmful.[7] Additionally, deceptive and manipulated content such as deepfake images and videos, voice cloning, generative texts, generated by AI may become indistinguishable from the original or non-manipulated content[8]. So, it is critical to analyse the potential threats that AI-enabled operations can create and outline steps to identify such risks.



To address the concerns related to the generation of disinformation through AI, technology organisations are taking measures through development of algorithms and technology models that can identify manipulative and fake content created by AI technologies, especially Generative AI, such as Generative Adversarial Networks (GANs), Natural Language Processing (NLP), etc. Several techniques have been adopted by GANs to address the issue of disinformation such as Detection, Provenance, Regulatory initiatives, open-source intelligence techniques[9]. NLP models have been used to develop techniques such as Tokenisation, Part-of-Speech (POS) tagging, Sentiment Analysis, Dependency Parsing, etc.[10].

In June 2023, the European Commission approved the draft proposal for the Artificial Intelligence Act, which provides for limited set of transparency obligation for AI systems that generate or manipulate image, video or audio content.[11] In June 2022, the European Commission released the Strengthened Code of Practice on Disinformation 2022[12], based on the work carried out

by the signatories including major online platforms, players from the advertising industry, fact-checkers, research and civil society organisations who have committed to take into consideration the transparency obligations and the list of manipulative practices prohibited under the draft Artificial Intelligence Act.

To create a defence against large-scale, automated disinformation attacks in the US, the Defense Advanced Research Project Agency (DARPA) has developed prototypes and technologies to identify and combat manipulated images or videos, including deepfake defensive models[13]. In 2019, Texas passed a law, making the distribution of deepfake videos intended to influence the result of an election illegal.[14]

The Deepfake Report Act, 2019, requires US Secretary of Homeland Security to publish an annual report on the extent deepfake technologies are being used to weaken national security, undermine nation's elections, and manipulate media.[15] Technology companies like Meta, Google and Microsoft are also using different models to address disinformation created by AI. Microsoft has created a Video Authenticator using a public dataset, which was tested on DeepFake Detection Challenge Dataset, for training and testing deepfake detection technologies.[16] Meta has also collaborated with industry leaders and academic experts to create an open, collaborative initiative- 'Deepfake Detection Challenge'[17], to spur the creation of innovative new technologies for detection of deepfakes and manipulated media.

Recommendations

AI-enabled content moderation tools are proving to be quite effective in identification, flagging and taking decisions for take down of harmful online content. With the growing use and deployment of automated tools and algorithms in content moderation processes, utmost importance should be given to building trust among stakeholders, including users, government, regulators, among others, on AI-based content moderation standards and performance of AI systems, ensuring that decision making by such automation tools, identification and actions taken on harmful content are unbiased and non-discriminatory. At the same time, urgent priority needs to be given on understanding the scope and extent of misuse of AI tools and AI's ability to generate disinformation through collaboration among different stakeholders.

To strengthen information integrity by online platforms and technology companies deploying AI for content moderation practices, we have come up with the following recommendations:

Ensuring user trust through transparency

- To build a trusted content moderation system, both AI systems and human moderators should be deployed in the moderation process. AI systems can help in filtering and taking down of harmful content at a scale and speed whereas human moderators can be deployed to understand the nuances of the content such as whether intent of the user sharing such information or content is in accordance with the law.
- Additionally, AI systems used for content moderation processes need to be designed in such a way that it can provide analytical data on the decisions taken by it for filtering harmful content so that the objectivity and bias (if any) of such technological systems can be analysed.
- As data and algorithms used to train AI and GANs evolve with different user behaviour taken into consideration, it is necessary that periodic audits and calibration of data sets are carried out and testing regimes for AI and GANs are established, so that data fed into such systems for training are well representative of the population.

Developing detection and take down policies for AI-generated disinformation

- Detection systems and takedown policies should be developed to keep AI-generated disinformation, such as deepfakes, at bay. Such detection systems will require rapid access to content or information samples produced online.
- Detection and takedown policies can also provide for the creation of "deepfake zoo" by technology companies and social media platforms, that can continuously aggregate freely available datasets of AI-generated disinformation, as they appear online.
- Such policies can also provide for encouraging 'radioactive' marking of public datasets that are used for training AI systems so that quick detection of AI-generated disinformation circulating online can be done.
- There should be a mechanism available to those impacted by AI or automated takedown systems, including transparent reasoning and context behined such takedown actions.

Labelling of AI-generated content

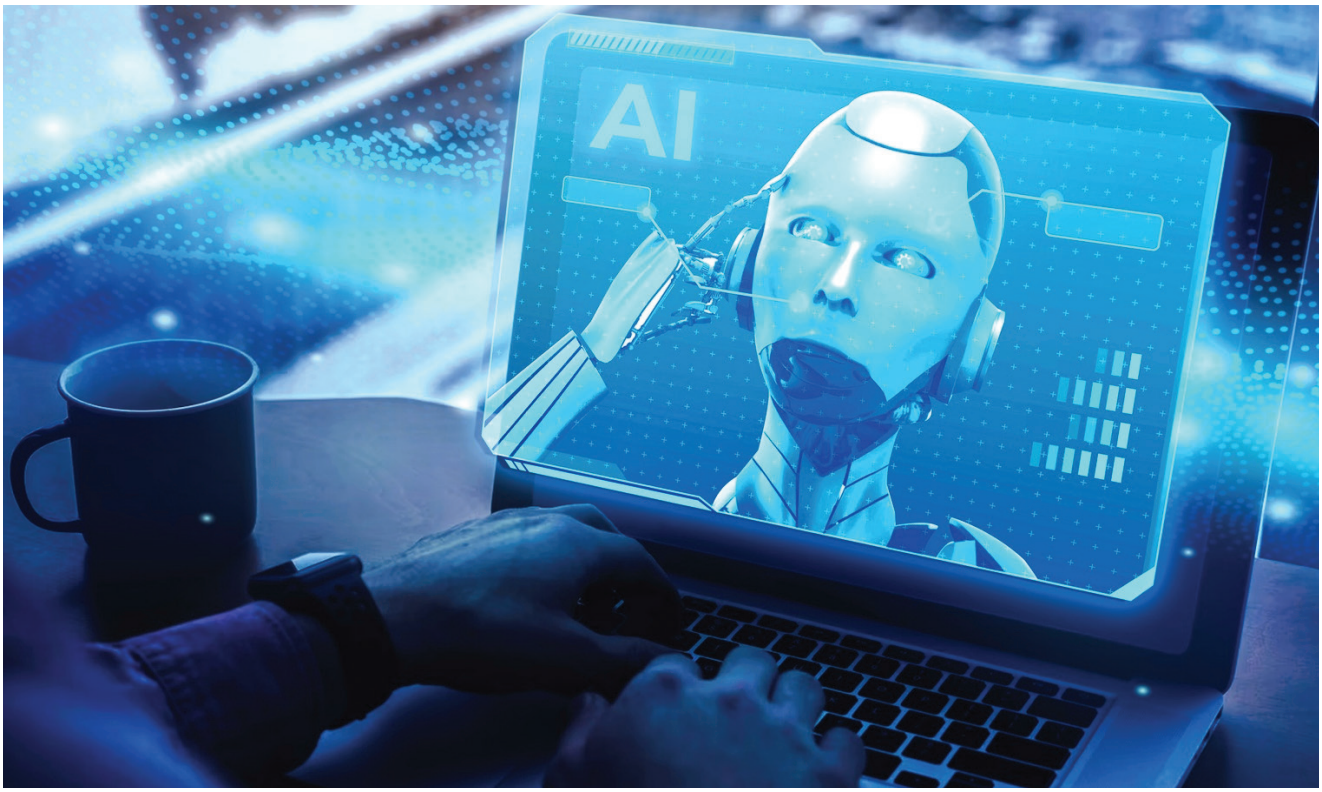
- In order to check and ensure that information or content generated by AI technologies are not biased or do not have inaccurate data which could potentially lead to harmful outcomes, all AI-generated content should be identifiable as AI generated content to the average person.
- Policies and laws that are formulated for regulating AI application such as actions taken by AI for moderating content or generation of content by AI such as GANs, should provide for obligations on the developers to design their systems in such a way that AI-generated content is identified as such. This should create a distinction between those content created by AI by itself and those created by humans through AI technologies.

Periodic assessment and classification of AI

- To ensure that AI systems used for content moderation are effective to identify, filter and take down harmful online content without any discrimination against fundamental rights like Freedom of Speech and Expression, end-to-end risk-based assessment and classification of these systems should be done. Some of the parameters that can be adopted for such risk-based assessment can include technical robustness and safety, transparency, diversity, non-discrimination and fairness and accountability.

Promoting investment in research and innovation

- To tackle growing online disinformation, there should be collaboration between government, academia and industry for development of AI models that can not only tackle the issue of disinformation but also create a trusted AI ecosystem by keeping a check on AI generated disinformation.
- As part of such collaborations, dedicated sections or divisions on AI R&D related to tackling disinformation can be created within existing Centre of Excellences (CoEs) for AI or new CoEs can be established.



Bibliography

1. Pratap, N., & Sahu, S. (2022). Content moderation through co-regulation. The Hindu. Accessed from: https://www.thehindu.com/opinion/op-ed/content-moderation-through-co-regulation/article66111242.ece

2. Rule 4 (4) of the IT Act, 2021. Accessed from https://mib.gov.in/sites/default/files/IT%28Intermediary%20Guidelines%20and%20Digital%20Media%20Ethics%20Code%29%20Rules%2C%202021%20English.pdf

3. Zhou, J., Zhang, Y., Luo, Q., Parker A.G., & Choudhury, M. D. (2023). Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA 20 Pages. Accessed from: https://dl.acm.org/doi/fullHtml/10.1145/3544548.3581318

4. Desjardins, J. (2019). How much data is generated each day? World Economic Forum. Accessed from: https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

5. How does Facebook use artificial intelligence to moderate content? Facebook Help Centre. Accessed from: https://www.facebook.com/help/1584908458516247

6. Is AI Content Moderation a Boon for Businesses Looking to Assure Quality? Expert Callers. Accessed from: https://www.expertcallers.com/blog/ai-content-moderation-good-or-bad/#:~:text=The%20use%20of%20AI%20for,or%20spammy%20content%20and%20more

7. Goldstein, J.A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova K, (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. Accessed from: https://arxiv.org/pdf/2301.04246.pdf

8. C2PA Releases Specification of World's First Industry Standard for Content Provenance. The Coalition for Content Provenance and Authenticity (C2PA). Accessed from: https://www.prnewswire.com/news-releases/c2pa-releases-specification-of-worlds-first-industry-standard-for-content-provenance-301468394.html

9. Helmus, T.C. (2022). Artificial Intelligence, Deepfakes, and Disinformation. Rand Corporation. Accessed from: https://www.rand.org/content/dam/rand/pubs/perspectives/PEA1000/PEA1043-1/RAND_PEA1043-1.pdf

10. Leveraging NLP Techniques for Effective Content Moderation. Accessed from: https://www.lettria.com/blogpost/nlp-techniques-for-content-moderation#:~:text=It%20uses%20machine%20learning%20algorithms,inappropriate%20material%20at%20massive%20scale.

11. Artificial Intelligence Act: Briefing. European Parliamentary Research Service. Accessed from: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

12. 2022 Strengthened Code of Practice on Disinformation. European Commission. Accessed from: https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

13. Sybert, S. (2021). DARPA Launches New Programs to Detect Falsified Media. GovCIO Media & Research. Accessed from: https://governmentciomedia.com/darpa-launches-new-programs-detect-falsified-media

14. Accessed from: https://capitol.texas.gov/tlodocs/86R/billtext/html/SB00751S.htm

15. S.2065 - Deepfake Report Act of 2019. Accessed from: https://www.congress.gov/bill/116th-congress/senate-bill/2065/text

16. Burt, T. (2020). New Steps to Combat Disinformation. Microsoft. Accessed from: https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/

17. Deepfake Detection Challenge Dataset. Meta AI. Accessed from: https://ai.meta.com/datasets/dfdc/

ABOUT CHASE INDIA

Founded in 2011, Chase India is a leading public policy research and advisory firm with growing practices in Technology & Fintech, Transport & Infrastructure, Healthcare & Life Sciences, Development and Sustainability. We provide consultancy services to organizations for mitigating business risks through insight-based policy advocacy. Over the years, Chase India has collaboratively worked with multiple stakeholders such as government, parliamentarians, civil society organizations, academia and corporates on several policy issues of critical importance. Chase India is committed to using its knowledge, high-ethical standards and result-oriented approach to drive positive action for our partners. Chase India has pan India presence with offices in New Delhi, Mumbai, Pune, Hyderabad, Chennai and Bengaluru and is a part of the WE Communications Group worldwide.
For more information, please visit www.chase-india.com.

ABOUT AUTHORS

Mrinmoy Deori Borah
Senior Associate
mrinmoyb@chase-india.com

Dhawal Gupta
Group Business Director
dhawalg@chase-india.com

Kaushal Mahan
Vice President
kaushal@chase-india.com

SUGGESTED CITATION

Chase India, 2024. AI and Content Moderation: An Analysis of their Interplay in the Era of Misinformation and Deepfake

CHASE
INDIA

DISCLAIMER

Neither Chase Avian Communications Private Limited (referred to as "Chase India"), nor agency thereof, nor any of their employees, nor any of their contractors, subcontractors or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or any third party's use or the results of such use of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific organization, commercial product, process or service by trade name, organizer trademark, manufacturer or otherwise does not necessarily constitute or imply its endorsement, recommendation or favoring by the organizer or any agency thereof or its contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of Chase India or, or any agency thereof.

For more information, please visit www.chase-india.com