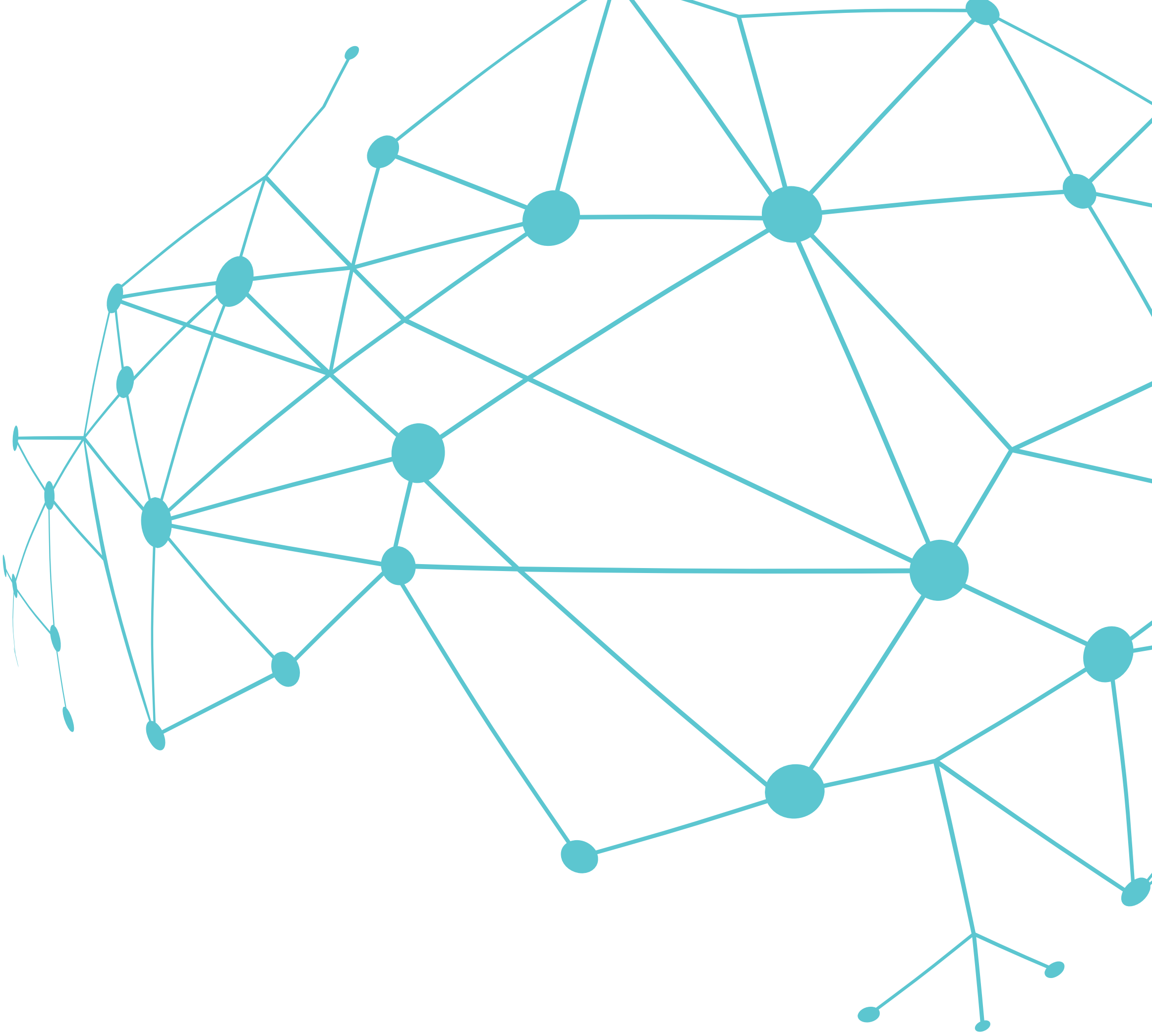
KOAN

Privacy and Security in the Age of AI

Takeaways from a Multistakeholder Roundtable held on 26th August 2023



KOAN

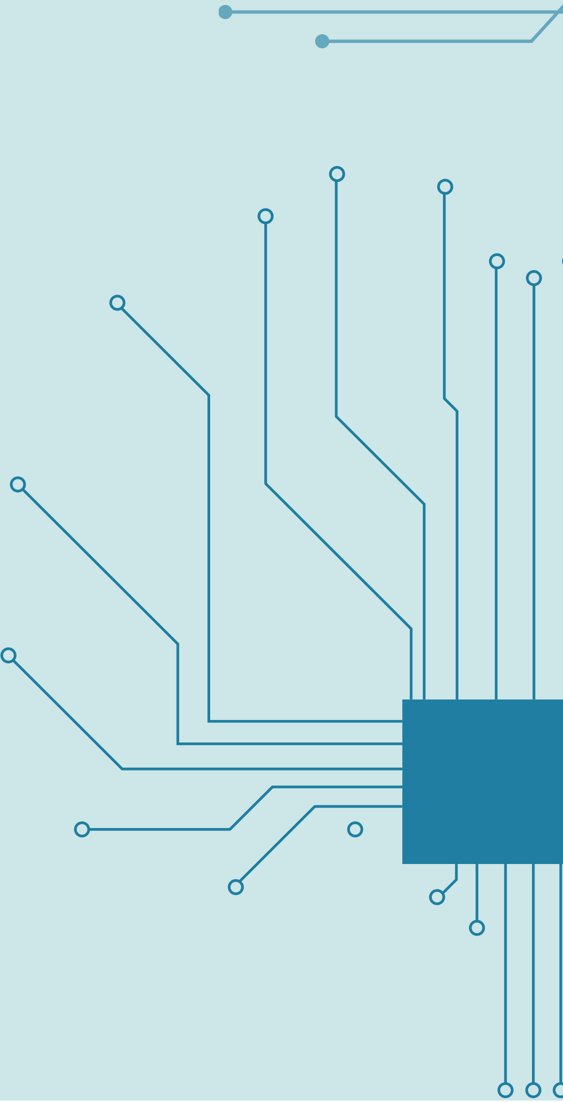
Koan Advisory Group (“Koan”), is a New Delhi based public policy consulting firm focused on new and emerging technology. It combines domain knowledge across diverse technical areas, with continuous engagement of decision makers. Koan is staffed by a multidisciplinary team of professionals, with expertise spanning law, economics, media and finance. The firm services the world’s most innovative companies, local government departments, and international organisations.

Overview

Artificial Intelligence (AI) stands as one of the most revolutionary technological advancements of the 21st century. Its potential to reshape industries, enhance productivity, and offer unprecedented solutions to complex problems is unparalleled. From healthcare diagnostics to financial forecasting, AI's applications are vast and varied. However, as with many transformative technologies, AI's integration into our digital infrastructure and daily lives also brings forth challenges.

The very nature of AI, which thrives on vast datasets and complex algorithms, makes it a focal point for privacy and cybersecurity threats. These threats range from unauthorized data access to sophisticated adversarial attacks, highlighting the dual-edged nature of AI: while it promises immense benefits, it also introduces vulnerabilities that can be exploited if not properly addressed.

With this context in mind, Koan Advisory Group organised discussion with representatives from industry, government and civil society on 26th August, in collaboration with the Information Technology Industry Council. This document captures key perspectives and recommendations from the discussion to chart a roadmap towards to promote better privacy and security standards in the AI ecosystem. This document represents Koan Advisory's synthesis of the discussions and should not be attributed to any particular speaker or institution.



Privacy and Security Concerns Surrounding AI

Privacy and security concerns surrounding AI can occur at three levels:

1. **Application:** Where the output of the AI either betrays the privacy of others or is used towards compromising security. Examples include:

   a. **Exposure of Sensitive Information**: AI can inadvertently expose or amplify sensitive information. For example, a model trained on medical or financial records might reveal personal details if not properly anonymized or secured. For instance, employees at Samsung unintentionally divulged confidential information when they used ChatGPT to generate and debug code.[1] Technically, a user could access this information by simply asking for it, which creates concerns about the exposure of sensitive information to generative AI models.[2]

   b. **Enhancing the capabilities of bad actors**: Generative AI models can be "tricked" into writing and providing malicious code that bad actors with no computer skills can use for cyber-attacks.[3] Bad actors may do this by tricking the AI to break away from protocol restrictions through indirect prompts.[4]

2 **Pre-Trained Model**: Large language models are complicated and expensive to train and create. As such, many of these models can be used to form the basis of other AI applications. For instance, Meta's Llama 2, a partially open-sourced large language models offers inventors the capability to create their own generative AI. The privacy and security risks from pre-trained large-language models can come in several forms, three of which include:

   a. **Data Poisoning**:[5] This involves "contaminating" the data used to train AI in order to make it generate erroneous responses.

   b. **Model Inversion Attacks**:[6] In open sourced models, where attackers have access to the parameters, model, as well generative AI, they can use the generative AI model to create synthetic images that may reveal the personal information within the open-sourced model's training data. For instance, this process could be used to generate synthetic images of faces to reveal the images of actual people that were used for training.

   c. **Data Collection and Usage:** AI systems often require vast datasets for training and operation. The collection, storage, and processing of this data can infringe on individual privacy, especially if done without explicit consent or transparency. For instance, ChatGPT's training process involved scraping data from various corners of the internet. It is likely that such data involved personal data of users, which was taken without procuring express consent.[7]

3. **Infrastructure:** AI infrastructure, which encompasses the hardware and cloud platforms powering AI applications, is becoming a critical component of modern digital ecosystems. However, this infrastructure is not immune to security vulnerabilities. The reliance on cloud platforms for AI processing and data storage introduces potential breach points where vast amounts of sensitive data can be compromised.

It stands to reason, then, that a comprehensive set of actions is required to guard against security and privacy vulnerabilities in different components of AI systems. The following recommendations are a synthesis of our discussions, and do not represent the views of any particular stakeholder. These are early ideas and will inevitably evolve.

Recommendations

1. **Identify Critical Areas**: Identify and define areas where AI, especially large language models, are used for critical purposes, such as biomedical sciences, health evaluations, and citizen safety. These areas may require stricter regulations and oversight.

2. **Establish Safety Breaks**: Drawing inspiration from the invention of the elevator, consider implementing safety breaks in generative AI, especially in high-risk areas. This would allow human intervention to halt or control AI processes if they go awry. In addition, companies must also create internal risk management protocols for the evaluation and redressal of privacy and security risks stemming from their AI systems.

3. **Facilitate Transparency through Disclosures**: Given that there is currently a black-box problem in AI i.e. it is not completely certain how AI models arrive at their output – disclosures are a more viable alternative to operational transparency for AI. Users and stakeholders should be aware of how AI systems operate and make decisions. There is a need to ensure that AI service providers offer clear disclosures to consumers about how AI systems operate, the data they use, and the protections in place.

4. **Apply the Know-Your-Customer Principle to AI Ecosystems**: Financial businesses are required to carry out Know-Your-Customer (KYC) procedures to identify customers, establish risk profiles, and keep a lookout for suspicious activity.[8] The KYC concept could be applied to the AI ecosystem through a licensing and KYC mandate, where deployers and developers may only deal with service providers such as data centres that were authorised and licensed.[9]

   KYC may also be considered for users of AI systems. Many harms prompted by AI systems are difficult to trace. For instance, in the context of deepfakes, which rely on generative AI systems to be created, detection methods are coming into place. However, these are not fool-proof. Knowing which individuals or entities have access to these systems could go towards minimising their misuse. It also puts entities in a position to be able to decide whether a prospective customer is a suitable and safe candidate to use the technology or not.

5. **Enact Precision Regulation**: Instead of broad, monolithic regulations, adopt a precision regulation approach that targets specific areas of concern without stifling innovation. Precision regulation has a better chance of address harms presented by different AI models. Conversely, scholars have critiqued omnibus AI regulation like the EU's AI Act as being overly prescriptive while also encompassing enough loopholes so as to be largely ineffective at mitigating harms.[10] Decision-makers must instead look to address the specific harms presented by different AI models.

6. **Examine How Current Laws Apply to AI**: AI development is not happening in a vacuum. There are laws pertaining to data protection and security that will currently apply. Policymakers must evaluate how current laws apply to AI development and deployment and endeavour to fill any gaps that may exist.

7. **Foster Development of Standards**: Policymakers must collaborate with global and domestic standards organizations to develop standards that focus on responsible AI deployment and consumer disclosures. Standards can go a long way towards building consumer trust and also enable better oversight of AI development by creating benchmarks against which AI systems can be audited. They also preserve operational and design autonomy, retaining flexibility for innovators. In terms of a vision for standards for AI, the World Trade Organisation's Principles for the Development of International Standards, Guides and Recommendations are instructive in this instance. They provide that standards must be relevant, coherent, devised in a transparent and open manner, and based on consensus.[11] Such standards should focus particularly on deployment issues such as trust and safety, which will be a key means of managing the fallout from the misuse of AI systems down the line.

8. **Adopt an All-of-Society Approach to AI Governance**: AI governance should not be the sole responsibility of one stakeholder. Instead, it should involve the participation and collaboration between different entities hailing from the private sector, civil society, academia, and government to ensure a holistic governance mechanism.

9. **Educate and Inform**: Given the rapid evolution of AI, continuous education and awareness campaigns are essential. This ensures that both the public and professionals understand the capabilities, limitations, and risks associated with AI. Building public awareness is particularly important for the success of transparency provisions such as disclosures.

10. **Distinguish Roles in AI Deployment**: Clearly define and distinguish the roles of AI developers, deployers, and users. Each has unique responsibilities and potential risks, and regulations and standards should address these distinctions.

11. **Introduce Measures to Restrict Adversaries and Bad Actors from Accessing AI Infrastructure:** Such measures could include licensing of models, restricting the operations of supply chains in territories or entities affiliated with adversaries, and restricting access to key components used to develop AI such as graphic processing units.

List of Speakers

- **Akash Tripathi,** CEO, MyGov, Ministry of Electronics and IT, Government of India
- **Ashutosh Chadha**, Director and Country Head for Government Affairs and Public Policy, Microsoft, India
- **Jason Oxman,** President and Chief Executive Officer, ITI Council
- **Kishore Balaji,** Executive Director for Government Affairs, IBM South Asia
- **Meghna Bal,** Head of Research, Esya Centre.
- **Sunil Abraham,** the Director for Data Governance and Emerging Tech, Meta India

Endnotes

[1] Gupta, Maanak & Akiri, CharanKumar & Aryal, Kshitiz & Parker, Eli & Praharaj, Lopamudra “From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy.” arXiv:2307.00691 (2023)

[2] Ibid

[3] Ibid

[4] Ibid

[5] Wu, Xiaodong, Ran Duan and Jianbing Ni. “Unveiling Security, Privacy, and Ethical Concerns of ChatGPT.” ArXiv abs/2307.14192 (2023)

[6] Hintersdorf Dominik, Struppek Lukas, Kersting Kristian,“Balancing Transparency and Risk: The Security and Privacy Risks of Open-Source Machine Learning Models.” arXiv:2308.09490 (2023)

[7] Wu, Xiaodong, Ran Duan and Jianbing Ni. “Unveiling Security, Privacy, and Ethical Concerns of ChatGPT.” ArXiv abs/2307.14192 (2023)

[8] https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2023/08/MSFT_Governing_AI_BlueprintFuture_India_web-2.pdf

[9] ibid.

[10] https://www.wto.org/english/tratop_e/tbt_e/principles_standards_tbt_e.htm

[11] https://www.wto.org/english/tratop_e/tbt_e/principles_standards_tbt_e.htm

KOAN

contactus@koanadvisory.com | www.koanadvisory.com