*For official use only*

No.40/AS (NEST)/2022
Ministry of External Affairs
AS (NEST)'s Office

I am enclosing below for perusal a paper on "*AI, Information Security & Ethics*" prepared in NEST Division, MEA under my supervision.

2. The AI Industry is set to see exponential growth at a pace which would be unprecedented. The role and applications of AI in development, especially in enhancing the sophistication of existing processes is a positive manifestation, which is helping build performance parameters to assist human capacities. The realisation of AI that is somewhat sentient is, however, still in the future.

3. Use of AI for advancing and mitigating the effects of disinformation will soon have implications for information security, synthetic media, use of bots, the problem of deep fakes, the issues of misuse of synthetic images and synthetic audio and speech and even texts, etc.

4. In the context of AI, we are already seeing issues in Machine Learning, Deep Learning, etc., highlighting the challenge of adequate explainability to make AI models comparative to human understanding and perception, not towards sentient AI which is still distant yet, but in the more immediate term where innovation and use of AI is reigniting the debate around issues of fairness, reliability and security.

5. All of this points to the need for developing norms and standards, rules, practices, processes and technological tools for AI governance; a governance structure that internalizes the elements of empathy, responsibility and accountability – the core of constructive human action - into AI.

6. The challenge, therefore, is to align AI to the best of human ambition. In this context, the enclosed NEST paper flags seven fundamental pillars or elements that are required to build robust AI governance models.

14/7/2022
D(KB)

Pl. examine & suggest
13/07/22
AS(RK)

Office of (KB) Scientist 'F' Diary No. 674 Dated 14/07/22

Office of Secretary, MeitY Dy. No. 1887735 Date 13/07/2022

Ministry of Electronics & Information Technology Diary No. 1887735 Date 14/7/22 Office of Addl. Secy

7. We look forward to your comments, inputs and guidance on building AI governance capacities in India, which would then also be shared with our Missions overseas, as relevant, to forge collaborative linkages in the development of AI.

(Renu Pall)
Additional Secretary
for New Emerging & Strategic Technologies
08.07.2022

*Separate copies to:*

1. Dr. P.K. Mishra, Principal Secretary to Prime Minister of India.
2. Dr. V.K. Saraswat, Member, NITI Aayog.
3. Prof. Ajay Kumar Sood, Principal Scientific Advisor to Government of India.
4. Dr. G. Satheesh Reddy, Secretary, Department of Defence R&D and Chairman, DRDO.
5. Dr. S. Somanath, Secretary, Department of Space & Chairman Space Commission.
6. Dr. Rajesh S. Gokhale, Secretary DSIR & DG CSIR.
7. Dr. Srivari Chandrasekhar, Secretary, Department of Science & Technology.
8. Shri K. Rajaraman, Chairman DCC and Secretary (T), Department of Telecommunications, Ministry of Communications.
9. Shri Alkesh Kumar Sharma, Secretary, Ministry of Electronics & Information Technology.
10. Shri Hari Ranjan Rao, Additional Secretary (Technology & Governance),
    PMO, New Delhi.

Ministry of External Affairs
NEST Division

# ***AI, Information Security and Ethics***

## *Introduction*

**1.** We have seen some staggering advances in AI applications. In just over a year, there have been starling AI enabled innovations in biotechnology and medicine, self-driving systems, surveillance, natural language processing to name a few. AI enabled systems have started to meet or sometimes exceed human capabilities in several areas. However, over the last two decades, AI enabled technological innovations especially in the digital realm are affecting the notions of power and influence. For instance, with digital technologies introducing multitude of communication channels resulting in unprecedented connectivity, information flow has moved to an unregulated, disaggregated, individual generated content. As an unintended consequence, trolling, disinformation campaigns, conspiracy theories have been thriving under such unregulated information flows. In another example, there is a growing response of nation states towards capitalization of emerging value of data while defending informational sovereignty. It is a reaction to the actions of the BigTech firms vis-à-vis exploitation of data through surveillance capitalism. Further, through network effects, these platforms accumulate user bases of billions and dominate global website traffic which has led the BigTech firms wielding unprecedented power over citizens' information streams and potential to manipulate political opinions. These platforms are susceptible to manipulation of content through fake news, deep fakes videos, privacy threats, digital discrimination that have the capabilities to disrupt social harmony. This is reshaping the online world with troubling implications for borderless information transmission, international collaboration, and freedom of speech on a global scale.

**2.** An un-deliberated adoption of such applications without a robust governance framework can plunge the world into an AI dystopia.

***Information Security***

**3.** Information is defined as a processed, organized and structured form of data that enables decision making. Information Security involves developing protocols/ measures to protect any form of data or information from unlawful / unauthorized access, use, misuse, disclosure, deletion, modification and/or disruption.

**4.** In the digital age, where quintillion bytes of data is being generated everyday and with almost every facet of society intrinsically interconnected with digital systems, information security becomes ever more critical. The role of AI in advancing and mitigating the effects of disinformation has serious implications on the changing threat landscape of information security. One such example is of the case of artificially generated media. It is sometimes referred to as synthetic media. They are artificially produced manipulated content typically in the form of audio, video, text, images generated through AI algorithms. The resulting content appears authentic and is used as a tool for propaganda with an intention to mislead and misinform people. The current information ecosystem has low barrier to entry such as seen in the case of digital media platforms (Twitter, Facebook, Youtube). A combination of bots that amplify such content at scale; AI algorithms that are used to target specific user base and amplify existing biases; and a lack of a robust framework to detect and penalize creation and dissemination of manipulated content; enable scalable, efficient and widespread propaganda.[1]

**5.** Examples of Synthetic media includes: 1) Deep Fakes – uses deep learning algorithms where an individual in an image or a video is replaced with someone else with a potential to deceive or cause harm. It has found widespread uses in generating fake news, fake celebrity pornographic videos, hoaxes and financial fraud. Here is an example of a deepfake in a video where major world leaders are seen singing in an audiovisual. Even though such a video appears to be harmless and satirical in nature, such type of videos with a malicious intent can be generated to create widespread panic,

[1] https://www.cnas.org/publications/reports/artificial-intelligence-and-international-security

**6.** 2) Synthetic Images – are artificial production of images using algorithms. Synthetic images are being generated for the better part of the last two decades. Computer generated imagery is common place in the entertainment industry. One of the sub-domains of synthetic images is human image synthesis i.e. near realistic human image compositions. However, in early 2000s largely expert systems were used to create such imagery. In the last decade or so, application of deep neural nets in image synthesis has allowed systems to create human like images without a human in the loop. Here is an example of an image synthesis website called This Person Does Not Exists. Whenever the website is refreshed, the back end algorithm called generative adversarial network (GAN) renders portraits of fake people. Such applications can help perpetuate misinformation and deception in an expeditious manner.

**7.** 3) Synthetic Audio and Speech – is artificial production of sound by manipulating audio waveforms. Like in the case of image and video synthesis, generation of synthetic sounds using programmable and non-programmable means have existed for decades. With advances in AI, seamless synthetic human speech that sounds close to actual human speech has become a reality. Some of the advances are seen in in Google's Deep Mind's WaveNet, a deep generative model of raw audio waveforms, that can produce natural sounding speech.[2] Chinese search company Baidu's DeepVoice system that uses deep neural network to convert text to speech. Its first version could produce short sentences indistinguishable from a human speech. Its second version could mimic a voice with 30 minutes of data and learn hundreds of accents. Its final version is said to mimic 2500 voices with 30 minutes of data each.[3] Another example is of a solution called resemble.ai that allows one to clone their voice for creating their digital avatars.[4] Such solutions are helping companies create relatable digital assistants and social robots, and help in democratizing music creation. However, production of synthetic audio especially the ones that clone human speech pose a threat. For instance, a cyber-security firm, Symantec reported three cases where fake audio of CEO's were used to trick financial controllers to transfer cash.[5] As it stands

[2] https://deepmind.com/research/case-studies/wavenet
[3] https://www.theverge.com/2017/10/24/16526370/baidu-deepvoice-3-ai-text-to-speech-voice
[4] https://www.resemble.ai/
[5] https://www.bbc.com/news/technology-48908736

currently, there are no tools that allow one to distinguish between real and fake audio, the security threat of synthetic audio is high.

**8.** 4) Text Synthesis – or sometimes referred to as Natural Language Generation uses systems to transform data into natural language. For instance, development of GPT 3 (Generative Pre-trained Transformer 3 ), a seminal development in natural language processing where the algorithms can write essays, translate languages, answer questions, and write codes. In real world applications, GPT 3 can be deployed, for instance, in generation and display of stories without a human in the loop. As language models' operational capabilities are dependent on the data that is fed into it, it can perpetuate bias. For instance, GPT 3 is trained on a massive amount of data that includes diverse websites like BBC, New York Times, Reddit, Wikipedia. As the information included in all these websites is written by humans, it incorporates best and worst of human qualities. Thus, GPT 3 learns to write like humans with all the inherent biases. Some real time examples can include in the area of content generation (like a media story) GPT 3 can pick up negative word associations with respect to particular gender or race as such connotations are littered in our historical datasets. In the domain of automated grading of papers, GPT 3 can reward papers written by students of a certain ethnicity and geography while penalizing students whose second language is English which is largely correlated to race. Thus making racial/ethnic minorities less likely to graduate. In the domain of healthcare, chatbots are increasingly used to understand patient health and recommend treatment. In case of a GPT 3 enabled chatbox and in an attempt to detect mental illness, these chatboxes might misinterpret patient's symptoms and present them with incorrect/offensive/dismissive comments putting the life of the patient at risk.[6] Despite its astounding capabilities in creating automated content, it can pose a serious threat to human security. It also poses a serious threat to the job market especially for journalists, writers and coders. Here is a perfectly written blog by GPT 3 which seems indistinguishable from humans. Another example is of a startup called decode.co that allows a non-programmer to build apps and websites using GPT 3.[7]

[6] https://techcrunch.com/2020/08/07/here-are-a-few-ways-gpt-3-can-go-wrong/

[7] https://debuild.co/

**9.** It is imperative to note that the challenges to information security is not limited to creation of human like artificial content but also the ability of AI systems to recognize patterns, predict behaviour of users, target specific user base with specific content, and artificially amplify and promote content through bots. For instances, advances in NLP or Natural Language Processing specially in the sub-domain of sentimental analysis has allowed evolution of AI systems to be able to recognize, analyse, interpret and develop emotionally relevant content. Coupled with the analysis of users' interaction with such content, it allows AI systems to further recalibrate the content for maximum impact. Thus, in effect creating a seamless propaganda tool.

***AI Governance***

**10.** While utilizing new age AI enabled technologies, we should not lose sight of notions of equity, inclusion, and other fundamental rights that our democratic societies are built on. Its impact poses novel challenges across disciplines. Hence governance of these emerging technologies poses questions around how should policies and regulations around AI to address the new and emerging digital barriers?

**11.** To understand, analyze and anticipate the impact of AI enabled systems, anticipatory governance models are needed to be constructed. However, there is a critical methodological hurdle vis-a-vis governance of such emerging technologies. This is illustrated by the Collingridge dilemma. It states that during the innovation process, interventions and course corrections are easy to make. However, the complete trajectory and impact of a technology is not understood and hence the need for intervention is not clear. When the need becomes apparent then interventions become costly and sometimes impossible to make. This is especially starker with AI based and data driven solutions, where decisions can be subjected to pre-exiting bias - consequence of societal prejudice as reflected by underlying data, technical bias - consequence of the constraint of the algorithm, and emergent bias - consequence of application of technology in an unanticipated context. Thus, it is imperative for an anticipatory technology governance framework to be armed with tools for auditability, and real-time course corrections. The objective of such a framework would be to facilitate the improvement of a desired outcome and

conversely, reduce the likelihood of an undesired one. Towards building an anticipatory governance models, understanding and operationalizing AI principles become crucial.

*Principles of AI*

**12.** The terms AI, Machine learning, and Deep Learning models are used interchangeably. However, it is important to understand the differences. AI models are a set of algorithms that utilizes data to recognize patterns and reach a conclusion or make predictions. Machine and Deep Learning are part of AI. Machine learning is a process where through algorithms, systems automatically learn from data and make decisions or predictions without explicitly being programmed for. Deep Learning is a class of Machine Learning where multilayered algorithms like advanced neural networks are used to extract patterns to make decisions or predictions. All machine learning models are AI models but the opposite is not necessarily true. Machine learning models can further be classified into white box models and black box models. White-box models provide solutions that are understandable while Black-box models produce solutions that are extremely hard to explain even for the designer of the model. When such models are deployed at scale, it becomes incumbent to ethically control it else it may cause catastrophic impact on the human civilization. Some of the risks due to AI deployment are job loss due to automation; privacy and security concern due to surveillance, profiling; and socio economic inequality due to bias.

*Explainability*

**13.** It is a concept where AI models and its outputs are interpretable to a human being at an acceptable level. Several machine learning algorithms like neural networks are very difficult to interpret on how the model came to a particular output. However, they remain to be a high performant. On the flip side, algorithms such as simple regression techniques or random forest methods are relatively easy to interpret but remain a low performant. Explaining outcomes of models to any stakeholder such as a user or a regulator becomes even more crucial when such models are applied to socio-economic or security applications. Due to lack of transparency and without testing for explainability, it will be difficult to assess if

AI models have been improperly applied or have misunderstood the context. Such a lack of transparency can lead to significant losses.

*Fairness*

**14.** It is a concept where AI models perform in an equitable manner. The notion of fairness can include protecting individuals from discrimination or mistreatment especially across religion, gender, race, caste, or economically disadvantaged, treats individuals impartially, and/or equitable allocation of resources. Bias in algorithms can enter into AI models due to technical limitation of its design, or used in an unanticipated context, or pre-existing social, cultural, institutional bias as captured by the data that is used to train the AI models. As AI models continue to expand and is increasingly being used to organise society, politics, and behaviour, testing for fairness becomes crucial as it can significantly impact the physical world.

*Security*

**15.** It is a concept where AI models ensure privacy and data protection of the user. Data and its analysis form one of the fundamental pillars of AI systems. In one of the more popular machine learning algorithms (supervised learning algorithms), AI systems are trained on a select set of data that have a labelled output. For instance, in some popular surveillance systems, data collected is that of faces and fingerprints. In many cases, systems around surveillance, finance, medicine, data sets contain personally identifiable information. In addition, such applications are at risk of rogue attacks in terms of theft or malicious data injection. In the next stage, the inference stage, the AI systems are deployed on the end application. Users must be protected against data leaks or misclassification. Thus, privacy and data protection become critical at the design stage of the AI systems.

*Reliability*

**16.** It is a concept where AI models operate according to their intended use case. There is a continuous requirement to monitor and validate AI systems post deployment to ensure adherence to its predefined goals. This is critical because there will be a mismatch between training data sets and real-world data. Thus, at

the very minimum recalibration and retraining of AI systems taking in user feedback will be essential to test for reliability of AI systems.

**17.** Thus, operationalization of frameworks of explainability, fairness, security, reliability into the governance architecture becomes critical to ethically control and manage AI. Towards this endeavor, measurements and indicators become increasingly important to evaluate outcomes before the technology is developed and adopted at scale. Development of such tools will facilitate informed accountability, auditability and decision-making capabilities in anticipatory governance models for emerging technologies.

*Skeleton of an anticipatory governance model*

**18.** To address the challenges of governance, anticipatory governance models should look to build capacities around the following pillars:

a. *Participatory and inclusive*: To integrate civil society considerations in line with social goals, values, and concerns into the dynamics of innovation.

b. *Ethical governance*: To operationalize guidelines vis-à-vis development of emerging technologies responsibly.

c. *Public-private coordination*: To ensure the government provides stewardship for new technologies while protecting its citizen and companies meet their social obligations.

d. *Agile and responsive regulation*: To ensure regulations are agile and responsive to changes. Currently the mechanisms of regulation are prescriptive in nature. They are rigid and take months or years to enact. In contrast, new age technologies developed in agile sprints, beta tested on early adopters and swiftly updated and adapted.

e. *Experimental*: To ensure the new age technologies are beta tested on sandboxes and accelerators. This will allow regulators to observe the

consequences of a new technology in the safety of an isolated environment.

f. *Data sharing/interoperability*: All AI enabled technologies rely on data to refine their operations. Secure data sharing and interoperable systems are needed to train, retrain, and recalibrate AI systems to continuously operate efficiently.

g. *Regulatory collaboration*: As effects of emerging technologies permeate national boundaries. Cross border collaboration is needed to manage second and third order effects that ripple out of the innovation.

*July 2022*