CoRE-AI
Coalition for Responsible Evolution of AI

Infosys
topaz

EVENT REPORT

Reimagining Data Protection for AI Landscape

CoRE-AI and Infosys

EVENT REPORT

REIMAGINING DATA PROTECTION FOR AI LANDSCAPE

CORE-AI AND INFOSYS

Established in July 2024, **CoRE-AI™ (Coalition for Responsible Evolution of AI)** is a prominent multi-stakeholder initiative hosted by The Dialogue, focused on fostering the responsible and ethical development of AI technologies. By bringing together stakeholders from government, industry, academia, startups, and civil society, CoRE-AI aims to drive collaborative efforts that address the risks associated with AI while maximizing its societal benefits. The initiative seeks to guide India's AI journey, ensuring that technological advancements align with ethical standards to benefit the broader public.

The Dialogue® is a public policy think tank with a vision to drive a progressive narrative in India's policy discourse. Founded in 2017, we believe in facilitating well-researched policy debates at various levels to help develop a more informed citizenry, on areas around technology and development issues. The Dialogue® has been ranked as the world's Top 10 think tanks to watch out for, by the Think Tank and Civil Societies Programme (TTCSP), University of Pennsylvania in their 2020 and 2021 rankings.

Infosys Topaz, launched on May 23, 2023, is an AI-first suite of services, solutions, and platforms utilizing generative AI technologies. It aims to enhance human potential, drive enterprise transformation, and benefit communities by enabling value creation through unprecedented innovations, connected ecosystems, and pervasive efficiencies. The platform leverages over 12,000 AI assets, 150+ pre-trained AI models, and 10+ AI platforms, all guided by AI-first specialists and data strategists. Infosys Topaz operates with a 'responsible by design' approach, ensuring strict adherence to ethics, trust, privacy, security, and regulatory compliance. By building an AI-first core through the Infosys Applied AI framework, it empowers individuals and organizations to deliver cognitive solutions that accelerate growth, build connected ecosystems, and unlock efficiencies at scale.

For more information
www. core-ai.in, www.thedialogue.co and https://www.infosys.com/services/data-ai-topaz.htm

Suggested Citation
Reimagining data protection for AI landscape (January, 2025). CoRE-AI and Infosys Topaz.

Catalogue No.
CAI/ER/0125/01

Publication Date
January 28, 2025

Contents

1. Introduction

The intersection of artificial intelligence (AI) and data protection has emerged as a critical area of concern, particularly in light of the increasing reliance on personal data to drive AI innovation. Artificial intelligence relies heavily on the ability of the models to scrape, collect, and process various forms of data, including publicly available personal information, semi-public or private data, and licensed information that may contain personal details. The more quality data a model processes, the more efficient it becomes. However, the imbalance between data utility and privacy protection has raised serious concerns.

To address these critical issues, CoRE-AI, in collaboration with Infosys, organized three in-person panel discussions at the Infosys campus in Bengaluru on the 18th of September, 2024. These discussions focused on AI's privacy challenges and examined how existing data protection laws apply to and regulate AI technologies. The sessions drew attendance from startups working in the AI space, established industry players, and students interested in the field. With 11 panelists participating across the three sessions, the discussions provided a comprehensive look at different aspects of the interface between data protection and artificial intelligence. Additionally, the panels also explored regulatory innovations and tools necessary to establish the legal basis for processing digital personal information in developing and deploying AI solutions.

The insights gathered during these sessions will provide a valuable roadmap for navigating the complexities of data protection in the age of AI, ultimately aiming to safeguard individual rights while enabling technological advancement.

2. Key Areas of Focus

A. Determining Legal Base for Processing Personal Data for AI Innovation

One of the most fundamental concerns regarding AI and data protection is the legal basis for processing personal data. Traditionally, consent has been the cornerstone of data protection laws. However, with the vast amounts of data involved in AI training, relying solely on consent presents significant challenges, such as consent fatigue and the inadequacy of traditional mechanisms to ensure informed, voluntary, and easily reversible consent.[1] Furthermore, users often feel compelled to provide consent due to their reliance on essential services,[2] including social media platforms and smartphones. Alternatives, such as providing users with clearer choices via access to advisory services, have been floated as potential solutions to improve the management of consent. Intermediary models, such as the Account Aggregator system—which facilitates structured data sharing—are part of the discourse as possible means to manage data flows in a transparent and controlled manner.

Consent must be accessible, easy to understand, and broken down into clear, manageable components. The introduction of advisory services is considered key to assisting users in navigating and negotiating consent. Furthermore, strategies to improve consent mechanisms must be focused on making consent more dynamic, rather than static. For example, cookie mechanisms have been proposed as a way to offer more restrictive data-sharing options,[3] with consent models allowing users to update their preferences over time. The concept of stewardship, as demonstrated by the Account Aggregator model licensed by the Reserve Bank of India (RBI), is cited as a successful example of consent intermediaries facilitating financial data sharing[4].

On the issue of processing sensitive personal data, there is broad consensus that using the "legitimate interest" clause is inappropriate, particularly in relation to sensitive data. Challenges surrounding publicly available data have also been highlighted, especially where such data is exempt from consent requirements. While anonymized aggregate datasets used in AI models help mitigate privacy concerns, the effectiveness of anonymization in protecting personal data must still be evaluated.

Beyond consent, the need for industry-wide technical standards has been underscored to ensure that data processing for AI innovation aligns with privacy norms. Data governance has emerged as a key area of focus, with calls to develop tools for data protection audits,

[1] Longpre, S., et al. (2024, April). Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them? [Research paper]. https://arxiv.org/abs/2404.12691

[2] Adams, B., Clark, A., & Craven, J. (2018, April 23). It is Free and Always Will Be - Trading Personal Information and Privacy for the Convenience of Online Services [Research Paper]. https://doi.org/10.48550/arXiv.1804.08491

[3] Zendata. (2024, July 4). Consent management 101: Navigating user consent for data collection and use. https://www.zendata.dev/post/consent-management-101-navigating-user-consent-for-data-collection-and-use

[4] Reserve Bank of India. (2021). Account Aggregator framework: Consent-based financial data sharing.. https://rbidocs.rbi.org.in/rdocs/Bulletin/PDFs/01SP_170520210603BF5000A54466AB160A1F9AF9F404.PDF

responsible AI checklists, data anonymization standards, and robust data security measures. Given the unstructured nature of many available datasets in India, there is an urgent need for stricter regulations to ensure the proper segregation of sensitive data.

B. Determining Lawful Utilization of Publicly Available Personal Information for AI Innovation

The vital role of accessing publicly available data while addressing privacy concerns is a critical issue in the development of AI models. The scalability and accuracy of AI heavily rely on data, and restricting the use of publicly available personal data could impede AI advancement. However, balancing the need for data accessibility with safeguarding privacy remains a significant challenge. The application of data protection laws to AI training data continues to be a topic of ongoing deliberation, with the legislative and interpretive challenge being to strike the right balance between protecting individuals and fostering innovation.

In India, the country's historical support for an open data policy,[5] even before the advent of the GDPR, has laid the foundation for promoting domestic research and innovation. Making aggregated and anonymized personal data available for research purposes can foster innovation without compromising privacy, operationalizing India's vision for an open data policy.

Additionally, a centralized digital data sandbox containing encrypted PII and consistent anonymization protocols could facilitate research. Properly anonymized and aggregated data from sectors like criminal justice, healthcare, and legal statutory bodies could serve as valuable resources for researchers and innovators. Similar to the Financial Services Information Sharing and Analysis Center (FS-ISAC)[6], a model could be adopted to process consumer grievance data to improve products and services without compromising privacy. This approach ensures that shared data does not violate privacy standards. There is also a growing recognition of the need to integrate techno-legal measures to protect privacy and data security while developing technology stacks for startups to access data securely.

Web scraping in India remains under-regulated[7], with loopholes in compliance measures. To facilitate research and innovation, there is a need to operationalize data scraping methods, ensuring that anonymized and aggregated datasets are accessible without the risk of misuse. However, clear guidelines and regulations must govern this process to prevent malicious exploitation of data. As data monetization grows, it becomes essential to establish clear frameworks that balance commercial interests with individual privacy rights, mitigate anti-competitive effects, and promote ethical data practices.

The handling of publicly available data also varies across jurisdictions. For instance, the EU's GDPR mandates that publicly available data is not exempt from privacy laws[8], requiring compliance with accountability and

[5] Verma, N., Mishra, A. (2018). Open Data: India's Initiative for Researchers, Research, and Innovation. In: Munshi, U., Verma, N. (eds) Data Science Landscape. Studies in Big Data, vol 38. Springer, Singapore. https://doi.org/10.1007/978-981-10-7515-5_20

[6] Financial Services Information Sharing and Analysis Center. https://www.fsisac.com

[7] S.S. Rana & Co. Advocates. (2020, February 3). Data scraping and legal issues in India. https://www.mondaq.com/india/copyright/900156/data-scraping-and-legal-issues-in-india

[8] European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). [Official Journal] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504

transparency measures in the absence of consent[9]. Legitimate business interests must be specified, and data must not infringe on individuals' fundamental rights. Conversely, Singapore's PDPA offers more flexibility, permitting publicly available PII for research and innovation[10], with recommendations for impact assessments and data minimization. In the US, publicly available data is exempt from privacy laws but must be disclosed if sold or shared with third parties.[11]

Non-privacy aspects of data processing, such as data collection practices, are also of concern. Financial institutions, for example, must be transparent about the data they use, ensuring that data minimization is prioritized. The collection of data for alternative lending has been a longstanding practice[12], but ethical AI use must adhere to principles like consent, fairness, confidentiality, and the avoidance of bias. These discussions highlight the need for balanced regulatory frameworks that protect privacy while enabling AI innovation both in India and globally.

C. Impact of the Digital Personal Data Protection (DPDP) Act 2023 on AI Innovations

The implementation of responsible AI and data protection regulations, particularly in the context of the Digital Personal Data Protection (DPDP) Act, has significant implications for AI innovation, especially for startups. There is a growing discourse around how the DPDP Act impacts AI-driven startups, with an emphasis on adopting a privacy-by-design approach from the outset.[13] This approach is critical due to the global nature of AI solutions, which must address data protection concerns across multiple jurisdictions.

Startups face particular challenges in data profiling, classification, and incident handling, necessitating comprehensive training for employees and vendors to effectively manage privacy concerns. Existing privacy technologies have been floated as a means to manage personal information within large language models, with privacy-by-design being viewed as a potential competitive advantage for startups.[14]

In terms of regulatory frameworks, there is a consensus that a tailored, sector-specific approach to AI regulation would be more beneficial for startups than a one-size-fits-all model. The development of common infrastructure and standardized assessment tools has been highlighted as a way to assist startups in navigating regulatory compliance. Certified authorities offering pre-packaged compliance solutions are part of the discourse, alongside the need to establish greater trust in larger players' adherence to privacy regulations. Standardizing data across various public sources and APIs, such as those related to PAN cards, remains a challenge, with regulatory sandboxes seen as a potential solution. These sandboxes could

[13] MediaNama. (2024, February). Understanding the essence and impact of Privacy by Design. https://www.medianama.com/2024/02/223-dpdp-act-understanding-privacy-by-design-2/

[14] Feretzakis, K., Papaspyridis, K., Gkoulalas-Divanis, A., & Verykios, V. S. (2024). Privacy-Preserving Techniques in Generative AI and Large Language Models. Information, 15(11), 697. https://doi.org/10.3390/info15110697

help address the distinct challenges startups face when managing consumer data versus enterprise data, the latter often involving sensitive information. Additionally, the legality of web scraping and data collection practices is complex and varies across jurisdictions. India needs to establish clear guidelines and regulations to govern these practices, ensuring that data is used ethically and responsibly.

The use of enterprise data, particularly in enhancing coding systems and addressing intellectual property violations, has also been a key area of discussion. Transparent and trustworthy systems are seen as essential, with enterprises encouraged to leverage upcoming data protection laws to build consumer trust. Demonstrating responsible AI practices through rigorous testing and benchmarking has been proposed as a way to foster innovation and ensure compliance with regulations.

There is a growing recognition of the importance of self-certification for AI systems, with the view that government support for

3. Recommendations

1. *Dynamic and Informed Consent:* Consent mechanisms must evolve beyond static models. Implement advisory services and real-time preference updates, similar to cookie consent systems, to empower users with clear, manageable choices and ensure informed, voluntary, and easily reversible consent.

2. *Stewardship Models for Data Sharing:* Adopt intermediary models like the Account Aggregator system to facilitate transparent, secure, and user-controlled data sharing, particularly for sensitive data, aligning with privacy norms while supporting innovation.

3. *Anonymization Standards:* Develop robust standards for data anonymization to address privacy concerns in AI training. While anonymization can mitigate risks, continuous evaluation of its effectiveness is necessary to safeguard personal data.

4. *Publicly Available Data Usage:* Establish clear, sector-specific guidelines for handling publicly available personal data to balance privacy with AI research needs. Incorporate principles like data minimization and impact assessments to ensure ethical data usage.

5. *Legal and Regulatory Framework:* The legal landscape for data access and usage is complex and varies across jurisdictions. India needs to establish clear guidelines and regulations to govern data practices, ensuring that data is used ethically and responsibly.

6. *Privacy-by-Design for Startups:* Encourage AI-driven startups to adopt privacy-by-design practices to manage personal data responsibly. Regulatory sandboxes can provide tailored compliance solutions, helping startups navigate complex privacy and data protection challenges.

7. *Techno-Legal Framework for Data Scraping:* Implement clear guidelines for web scraping practices to ensure ethical and responsible data usage, protecting both privacy and commercial interests. Regulations must prevent misuse and ensure compliance with privacy laws.

8. *Sector-Specific Regulatory Frameworks:* Advocate for sector-specific AI regulations rather than a one-size-fits-all approach, ensuring startups can comply with privacy standards while fostering innovation. Certified authorities can assist in providing pre-packaged compliance solutions.

9. *Third-Party Certification and Self-Certification:* Promote responsible AI practices through self-certification and third-party certifications. Establishing globally recognized standards, such as those advocated by CoRE-AI, will help build trust and credibility in AI systems.

4. Annexure: Panelists

Session 1: Determining Legal Base for Processing Personal Data for AI Innovation

The Panel was moderated by *Kamesh Shekar, Senior Programme Manager, at The Dialogue*. The discussion fractured the following panelists:

- *Ms. Soujanya Sridharan*, Senior Manager, Aapti Institute
- *Mr. Rajesh Kumar Viswanathan*, Senior Group Manager - Privacy and Data Protection, Infosys
- *Mr. Vinay Kesari, Director* - Operations & Strategy, Setu Account Aggregator
- *Ms. Shweta Mohandas*, Researcher, Centre for Internet and Society

Session 2: Determining Lawful Utilization of Publicly Available Personal Information for AI Innovation

The panel was moderated by *Jameela Sahiba, Senior Programme Manager, at The Dialogue*. The discussion fractured the following panelists:

- *Ms. Rama Vedashree*, Former CEO, DSCI
- *Mr. Subhashis Nath*, AVP, Delivery Head, Analytics and AI, Infosys
- *Ms. Aadya Misra,* Counsel - Privacy, Data Protection and Cybersecurity, Spiceroute Legal
- *Ms. Anubhutie Singh*, Research Associate, Dvara Research

Session 3: Impact of DPDP Act 2023 on AI Innovations

The panel was moderated by *Kazim Rizvi, Founding Director, The Dialogue*. The discussion fractured the following panelists:

- *Mr. Manjunatha Gurulingaiah Kukkuru*, VP - Principal Research Analyst, iCETS, Infosys
- *Mr. Sridhar Vaidyanathan*, Chief Revenue Officer, Myelin Foundry
- *Mr. Murugan Chidhambaram*, Head of Digital Transformation, AquaConnect

core-ai.in

@Core-Ai

@CoalitionforResponsibleEvolutionofAI

thedialogue.co

@_DialogueIndia

@The-Dialogue-India

@TheDialogue_Official

@TheDialogue

infosys.com/services/data-ai-topaz

@InfosysTopaz