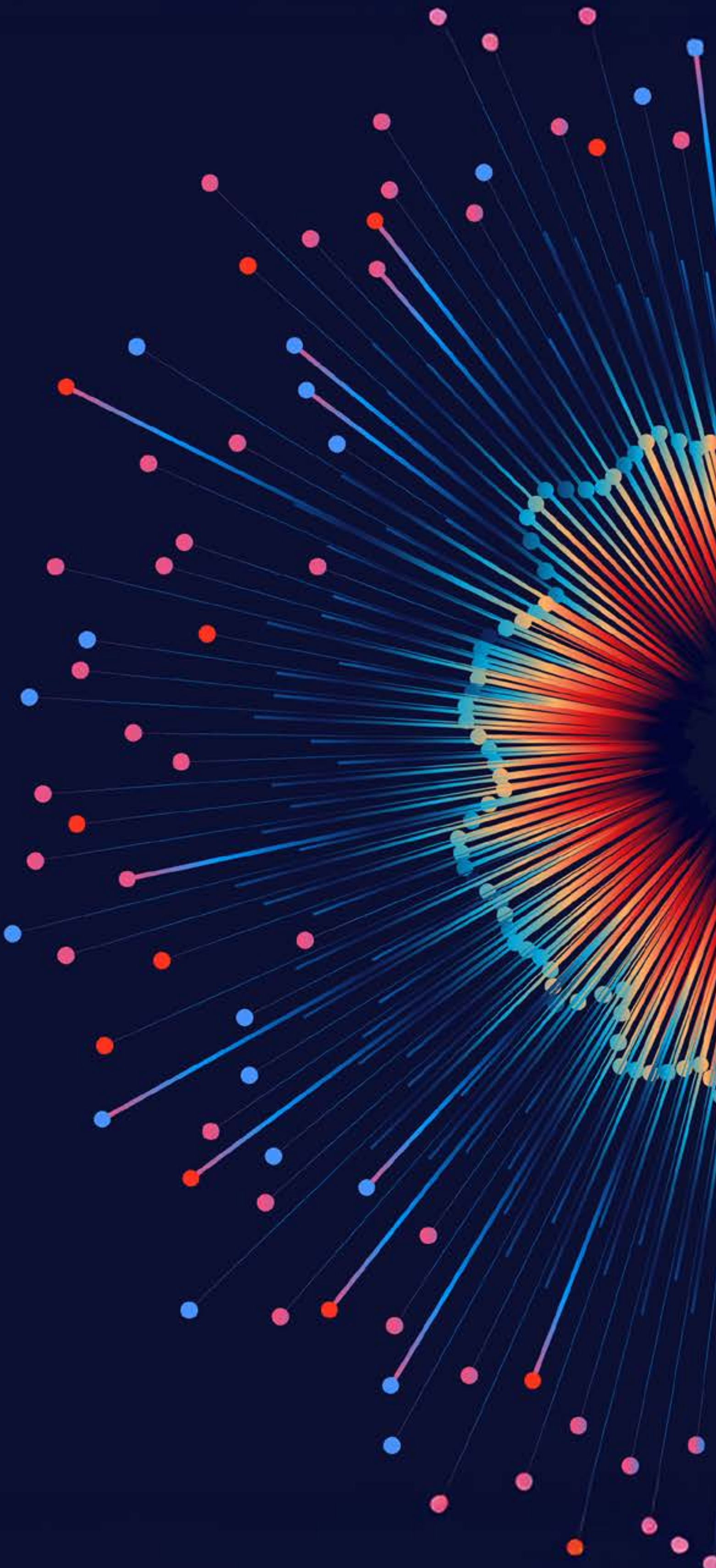
german cooperation
DEUTSCHE ZUSAMMENARBEIT
Implemented by giz
FAIR FORWARD Artificial Intelligence for all.
IKIGAI LAW
DSCI PROMOTING DATA PROTECTION
nasscom

Handbook on **Data Protection and Privacy** *for* **Developers of Artificial Intelligence (AI)** *in India*



January 2026

Published by:
Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

Registered offices:
Bonn and Eschborn, Germany

FAIR Forward Artificial Intelligence for All
A2/18, Safdarjung Enclave,
New Delhi-110029, India
T: +91 11 4949 5353
F: + 91 11 4949 5391
E: fairforward@giz.de
W: https://www.bmz-digital.global/en/overview-of-initiatives/fair-forward/

Authors: Ikigai Law (Ajey Karthik, Aman Taneja, Sreenidhi Srinivasan and Vidushi Sinha)

Design, Layout and Printing: MKM Creatives

On behalf of
German Federal Ministry for Economic Cooperation and Development (BMZ)

Disclaimer: The data in the publication has been collected, analysed and compiled with due care; and has been prepared in good faith based on information available at the date of publication without any independent verification. However, Deutsche Gesellschaft für In-ternationale Zusammenarbeit (GIZ) GmbH does not guarantee the accuracy, reliability, completeness or currency of the information in this publication. GIZ shall not be held liable for any loss, damage, cost or expense incurred or arising by reason of any person using or relying on information in this publication.

New Delhi, India
January 2026

Table of Contents

जितिन प्रसाद
JITIN PRASADA

सत्यमेव जयते

राज्य मंत्री
वाणिज्य एवं उद्योग,
इलेक्ट्रॉनिकी और सूचना प्रौद्योगिकी
भारत सरकार
Minister of State
Commerce & Industry,
Electronics and Information Technology
Government of India

MESSAGE

India's vibrant startup ecosystem is a testament to our innovation capabilities and entrepreneurial spirit. Within this dynamic ecosystem, AI startups are driving transformative changes across key sectors such as healthcare, agriculture, education, and governance. These innovations reflect India's commitment to inclusive and sustainable development, powered by technology.

Positioned as a voice for the Global South, India is uniquely placed to shape the future of AI by championing a development-centric and rights-based approach. Through the *IndiaAI Mission*, the Government of India is advancing an AI vision rooted in public trust, ethical design and responsible deployment. As we approach the 2026 AI Impact Summit, these values will be central to shaping both national priorities and the global dialogue on AI governance.

The responsible use of data lies at the heart of this vision. As AI systems become increasingly complex and data-driven, safeguarding privacy and ensuring accountability are imperative. The Digital Personal Data Protection Act, 2023 (DPDP Act), and its forthcoming Rules mark a watershed moment in India's digital journey, balancing developer obligations with Data Principal rights. With this background, the *Handbook on Data Protection and Privacy for Developers of AI in India* comes at a timely juncture.

The Handbook offers practical guidance to AI developers to help operationalise the DPDP Act at an organisational and model level. It translates the principles of the DPDP Act into actionable guidance tailored for real-world AI development. With practical checklists, sector-relevant case studies, and a clear focus on implementation, it empowers developers, startups, and organizations to embed privacy and accountability into the core of their AI systems.

I commend the initiative led by GIZ's FAIR Forward – AI for All project (funded by the German Federal Ministry for Economic Cooperation and Development (BMZ)), developed in close collaboration with Ikigai Law, NASSCOM, and the Data Security Council of India (DSCI). By grounding its recommendations in India's legal and policy frameworks, this Handbook contributes meaningfully to the "Trust and Safety" pillar of the IndiaAI Mission. Built on wide-ranging stakeholder consultations, this resource is both contextually grounded and uniquely suited to the needs of India's A! startup ecosystem. I extend my sincere thanks to the legal experts, industry leaders, technologists, and civil society voices who contributed their insights to this effort.

This Handbook is a valuable resource for India's growing AI ecosystem. I hope it serves as a model for translating regulatory vision into practical and impactful action for AI developers.

(Jitin Prasada)

Digital India
Power To Empower

स्वच्छ भारत
एक कदम स्वच्छता की ओर

Acknowledgements

This Handbook has been developed as part of the FAIR Forward – AI for All initiative by GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit), funded by the German Federal Ministry for Economic Cooperation and Development (BMZ), in close collaboration with Ikigai Law, the Data Security Council of India (DSCI), and NASSCOM.

The Handbook was researched and authored by the team at Ikigai Law, who also led the consultation process. We acknowledge the contributions of Ajey Karthik, Aman Taneja, Sreenidhi Srinivasan, and Vidushi Sinha. The Ikigai team also thanks Dr. Debayan Gupta for his support, and Amol Kulkarni for his inputs on an earlier version of the draft.

We are grateful to Wadhwani AI and Digital Green for sharing their case studies that feature in this Handbook.

We also extend our sincere thanks to Anirudh Rastogi, Nehaa Chaudhari, Rutuja Pol, Chikita Shukla, Himali Sylvester and Aarya Pachisia from the Ikigai team; Shivangi Malhotra and Deepa Ojha from DSCI; Bhavika Nanawati, Phillip Olbrich, and Aishwarya Salvi from GIZ; and Ankit Bose, M.Chockalingam, Raj Shekhar and Simrandeep Singh from Nasscom for their guidance, feedback, and close collaboration throughout the project. Their feedback, support, and close collaboration were instrumental to the development of this Handbook.

We also thank the many experts, industry bodies, and civil society representatives who participated in the consultations, including: Abhishek Jain, Abilash Soundarajan, Akash Kulgod, Alka Pachori, Amber Sinha, Amit Jain, Amoha Basrur, Ankit Bajaj, Aparna Pathak, Ashwathy Babu Pillai, Asmi John Akkara, Aswathy Asok, Ayush Singhal, Balaraman Ravindran, Bilal Mohamed, Brinda Lashkari, Bunty Kundnani, Chaitanya Nittala, Danielle Gilliam Moore, Dhanyashri Kamalakannan, Fahad, Fawaz, Harsha Vardhan Jonnalagadda, Hemanth, Isha Suri, Jagannath PV, Jhalak Kakkar, Jishnu Nair, Kapil Chaudhary, Kartikeya Raman, Krity Kansara, M S H Shankar Datta, Mayank Manish, Mihir Kulkarni, Mira Swaminathan, Naveen Setia, Neha Misra, Nisha Sethi Girdhar, Nitendra Rajput, Parminder Singh Kakria, Pavan M Laxmeshwar, Piyush Srivastav, Pooja M, Pranav Bhargava, Prasaanth Balraj, Prekshaa Arunachalam, Priya Muku Gora, Priyanka, Prof Aaditeshwar Seth, Prof Venkata Pingali, Rajdeep Endow, Rohan Mitra, Roli Sharma, Satyajit Gupta, Saurabh Singh, Saurav Banerjee, Shyam Ravikumar, Soumya, Sowmya, Sriram Raghavan, Supratik, Sushen Duhan, Tanvi Pillai, Titiksha Vashisht, Trishee Goyal, Vakasha Sachdev, Venkatesh Krishnamoorthy, Vineet Singh, Vishnu Ramesh, Vivek Abraham, Vivek Seshadri. Their thoughtful inputs helped shape the practical orientation and relevance of this document.

Acknowledgement of these individuals does not imply their endorsement, nor the endorsement of any organisations with which they are affiliated, of the contents of this Handbook or its conclusions.

List of Abbreviations

AI	Artificial Intelligence
ABAC	Attribute-Based Access Control
API	Application Programming Interface
BBQ	Bias Benchmark for QA
CCTV	Closed Circuit Television
CERT-In	Indian Computer Emergency Response Team
CMA	Competition and Markets Authority
CNIL	Commission Nationale Informatique & Libertés
COPPA	Children’s Online Protection Act
CT Scan	Computed Tomography Scan
DF	Data Fiduciary
DP	Data Principal
DPA	Data Protection Authority
DPDP Act	Digital Personal Data Protection Act, 2023
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
DSA	Digital Service Act
EU	European Union
FIR	First Information Report
FTA	Farm Tele Advisors
FTC	United States Federal Trade Commission
GDPR	General Data Protection Regulation
HBNC	Home-Based Newborn Care
ICMR	Indian Council of Medical Research
ICO	Information Commissioner’s Office

ICRISAT	International Crops Research Institute for the Semi-Arid Tropics
ID proof	Identity Proof
IMD	Indian Meteorological Department
IMDb	Internet Movie Database
INR	Indian Rupee
IP Address	Internet Protocol Address
IP Infringement	Intellectual Property Infringement
ISO	International Organization for Standardization
IT Act	Information Technology Act
IVR	Interactive Voice Response
KYC	Know Your Customer
LLM	Large Language Model
MFA	Multi-Factor Authentication
ML	Machine Learning
MMLU	Massive Multitask Language Understanding
OECD	Organisation for Economic Co-operation and Development
OOD	Out of Distribution
PAN	Permanent Account Number
PDPA	Personal Data Protection Act
PDPC	Personal Data Protection Commission
PETs	Privacy Enhancing Technologies
PIPEDA	Personal Information Protection and Electronic Documents Act
POC	Point of Contact
PTM	Privacy Threat Modelling
PwD	Persons with Disability
RAG	Retrieval-Augmented Generation
RBAC	Role-Based Access Control

RBI	Reserve Bank of India
SDF	Significant Data Fiduciary
SOP	Standard Operating Procedure
TEC	Telecommunication Engineering Centre
UK	United Kingdom
UK ICO	United Kingdom Information Commissioner's Office
USD	United States Dollar
UT	Union Territory

How to read this Handbook?

This Handbook is designed as a practical guide for developers of AI systems, especially from early-stage startups, to navigate data protection obligations and ethical considerations, in a clear and actionable manner.

Rather than serving as a comprehensive legal or technical manual, the Handbook complements existing global and domestic resources on data protection and responsible AI. It offers a context-aware framework grounded in core legal and ethical principles, encouraging developers, product teams, and founders to interpret and apply these principles in ways best suited to their technology and user base.

Drawing from existing frameworks, the Handbook offers recommendations across the lifecycle of an AI system. For the purposes of this Handbook, these stages include:[1]

01 **Conception and design:** This stage involves defining the AI system's purpose, intended users, and overall goals. Key decisions are made about its scope, users, functionality and performance expectation, model choice, among others. Teams also identify the types of data needed, determine appropriate sources, and ensure that data is collected and used lawfully, through consent, or other appropriate legal basis.

02 **Development:** In this phase, the AI system is built, refined, and tested. Teams focus on how the system functions - ensuring it works as intended across different user groups, and addressing potential risks such as unfair outcomes, lack of transparency, or security vulnerabilities.

03 **Deployment:** Real-world deployment of the AI system, with ongoing monitoring to catch model drift, performance issues, and emerging risks.

The Handbook is divided into two main sections:

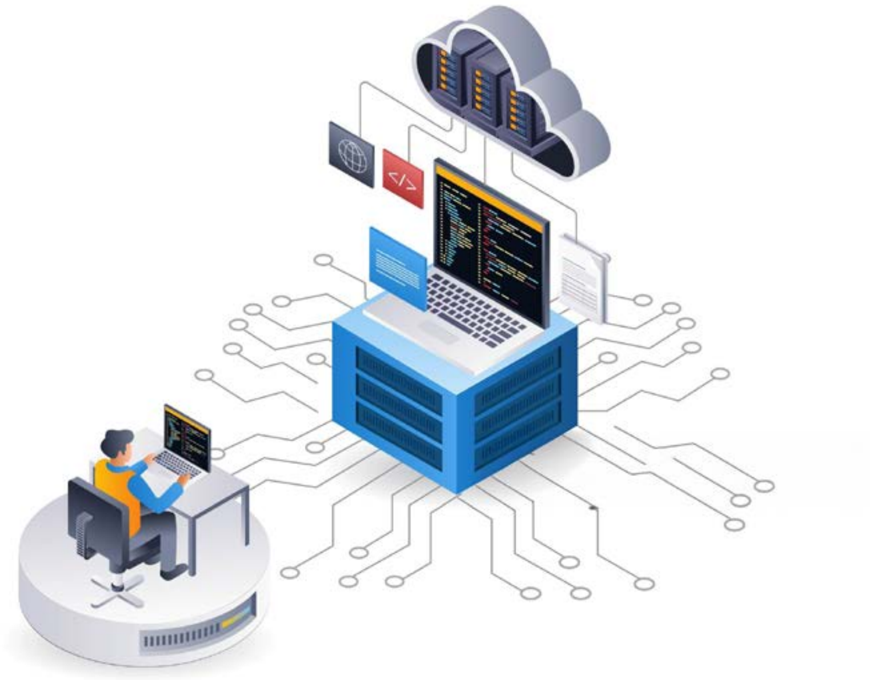
- **Section I: Data Protection** – which unpacks key concepts and compliance requirements under India's data protection law, with a focus on their relevance to AI development.
- **Section II: Responsible AI –** which explores AI development through widely recognised responsible AI principles and provides a framework for their practical application.

Each section concludes with a checklist of actionable takeaways intended to support developers in embedding privacy and ethical safeguards from the earliest stages of product design through to deployment.

We have annexed a few case studies at the end, which demonstrate how developers adopt privacy and responsible AI principles in real-world applications.

Section I

Data Protection



Data Protection

Data protection laws are designed to safeguard individual privacy and regulate the authorized use of data. The Indian Constitution recognises the right to privacy[2] as a part of the right to life and personal liberty. This includes the right to information privacy which allows an individual to control how their data is used and disclosed.[3]

India enacted the Digital Personal Data Protection Act, 2023 (DPDP Act/ Act)[4] to govern how companies collect and process individuals' data. This law requires careful compliance, even if companies seek to use data purely for business purposes without the intent to cause harm. Non-compliance could result in hefty financial penalties going up to INR 250 crores.[5] The government published rules under the Act in November 2025 - the Digital Personal Data Protection Rules 2025 (Rules),[6] which set out specifics of implementation on certain aspects. Substantive provisions of the law will take effect in May 2027- giving companies a runway of 18 months to comply.

Globally, there is increasing regulatory attention on the use of individuals' data in AI systems. Data protection regulators, such as those in the UK,[7] Netherlands[8], Germany[9], Singapore[10], France[11], and others have issued guidance on the application of data protection laws to AI systems. There have also been an increasing number of enforcement actions involving the use of personal data in AI systems, particularly in the EU. Some notable themes emerging from these global developments include ensuring transparency in data collection and processing for AI, minimising collection and use of personal data, and securing effective consent when using individuals' data to train AI models.

Over the next few sections, we explain India's DPDP Act and what companies must do to comply. We discuss the scope of the law, notable exceptions such as for personal data that is made "publicly available", how to provide notice and get individuals' consent, and other organisational measures that AI companies must adopt to comply with the law. Where relevant, we draw from global regimes for interpretative guidance. While this handbook is primarily for developers, we discuss application of the Act to both- AI model development (development, testing, validating) and deployment in user-facing contexts.

Summary of

Law and Key concepts

1. Scope

The Act covers personal data[12], i.e., data about an individual that can identify them. This includes identifiers like name, phone number, email address, postal address and Aadhaar number (i.e. national ID). It also includes profiling data or usage data, for example, a user's preferences. It only covers 'digital' data, not offline records unless they are digitised. It does not cover non-personal data (business insights, anonymized data). It doesn't apply to data that is made or caused to be made "publicly available" by the individual or any other person under a legal obligation to do so.[13] For example, a blogger posts about her spending habits on social media. This exception creates some room to use data that is made publicly available by an individual on the internet for training of AI/ML models. However, the scope of this exception must be carefully considered as it is only personal data that is made or caused to be made publicly available by an individual that is exempt, and not all data available on the Internet. This position is also reflected in Ministry of Electronics and Information Technology's (MeitY) India AI Governance Guidelines, which highlight that the scope of the "publicly available" data exemption under the DPDP Act remains unclear when applied to AI training. The Guidelines note open questions around purpose limitation, consent, and whether research or legitimate-use exceptions can support AI development, and suggest that further guidance or even legislative changes may be required.[14]

2. Who does the law apply to?

Anyone who processes digital personal data will be impacted, barring some exceptions. Processing means collecting, recording, structuring, storing, sharing, or any other automated action on the data.[15] The data could be processed in India or abroad. If data is processed abroad, the law will apply if it relates to "offering" goods and services in India. So, *if offshore businesses offer goods or services in India, the law applies to them.*[16]

The law recognises two entities – data fiduciaries and data processors.

Data Fiduciary (DF)

Entity determining purpose and means of data-processing. Known as data controllers in other parts of the world.

Data Processor (DP)

Entity using and processing the Personal Data on behalf of DF.

Data fiduciaries: Businesses that define "purpose and means" of processing. They are also called data controllers in other parts of the world. These are businesses that determine why user data is needed, how it is used, how long it is to be retained, etc.[17] They are responsible for the data and assume responsibility under the law. For example, an e-commerce platform that collects customer data to fulfil orders and provide personalized recommendations is considered a fiduciary, as it determines the purpose and method of processing the data. Similarly, a healthcare provider that decides how patient data is collected, stored, and shared for diagnostic purposes is also a fiduciary, given its control over data use and management.

Data processors: Businesses that process data on behalf of fiduciaries.[18] For example, cloud service providers who host data for their customers or 'know-your-customer' (KYC) service providers who conduct users' KYC on behalf of a payments company. Fiduciaries tell them what to do.

3. How should data fiduciaries collect personal data?

Fiduciaries must either get an individual's consent or the collection/ processing must be for certain "legitimate uses" recognised in the law.[19] To be clear, this is not required for publicly available data exempt under the law.

Consent: Fiduciaries must give users a notice describing what data is collected, for what purpose, users' rights, and how they can complain to the enforcing authority – the Data Protection Board (or "Board"). Fiduciaries must give users the option to access the notice in English and local languages (recognised in the Indian Constitution).[20] The Rules also require that the notice must be understandable independently of other information given to users, and that it must contain a fair account of the details necessary for processing, such as data, the specified purpose/ purposes of processing, services/ uses enabled by the processing, indicating that some level of detailing is required in the notice.[21] On reading this notice, individuals must give clear and affirmative consent confirming that their data can be processed for the specified purpose.[22] They must also allow individuals to withdraw their consent.[23]

For data collected before the law kicks in, fiduciaries must send individuals a fresh notice, which sets out what data is processed, purpose, how individuals can exercise their rights and make complaints to the Data Protection Board (DPB/ Board).[24] For instance, an e-commerce platform may have previously collected customers' names, delivery addresses, and purchase history to fulfill orders and offer personalized product recommendations. Once the DPDP Act comes into force, the platform must issue a fresh notice to these users, explaining how their data is used and informing them of their rights under the new law.

Legitimate uses: If fiduciaries process data for certain "legitimate uses" recognised in law, they do not need to obtain user consent separately. This includes situations where the individual voluntarily provides her data for a specific purpose; or data is processed to meet legal obligations or to comply with a court order, among other things.[25] For instance, if a court orders a company to provide certain user data as part of an investigation, the company can process and share this data without obtaining the user's consent, as it is for compliance with a court order.

The law also recognises some circumstances (exemptions), where the law does not apply. The exemptions under the DPDP Act are primarily bound to the purpose or actions for which the data is being processed, rather than the type of data fiduciary involved. This means that both private and public entities can invoke these exemptions if their processing activities align with the specific purposes recognized under the law. This includes processing data to detect or prevent an offence, for enforcing a legal right or claim, among others.[26]

That said, there are certain situations under the DPDP Act where exemptions are likely to apply only to government entities or public authorities. These exemptions are typically tied to functions that are inherently governmental in nature, such as national security, public order, and certain regulatory or sovereign functions.[27]

4. What else should fiduciaries do?

a. Implement organisational and technical measures;[28]

b. adopt reasonable security safeguards;[29]

c. notify personal data breaches to the Data Protection Board and affected individuals;[30]

d. ensure accuracy, completeness, and consistency of the personal data, in certain situations;

e. erase personal data once the purpose is met or if the individual withdraws consent;[31]

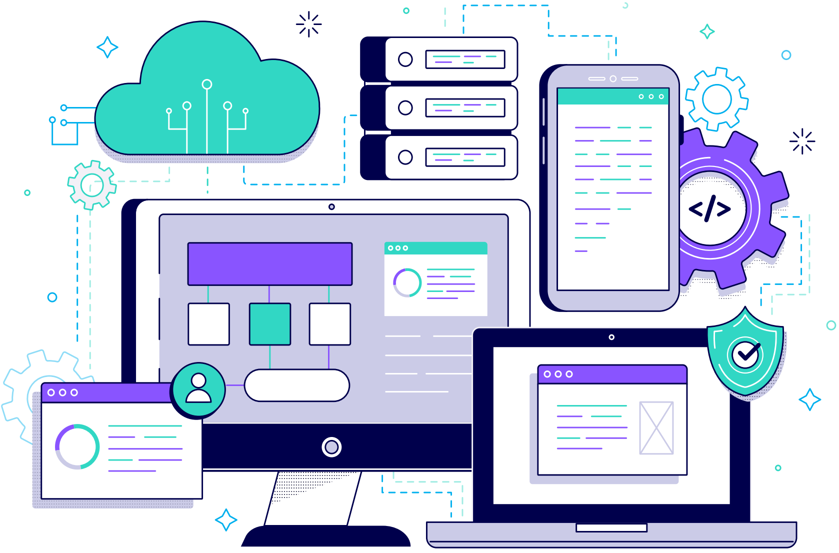
f. implement a mechanism to resolve grievances[32] and enable Data Principals to exercise their rights under the Act;[33]

g. appoint vendors only under a contract that describes how they'll use and protect the data, among other things;

h. publish the contact details on their website or app of a Data Protection Officer or designated person to answer user queries on processing of their personal data[34] and clearly outline the process for users to exercise their rights.[35] (*under the Rules*)

Fiduciaries that process large volumes of data or sensitive data could be designated as "significant data fiduciaries" (SDFs) by the government.[36] SDFs must: (a) appoint a data protection officer based in India;[37] (b) appoint an independent data auditor and do periodic data audits;[38] (c) carry out periodic data protection impact assessments[39]; (d) ensure due diligence in deploying algorithmic software to mitigate risks to data principals' rights.[40] Notably, the major findings from the data audits and data protection impact assessments must be reported to the Board.[41]

Processing Persons with Disabilities (PwDs) & children's data: Companies that collect the data of PwDs and children must get their parent/ guardian's consent.[42] They also cannot track, monitor a child's behaviour, or serve targeted ads directed to children.[43] The central government can provide exemptions to comply with these obligations.



5. What should data processors do?

The law does not spell out specific obligations for data processors or penalties for them. Fiduciaries may pass these on to processors through contracts.[44] So, processors must review their contracts with fiduciaries closely.

Indicatively, fiduciaries will seek clauses such as:

Purpose limitation: Processors must not process data beyond the purpose of the agreement and the agreement generally will set out the rights and obligations of the data fiduciary and processor.

Security safeguards: Processors must implement appropriate security safeguards relevant for data fiduciary's purpose of processing. Processors typically allow fiduciaries access to security documentation/ certifications and audits to verify compliance, with conditions on the information to be furnished/ details of audit typically included in the agreement.

Sub-processors: Clauses that bring clarity to fiduciaries on sub-processor arrangements of the processors and the continued responsibility of processors.

Indemnity: Fiduciaries will seek to be indemnified for any data breaches at the processor's end. Such clauses are also usually heavily negotiated, since processors would want to limit the extent of liability under their agreements.

6. Can companies transfer/process data outside India?

Yes, but the Indian government can restrict transfers to certain countries through notifications.[45] The Rules state that even when personal data is sent abroad, businesses may need to meet certain conditions set by the government, especially when dealing with foreign governments or government-controlled entities.[46] Specifically, for SDFs, the Rules state that a government-constituted committee has the power to recommend types of personal data and associated traffic data that cannot be transferred outside India.[47] This is separate from sector-specific directions on local storage of data/ restrictions on cross-border data flows, such as the Reserve Bank of India's (RBI) direction to payment businesses mandating storage of payment data on Indian servers.[48] Therefore, fiduciaries must evaluate whether any sector-specific obligations impose local storage requirements for their use-case.

7. What rights do individuals have over their personal data?

Individuals can ask fiduciaries to give them information on the personal data being processed, processing activities, and identities of all organizations with whom their data has been shared.[49] They can also ask for their information to be corrected/erased[50] - which can be challenging in the context of AI systems. They can nominate someone else to exercise their rights on their behalf in case they die or are incapacitated.[51] Companies should allow individuals to easily access grievance redressal mechanisms.[52] The law also places duties on individuals, such as, not making false or frivolous claims, not impersonating another person, among other things.[53]

8. What happens if companies do not comply?

The Act sets up the DPB to enforce the law and hand out penalties.[54] Individuals can approach the Board if a data fiduciary doesn't comply with the law.[55] The Board can award penalties up to INR 250 crore (USD 30 million) for some breaches. For example, penalties can be levied for failing to secure personal data, resulting in a breach, for processing data without obtaining proper consent from individuals, for not adhering to additional obligations set out in the law for processing children's data or for failing to observe the additional obligations applicable to SDFs.[56] There is no criminal liability. In awarding penalties, the Board will assess any steps the company took to mitigate the impact of the breach or non-compliance.[57] Notably, the Board can also ask the government to issue directions to block access to a fiduciary's platform in certain cases.[58] Complaints to the board can also be resolved through mediation,[59] or with the fiduciary committing to voluntary undertakings to rectify non-compliance with the law.[60]

Personal Data
Use for training AI models

Data is used to create code, which 'learns' from data patterns and makes calculated predictions or decisions. Larger datasets are able to provide more inputs to the AI model to learn and generate responses.[61] So, a large language model (LLM), which is trained on more data,[62] will have more relevant user-specific examples, complex patterns, and relationships to learn from. For example, in traditional deterministic AI systems, such as fraud detection algorithms in banking, large datasets of historical transactions are essential for the model to recognize patterns of legitimate and fraudulent behavior.[63]

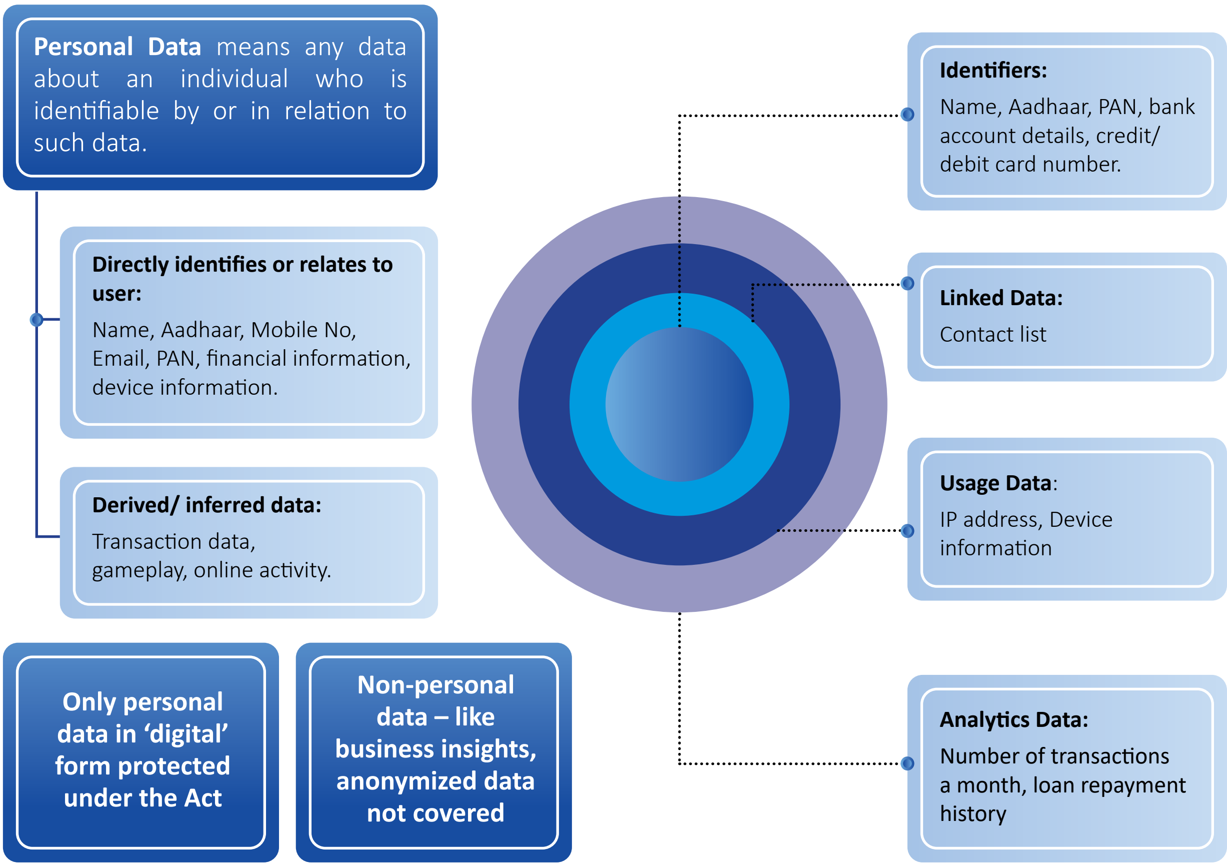
Similarly, generative AI applications, such as ChatGPT and DALL-E, are trained on massive scale data, allowing these systems to understand complex patterns, relationships, and user-specific contexts to generate more relevant responses or outputs.[64] Regardless of the type, the scale and quality of data directly influence the effectiveness of AI systems.

Developers may collect data from various sources – scraping data from the Internet, government/ public databases, deployers (in specific contexts), end users (if deployed in a consumer-facing application), third party platforms through data license arrangements/ platform APIs, other data providers, and so on. The type of data required, and source chosen, may vary depending on the stage of model development.

- **Training phase** - Developers need large and diverse datasets to build a strong and effective model;[65]
- **Testing phase -** Separate datasets are selected that simulate real-world scenarios to assess the model's performance accurately;[66]
- **Validation phase -** Data not included in the training set is used to fine-tune the model and enhance its accuracy.

Against this backdrop, developers must first understand the scope of the Act – since it is concerned only with "personal data". Developers should evaluate whether they need personal data for their AI models, and if not, identify ways to minimise collection of personal data at source.

Scope of "personal data" under the DPDP Act[64]



Personal Data means any data about an individual who is identifiable by or in relation to such data.

Directly identifies or relates to user:
Name, Aadhaar, Mobile No, Email, PAN, financial information, device information.

Derived/ inferred data:
Transaction data, gameplay, online activity.

Only personal data in 'digital' form protected under the Act

Non-personal data – like business insights, anonymized data not covered

Identifiers:
Name, Aadhaar, PAN, bank account details, credit/ debit card number.

Linked Data:
Contact list

Usage Data:
IP address, Device information

Analytics Data:
Number of transactions a month, loan repayment history

The DPDP Act is concerned only with digital "personal data", i.e. data about an individual who is identifiable by or in relation to such data.[67] It does not extend to non-personal or anonymised data, i.e. data that does not relate to or identify an individual.

For data to be considered personal data, it must:

- be about an individual, i.e. a natural person;
- directly or indirectly identify the individual.

Direct identification: means direct references to or identification of a person. Example: their name or their phone number or photograph or unique government identification number.[68]

Indirect identification: likely to mean when individual pieces of data do not directly identify an individual on their own, but other information (which may either already be with the fiduciary or can be reasonably accessed from another source) can contribute to revealing their identity.[69] Drawing from global regimes, an individual can be "indirectly" identified from a dataset when datasets are:

- ***Combined with additional data:*** Example: car registration number, age group related data, geographical location — can be co-related or linked to master databases, to identify individuals.[70]

- ***Analyzed using advanced methods***: These include data aggregation (synthesizing in a larger dataset in summary form)[71], cross-referencing (comparing information and finding correlations between different datasets) and inferences (drawing conclusions and deriving insights based on the information collected).

Examples of personal data

- Personal details like name, date of birth, gender, marital status, religion.
- Contact details like postal address, phone number, IP address, email address.
- Biometric details like retinal scan, fingerprint.
- Unique identification numbers like Aadhaar, passport, driving license number, PAN details.
- Media like voice recording, videos, images, CCTV footage.
- Financial data like bank account numbers, credit card numbers, transaction history.
- Health information like medical records, health insurance, genetic information.
- Employment data like salary details.
- Educational details like academic transcripts, student ID numbers, enrolment records.
- An identified user's interactions with a service for instance usage patterns, preferences, chat history, prompt history, etc.
- Inferences such as person X is likely to respond to a promotional offer on a Sunday evening or person Y is interested in luxury handbags.

In determining whether a dataset or an attribute is personal data, context is key. While the name 'Rahul' by itself may not identify a person, they may be identified with additional information like their job title, location, and name of company. There may also be special circumstances — while Rahul's occupation alone may not be considered personal data (since job titles are typically not unique), it can still help narrow down and identify a person. Especially, in situations of a one-person company (sole proprietorship) or when the job title in question is of a founder, identification of individuals linked with such companies may be easier.

For instance, a company conducts a survey and collects details such as age, gender, occupation, and place of work from respondents. Each attribute alone may not identify a person, but when combined, they can. For instance, the company collects data from respondent A, a female in her 20s working in marketing in Connaught Place. This combination is common and may not identify her. In contrast, respondent B, a male security officer in his 20s at a specific office in Nehru Place, could be identified because his combination of attributes is rare. In cases where individuals can be identified from the dataset, the company should consider treating it as personal data.[72]

The scope of personal data is wide. Personal data includes data in any form - video, audio, text, image, documents.[73] It could also cover subjective information such as opinions - taking cue from global regimes – as long as it relates to an identifiable individual. Example: employment evaluations or a drawing of someone's family made as part of a psychiatric evaluation may also be personal data if it relates to an individual. Essentially, any information that relates to an identifiable individual, whether objective or subjective, can be considered personal data.

Not in scope: Non-personal or anonymised data

The Act covers only personal data; it does not govern data that cannot identify or trace back to an individual or "non-personal data". While the law does not define non-personal data, it is understood to be of two types:[74]

- **Data that was always non-personal**: This is data that at no point was related to any identifiable person. For example:[75] soil data, climate conditions or weather patterns, aggregate of number of cabs on the road in an Indian city.
- **Data that used to be personal data but has been anonymised**: This data originally was linked to a person; who, however, is no longer identifiable since all identifiers have been removed. This process of removing identifiers from a personal dataset is called anonymization. Data that cannot be linked back to a person, or has been fully anonymized, does not fall under data protection laws.

Examples of non-personal data[76]

- Business information like total sales figures, revenue, production volumes.
- Performance data like error rate percentage or usage statistics of a product.
- Aggregated statistical data depicting broad trends like average temperatures for a city, total number of website visitors.
- E-commerce data like conversion rates, attributes.
- Raw data like readings from sensors tracking air quality or temperature.
- Anonymous feedback or reviews like comments or e-commerce product reviews.
- Inferences such as users in a residential area are more likely to respond to a marketing notification on a Sunday evening.
- Aggregate purchase data of a retail store.

Anonymisation process

The DPDP Act does not refer to, or offer guidance on, anonymisation. Earlier drafts of the law defined anonymisation and required "irreversibility" for data to be considered anonymised.[77] However, the DPDP Act avoids making such references. Standards for anonymisation may evolve through market practice and enforcement actions. For reference, Singapore's data protection regulator recognises techniques such as de-identification, record suppression, character masking, generalisation, swapping, data perturbation, k-anonymisation, differential privacy, and data aggregation.[78] These methods help ensure data anonymity and prevent re-identification. The United Kingdom Information Commissioner's Office (UK ICO) sets out a threshold for when data is considered anonymised – when a "motivated intruder", using public resources and investigative techniques without any prior knowledge, cannot re-identify individuals from the data. This test helps a company evaluate whether the data is effectively anonymised.[79]

Anonymisation may not always be fool-proof. Research increasingly shows de-identified data can be re-identified.[80] The ability to re-identify individuals from an anonymised dataset depends on factors such as: the nature of the original dataset, the advanced methods used, the skill and resources of potential attackers, and the availability of additional data that could be linked to the de-identified information. For instance, when anonymized data from New York taxi rides was released — showing origins, destinations, times, and payments, but with passenger info omitted and taxi IDs hashed — it was initially thought to be anonymous. However, hashed IDs were easily decoded, and photos of celebrities in taxis published by Google revealed the taxi IDs. By linking these photos to the decoded data, the destinations and payments of many celebrities were exposed.[81]

Similarly, Netflix released a dataset of 100 million anonymous movie ratings[82], offering a prize of USD 1 million to the developer community for improving its recommendation algorithm. However, despite Netflix's efforts to anonymise this data, researchers from the University of Texas re-identified most users by cross-referencing with publicly available movie ratings on IMDb.[83] This showed the risks of re-identification even with advanced anonymisation techniques.[84]

While absolute anonymity might not always be possible, the data must be protected so that the risk of re-identification is very low. This means the anonymisation techniques used should make it highly unlikely that someone could successfully uncover identities. This involves two steps: de-identification, and identifying and containing re-identification risks.

Pseudonymisation

Pseudonymisation involves replacing identifiers with fake values.[85]

A common approach is to pre-generate a list of fake values and randomly select from this list to replace the original data.[86] Pseudonymisation protects privacy while keeping the data useful for analysis or other purposes. Pseudonymisation would not automatically mean that the data is not personal data; in the EU, the test for assessing whether data is personal or not is still whether the entity can identify an individual using all reasonable means available.[87]

Question for developers: Evaluate whether the data you process is personal data

- ***Does the data relate to an individual?***: Check if there are any direct identifiers, which can identify an individual.
- ***Will it relate to an individual if combined with any additional information?***: Check if with some additional information, there is an identifier — which helps in tracing back to the individual.
- ***Will it relate to an individual if advanced methods are applied?***: Check if with the application of any advanced method – like aggregation, cross-referencing, inference, etc, the information can be traced back to the individual. Example: If an online platform releases statistical data about its services' usage (which does not include users' personal identifiers), the usage patterns and public comments made by users can be cross-referenced. This can identify individuals and fall under the ambit of personal data.

DPDP Act extends only to personal data

Consider a developer who is creating a model for an AI-based agri-support tool that aims to identify appropriate interventions for farmers. The developer collects data directly from farmers across different regions of Northern India, which includes identifiers like names, locations or contact details. In this pre-processing stage of the AI model, the collected data is classified as personal data, since it can be used to link to a particular individual, and therefore, is subject to the rights conferred by the DPDP Act.

However, before the data is fed into the AI model for training, the developer anonymises the dataset, stripping it of identifiers that could link it back to individual farmers. This data has now been transformed into non-personal data. This is the processing stage for the AI model, where it processes only anonymized data to generate insights. Since the data no longer qualifies as personal data, the DPDP Act will not apply to this dataset.

```
private function _build()
{
   $deploy = $this->_get_deploy();
   // get domains
   $domains = array();
   foreach ($deploy->{'domains'} as $d) {
      $domain = Domain::fromJson($d);
      $domains[$domain->getId()] = $domain;
      Logger::debug("load domain %s", $domain
   }
   // get layers
   foreach ($deploy->{'layers'} as $l) {
      $domain = $domains[$l->{'domain_
      $layer = new Layer($l->{'id'}, $dom
      $this->_layers[$l->{'id'}] = $layer;
```

```
if ($prop     e'] insta      Time) {
   $tim     $props['ti     ];
} else
   $time = Time::constructFromArray($props['time']);
}
if ($props['clef'] instanceof Clef) {
   $clef = $props['clef'];
} else {
   $clef = Clef::constructFromArray($props['clef']);
}
if ($props['key'] instanceof Key) {
   $key = $props['key'];
    lse {
    key = Key::constructFromArray($props['key']);

        s = $props['divisions'];
        ps['barline'] instanceof Barline) {
        ine = $props['barline'];
  else {
   $barline = Barline::constructFromArray
```

Risks with using personal data

When a fiduciary collects and uses personal data to develop an AI model, the collection and processing of the data is governed by the DPDP Act. Consequently, the fiduciary must meet the various requirements of the DPDP Act. These include providing individuals with a notice about data collection, obtaining their consent, and allowing them rights over their data. Fiduciaries must also notify individuals in case of a breach. However, these obligations may not apply if the processing or dataset is exempt under the DPDP Act (more on exemptions, including for publicly available data below). Retaining personal data also entails other risks- personal datasets can be prime targets for cyberattacks.[88] Data breaches involving individuals' data can expose a company to liability under the law and cause significant reputational harm.[89]

Question for developers: Evaluate whether you need personal data

Given these risks, where possible, developers could consider whether they even need personal data – if the purpose can be met by using anonymised data. For instance, to develop an AI-based radiology assistance tool, the developer needs CT scans of lungs of individuals with cancerous nodes. The model does not need identifiable information about an individual, only the relationship between the input (the scan) and the output (diagnosis). When sourcing such data from radiology labs, the fiduciary could require the lab to only provide anonymised data with patient details redacted.

However, it may not always be feasible to avoid use of personal data for training the AI model. The context may require the use of identifiable data. For instance, a wearables company intends to develop and provide a new functionality in its health tracking mobile application to give timely reminders based on changes in users' vital signs. The company uses personal data, like heart rate and step count, to train its machine learning model. While anonymised data could be used for general model training, personal data may be required for personalizing the reminders for each user.[90]

Also, even where they do not need personal data to train the model, developers may end up collecting personal data- as part of a dataset that they need. For instance, if you are training your AI model to simplify tax reporting and collect tax records from various companies, these records might contain personal data such as names, job titles, contact information, and addresses of employees and directors; in addition to, non-personal data like asset and revenue details. These portions are 'inextricably linked'[91] to the non-personal data components. A mixed dataset will attract DPDP obligations unless it is a public dataset exempt from the Act.

In scenarios where fiduciaries collect personal data or mixed datasets, they should consider anonymising the data before further use in training the AI model – since use of anonymised data is not governed by the Act. Therefore, while the underlying raw dataset collected from any source is personal, the subsequent anonymised data does not attract DPDP compliance obligations.

To anonymise, developers must:

- Identify and apply the appropriate technique for de-identification.
- Identify risks of re-identification and manage risks.

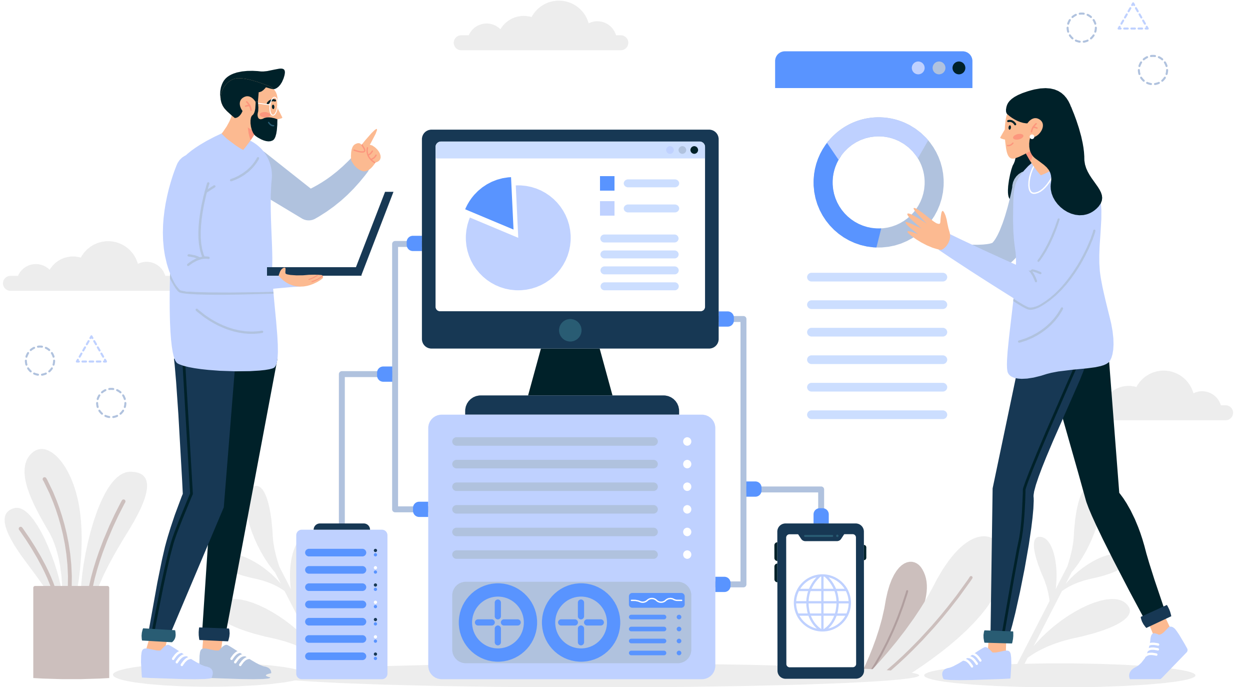
For more detailed guidance on techniques and process of anonymisation, refer to Singapore's data protection regulator's "Guide to basic anonymisation".[92]

Chapter Summary

- AI learns by analysing vast amounts of data, recognizing patterns, and making connections to generate relevant outputs. To train, test, and refine their performance, AI models are trained on data from a variety of sources; web scraping, public databases, user inputs, and licensed datasets.
- Under the DPDP Act, only "personal data" falls under regulation. This includes any digital information that can identify a person, whether directly (like names and phone numbers) or indirectly (through linked datasets). Non-personal and anonymized data are excluded.
- Sometimes, collecting personal data is unavoidable, especially when working with mixed datasets that contain both personal and non-personal information. In such cases, organizations should take a layered approach: separating personal data where possible, using anonymization or pseudonymization techniques, and ensuring compliance with legal requirements.

Checklist

1. **Decide whether you need personal data** (This is to minimise risk exposure and scope of data that is regulated).
   - Assess if you need personal data in the first place.
   - Check if the same result can be achieved using non-personal or anonymised data.
2. **Implement processes to limit identification of individuals**
   - If possible, anonymise data at source of collection.
   - Where technically feasible, implement filters to screen out identifiers or other personal data before data is fed into the AI model.
   - Document this in your agreements with data providers — meaning seek representations from your data providers that they will only provide you anonymised datasets.
3. **Exercise caution when using anonymised datasets**
   - Evaluate risks of re-identification when using anonymised data.
   - Explore different methods for achieving optimal anonymization.



Data Sources

Data is required throughout the lifecycle of AI model development, encompassing[93] stages such as training, validating, testing, operation, and enhancement, and subsequently in deployment. Across these stages, developers may collect data from different sources. Sources could include – scraping data from the Internet, government/ public databases, deployers (in specific contexts), end users (if deployed in a consumer-facing application), third-party platforms through license arrangements/ platform APIs, other data providers, and so on.

Sourcing publicly available data

The DPDP Act does not extend to the processing of certain types of publicly available data.[94] This is a useful exception for AI developers looking to source data from public sources, as access to vast and diverse datasets is foundational for achieving model quality and functionality However, the scope of the exemption will evolve through enforcement and guidance from the regulator/ government and therefore developers should not assume that any personal data that is available publicly is covered by the exemption.

The exemption extends to:

(a) Data that an individual herself has made public or caused to be made public, for instance, personal information posted on a public blog; and

(b) Data that is made public under a legal obligation.

Singapore's law also has an exception for publicly available data.[95] However, it is narrower than India's DPDP Act because it only provides an exemption from obtaining consent, but all other obligations under Singapore's law still apply to the dataset. At the same time, it is wider than India's law - since it extends to any data that is "generally made publicly available".[96]

There are two parts to the exception in the Indian law: one where the individual "makes [the data] publicly available", and another where the individual "causes [the data] to be made publicly available."

The first part is relatively straightforward - it covers situations where the individual directly publishes their personal data, such as by posting a blog, commenting on a forum, or including information on a public social media profile.

The second part is more nuanced. One interpretation is that it applies when the individual intentionally instructs or authorizes a third party (e.g., a platform or service) to make the data public — for example, by choosing public visibility settings when uploading content.

Another possible reading is broader: it could apply when the data is publicly accessible, the individual is aware of its availability, and has not taken steps to remove it - thereby effectively "causing" it to remain public.

Since the scope of this exemption is likely to be a contentious issue, developers should adopt a considered interpretation of the provision, apply it consistently, and document their reasoning. The government's India AI Governance Guidelines acknowledge that this position remains unresolved[97], and therefore developers should remain alert to regulatory guidance or enforcement actions that may clarify or narrow the exemption, and be prepared to adjust their practices accordingly.

Personal data that is made publicly available under a law is also exempt. Examples could include: court records, First Information Report (FIR) registries, land records made public by state land revenue departments. Government websites that make data available on payment of fee are also likely to be covered in this exemption- taking cue from Singapore regulatory guidance, which notes that where a database is made accessible to the public, the personal data contained in such a database would "generally be considered publicly available, even if a nominal fee is payable in order to access the data."[98]

While the DPDP Act may exempt publicly available data, businesses should not assume that they are free from all legal requirements. Other Indian laws and regulations, such as the IT Act 2000, sector-specific regulations, and CERT-In guidelines, may still impose obligations on organizations to protect data—regardless of whether it is publicly available. For instance, these laws may require businesses to implement security measures, prevent unauthorized access, or report breaches. Therefore, even when processing publicly available data, companies should assess their broader legal responsibilities and adopt appropriate safeguards to mitigate potential risks. In any case, if scraping data, companies must evaluate other risks associated with it, such as, breach of platform terms or IP infringement.

Personal data from other sources: In scope

If personal data is sourced from certain public sources, it is exempt. But, to collect and use personal data from any other source, the fiduciary must comply with the requirements of the Act.

Let's consider an example:

A developer is creating an AI model that can aid healthcare professionals identify suitable mental health management interventions for their patients. The developer identifies the following sources:

- A public community of individuals that discuss their counselling experiences – on a social media platform.
- A survey among psychologists to share their experiences without any patient names.
- A mobile/ web application created by the developer to collect data for this use-case.
- An existing mobile application run by the same developer which allows individuals to journal, connects them to mental health professionals, set up calls, etc.

In the first scenario, the fiduciary should assess whether they can seek to use the publicly available data exemption.

For the second scenario, the fiduciary must evaluate whether the dataset has any personal data. To minimise exposure to personal data - when sourcing data from the respondents, it must seek that they refrain from providing any individuals' names or other identifiers that could potentially identify individuals. In both these scenarios, since the company relies on an exception, the Act will not apply to the collection and use of the data, if the conditions set out in the law for each of those exceptions are met.

In the third and fourth scenarios, the data in question is personal data since it relates to an individual who is identifiable. The fiduciary must then comply with the requirements in the Act - the first of these is to identify a legal basis for the processing of such data, i.e. a legitimate use or consent (discussed in the next two chapters).

Takeaway for developers

For each different source, the developer must consider the extent to which the DPDP Act applies, i.e. whether any exception is available. The developer must also consider additional risks, such as breach of platform terms, IP infringement, etc. A summary of positions under the DPDP Act and other risks is below:

Source	DPDP Act
Public datasets – from government websites	Likely to be understood as data made publicly available under a law.
Scraping data from third party platforms	Verify that scraping is not prohibited by platform terms.
Licensing data from third party data providers	Must identify legal basis, i.e. legitimate use or consent. Must require the third party to have taken appropriate consents. Seek indemnities in the agreement with the provider.
Licensing data from third party data providers	Must identify legal basis, i.e. legitimate use or consent. Must require the third party to have taken appropriate consents. Seek indemnities in the agreement with the provider.
Deployers	Must evaluate the relationship between developer and deployer.
End-users	Must identify legal basis, i.e. a legitimate use or consent, and abide by other DPDP compliance requirements.

Chapter Summary

- The DPDP Act provides an exemption for some publicly available data but only in specific cases: (1) data an individual has made public themselves or caused to be made public (e.g., a public blog post) and (2) data made public under legal obligations (e.g., court records).
- For personal data collected from other sources, compliance with the DPDP Act is mandatory. Developers must determine if an exemption applies, consider additional risks like platform policy violations, and ensure compliance with broader legal obligations.

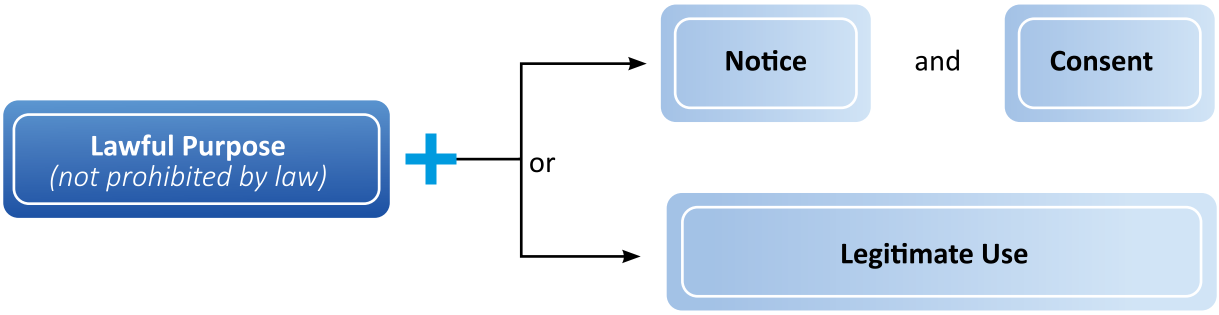
Checklist

Evaluate risks associated with the data source

- If relying on publicly available data, evaluate if it is actually covered under the exemption in the DPDP Act.
    - Identify legal risks associated with each data source.
- If using scraped or third-party data, document provenance and obtain necessary permission.
- Validate that data collection practices align with platform policies and licensing terms.

Notice and Consent

To process any personal data, a fiduciary must identify a legal basis for the processing. This means the fiduciary must either: (i) give individuals a notice and get their consent; or (ii) justify the data processing under one of the nine legitimate uses recognised in the Act.[99] We first discuss notice and consent.



A fiduciary must give users a notice describing the data and purpose of its processing,[100] and seek their consent to the processing of data for that purpose.[101] The form and manner of providing a notice and getting users' consent will depend on the stage of AI development. For instance, if data is collected from a consumer-facing application for re-training or continuous training of an already deployed model, notice and consent can be embedded in the platform's user interface. However, if the data is collected for initial training directly from individuals, the developer must provide notice and obtain explicit consent at the time of collection.

Can data originally collected for another purpose (e.g., tracking app usage) be used for AI training later? If the data was originally collected for a specific purpose, such as tracking the use of an application, it generally cannot be repurposed for AI training without providing users with renewed notice and obtaining consent. This is because repurposing data for AI training constitutes a new and distinct purpose.

Elements of a Notice

A fiduciary must disclose the following details in the notice:[102]

- Personal data to be processed;[103]
- Purpose for processing personal data and fair account of details necessary of the goods, services or uses enabled by such processing;[104]
- Manner in which the individual can exercise her rights to correct, complete, update or erase personal data;
- Manner in which the individual can complain to the Data Protection Board of India.

The Rules set out further details. They provide that the notice must be clear, independent and understandable on its own, without requiring reference to other information provided by the data fiduciary. It should also use clear and plain language to ensure that the user can make an informed and specific decision regarding consent. They also call for a fair account of details necessary of the personal data – such as description of data, purpose or purposes, and services/ uses.[105]

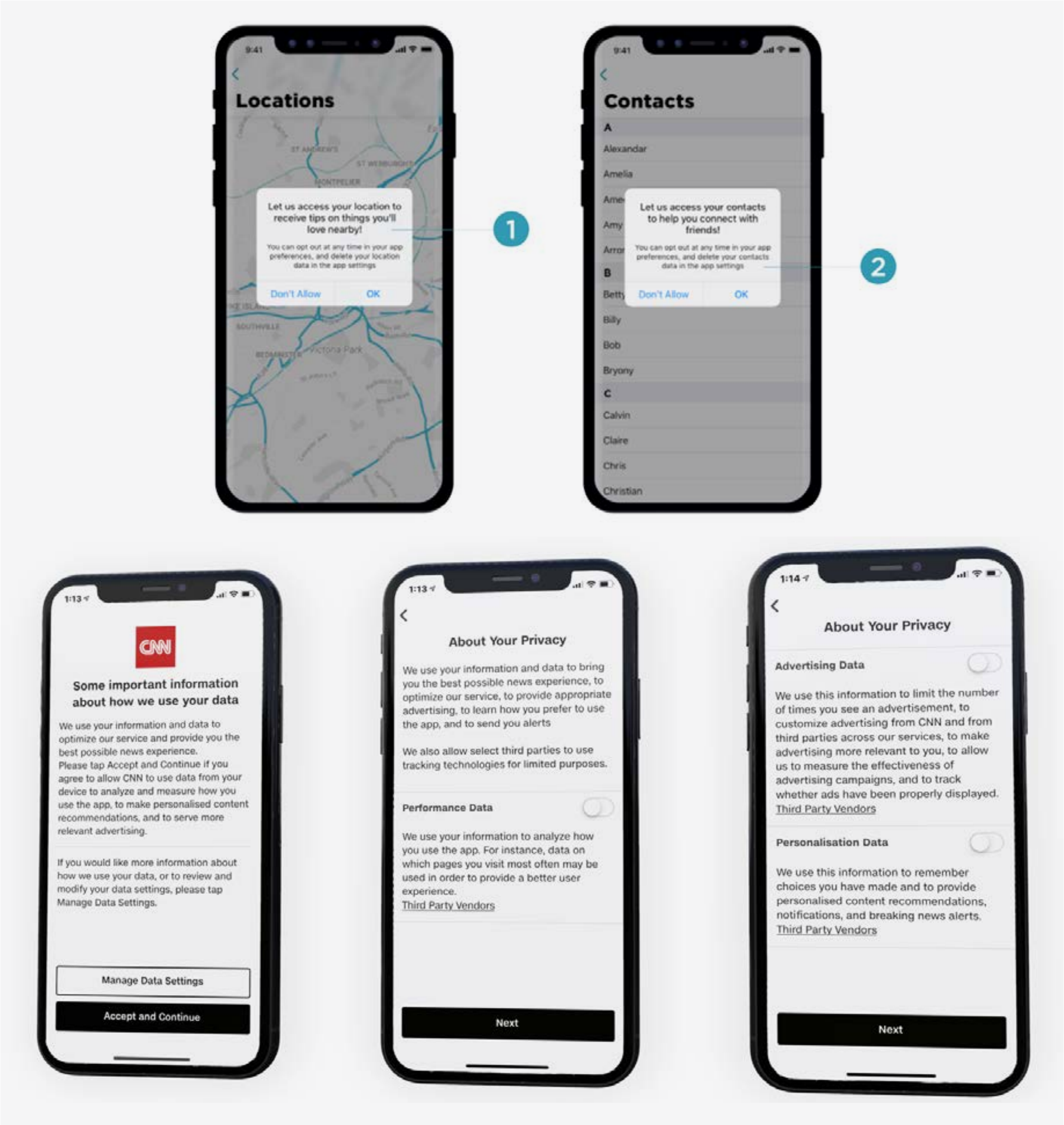
Organisations' privacy notices usually also include information about sources of data, contact details of the company, categories of recipients of the data, the company's retention policy, security safeguards, among others.

Typically, notices are understood as the user-facing privacy policy of an organisation that describes in detail the organisation's data handling practices. While fiduciaries must describe such details in a notice,[106] the requirement of a notice under the Act could also mean providing a more upfront notification or disclosure to the user, possibly through a short form notice. (See figures below).

How We Use Information

What do we do with the information we collect? The short answer is: Provide you with an amazing set of products and services that we relentlessly improve. Here are the ways we do that:

- ***Develop, operate, improve, deliver, maintain, and protect our products and services.***
- ***Send you communications, including by email.*** For example, we may use email to respond to support inquiries or to share information about our products, services, and promotional offers that we think may interest you.
- ***Monitor and analyze trends and usage.***
- ***Personalize our services*** by, among other things, suggesting friends, profile information, or Bitmoji stickers, helping Snapchatters find each other in Snapchat, affiliate and third-party apps and services, or customizing the content we show you, including ads.
- ***Contextualize your experience*** by, among other things, tagging your Memories content using your precise location information (if, of course, you've given us permission to collect that information) and applying other labels based on the content.



Let us access your location to receive tips on things you'll love nearby!

You can opt out at any time in your app preferences, and delete your location data in the app settings

Don't Allow OK

Let us access your contacts to help you connect with friends!

You can opt out at any time in your app preferences, and delete your contacts data in the app settings

Don't Allow OK

CNN

Some important information about how we use your data

We use your information and data to optimize our service and provide you the best possible news experience.
Please tap Accept and Continue if you agree to allow CNN to use data from your device to analyze and measure how you use the app, to make personalised content recommendations, and to serve more relevant advertising.

If you would like more information about how we use your data, or to review and modify your data settings, please tap Manage Data Settings.

Manage Data Settings

Accept and Continue

About Your Privacy

We use your information and data to bring you the best possible news experience, to optimize our service, to provide appropriate advertising, to learn how you prefer to use the app, and to send you alerts

We also allow select third parties to use tracking technologies for limited purposes.

Performance Data ○

We use your information to analyze how you use the app. For instance, data on which pages you visit most often may be used in order to provide a better user experience.
Third Party Vendors

Next

About Your Privacy

Advertising Data ○

We use this information to limit the number of times you see an advertisement, to customize advertising from CNN and from third parties across our services, to make advertising more relevant to you, to allow us to measure the effectiveness of advertising campaigns, and to track whether ads have been properly displayed.
Third Party Vendors

Personalisation Data ○

We use this information to remember choices you have made and to provide personalised content recommendations, notifications, and breaking news alerts.
Third Party Vendors

Next

In recent enforcement actions, such as the Italian data protection agency's direction to OpenAI for ChatGPT, the regulator asked OpenAI to make available on its website: (a). an information notice describing the arrangements and; (b). logic of the data processing required for the operation of ChatGPT, along with other details. It also required that the notice must be easily accessible and placed in such a way as to be read before signing up to the service.[107]

So, it becomes important to give this notice to users at the earliest available opportunity. This will typically be when they are signing up for a service or first interacting with a product.

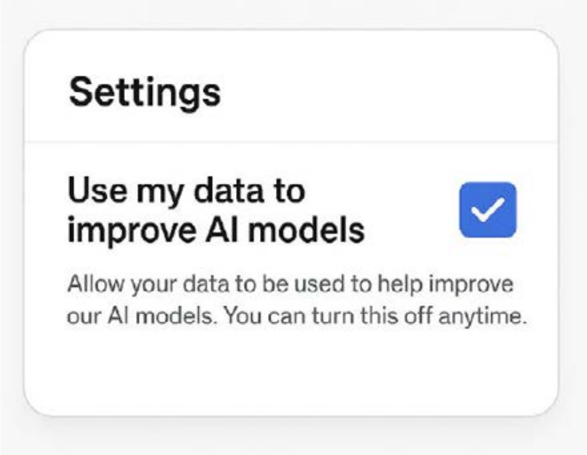
Form of Consent

After providing a notice, the fiduciary must get individuals' consent. Consent must be freely given, specific, informed, unambiguous and expressed through a clear affirmative action.[108] While these terms are not defined in the DPDP Act/ Rules, drawing from global regimes to interpret them:

- **Freely given**: Provided voluntarily without any form of coercion or undue pressure.
- **Specific**: Tailored to specific data processing activities, and not be 'catch-all' clauses. For instance, if a telemedicine app requests consent to process data for its services and to access contacts, but it does not need access to contacts — consent should only cover the services.
- **Informed**: Provide all relevant information required for the data principal to make an informed decision, presumably through the "notice".
- **Unambiguous**: Give options to data principals to exercise and express clear, affirmative action – which unequivocally indicates agreement. Consent should not be inferred.

The requirement of consent being "specific" may be of particular importance in the context of AI systems. The illustration in the Act (regarding a telemedicine app) indicates that consent to collect and use ancillary data (not strictly required for provision of the service) cannot be bundled with the consent to use data that is required to provide the service. So, if the core service can be provided without contact book access, the telemedicine app cannot bundle both consents together - and make the access to the app conditional on the user providing contact book access.

Fiduciaries may evaluate what kind of choices they provide users. For instance, fiduciaries could consider seeking consent for AI training:



Settings

Use my data to improve AI models ☑

Allow your data to be used to help improve our AI models. You can turn this off anytime.

Let's consider an example:

An ed-tech platform is deploying an AI-based anti-cheating solution for proctoring. As a condition for students to take the exam, the company requires them to consent to their data being used to further train and test their AI model. Fiduciaries must evaluate whether consent taken in this manner is "freely given" and valid under the Act.

Fiduciaries must also keep a record of the consents they get, along with time-stamp and other details.

Purpose specification

The DPDP Act requires fiduciaries to specify the purpose of processing in the notice that is shown to users to get their consent.[109] Purpose cannot be an afterthought. For instance, if when the user originally signed up to the service, the organisation did not intend to use the individuals' data for AI training purposes, it may not be able to rely on the original consent for re-using the same data for AI training. In such cases, a fresh notice may need to be provided and fresh consent sought.

Fiduciaries must delete an individual's data, once it no longer serves its original purpose.[110] The law provides some flexibility to businesses in determining this timeline. However, under the Rules, certain businesses such as online gaming platforms (with > 50 L users), and social media/ e-commerce platforms (with over 2 Cr users) must delete personal data after three years of user inactivity.[111] The 3-year timeline kicks in from the last user interaction. Users must be notified 48 hours prior to data deletion.[112] However, businesses can retain data in certain circumstances - for legal compliance, account access or maintaining virtual tokens issued to the user.[113]

Withdrawal of consent

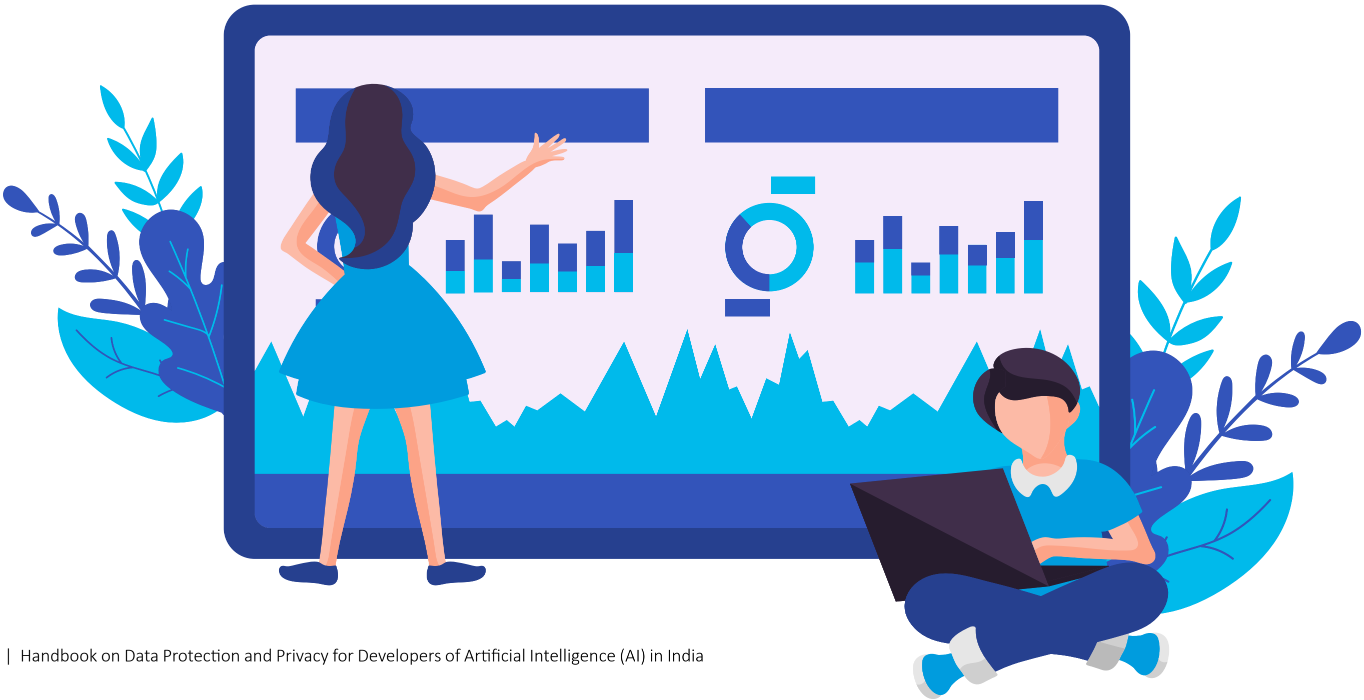
Under the DPDP Act, users have the right to withdraw their consent at any time after it has been given.[114] Importantly, the process for withdrawing consent must be as easy as the process for giving it.[115] This means organisations must allow individuals to revoke consent without unnecessary friction or delay.

Once consent is withdrawn, the fiduciary must stop processing the personal data that was collected or used based on that consent, unless the processing is required or authorised by law. Notably, this withdrawal does not affect the legality of personal data processed when the individual's consent was active.[116] In the context of AI systems, this may require building mechanisms to halt further data use or training. (Refer to the Section 'Individuals' Rights' of this Handbook, specifically the discussion on right to seek correction/ erasure)

Using children's data

Organisations must exercise caution while using children's data in AI systems. The use of children's data, including in the context of AI systems, has garnered attention world over. Notably, in 2022, the US Federal Trade Commission (FTC) directed algorithmic disgorgement as penalty for a company that failed to take adequate parental consent when collecting health and other information of children, in line with the US Children's Online Protection Act (COPPA).[117]

Similar to COPPA, under India's DPDP Act, a fiduciary must get parents' verifiable consent before processing children's data.[118] The Rules provide some guidance here, without being prescriptive on the specific method to be used. Businesses must adopt appropriate technical and organizational measures to obtain parents' verifiable consent before they process children's data. They can use "reliable details" of identity and age of the parent available with the business



or age and identity details provided by parents or tokens issued by government-authorized entities.

Globally, methods for parental consent depend on the context. Some examples include additional email verification, Interactive Voice Response (IVR) calling, requiring a credit card on file, knowledge-based tests, etc.[119]

Organisations that are using children's data for training models must implement appropriate age-gates to ensure they are not picking up children's data accidentally, and must secure parents' verifiable consent in the manner prescribed in the rules. Similarly, when deploying AI systems for children, appropriate notice and consent mechanisms must be put in place.

The DPDP Act further bars organisations from tracking, behaviourally monitoring, or deploying targeted advertisements directed at children, except in scenarios specifically called out in the rules.[120] Unless you fall within one of the listed exceptions, these activities cannot be undertaken even with parental consent. Exceptions include educational institutions may monitor students' data for academic purposes and safety measures, while clinical establishments, healthcare professionals, and mental health practitioners can collect children's personal data to safeguard their health. Additionally, tracking and behavioural monitoring of children is allowed for specific purposes such as managing email communication accounts, preventing exposure to harmful content, and conducting age verification.[121]

Consent Managers

The DPDPA introduces 'consent managers' to facilitate the management of consent through a registered, transparent platform, which allows them to act on behalf of data principals.[122] The Rules lay down some general obligations for consent managers. Consent managers act as a neutral intermediary, helping individuals manage consent for data processing, while ensuring they remain "data blind" (i.e., they cannot read or access the data themselves).[123] Consent managers must meet stringent requirements to operate- they must be registered with the Board, be an India-incorporated company, possess the technical and operational capacity to handle consent management, and maintain a minimum net-worth of INR 2 crore.[124] While the concept has been newly introduced under the DPDP Act, similar frameworks for consent management exist in India's financial sector, such as the Account Aggregator system. Additionally, India's IT Ministry had previously introduced an electronic consent framework for user consent management.[125]

The consent management system assists in managing consent requests — both on scale and specificity. The system is not only capable of dealing with a mass number of consent requests, but also provides a mechanism for granular consent. However, even if a consent management system is utilized, mechanisms for limiting liability and indemnity may still be required.

Chapter Summary

- To process personal data, businesses must have a legal basis for processing (which may be either get clear consent or legitimate uses as per DPDP Act from users or justify the processing under legally recognized grounds). Users need to be given a straightforward, upfront notice explaining what data is being collected, why it's needed, and how they can exercise their rights.
- Consent isn't just a checkbox; it has to be freely given, specific, and unambiguous, meaning users should actively agree to how their data is used. If data was collected for one purpose, like tracking app usage, it can't later be used for AI training without fresh consent.
- Consent managers help streamline this process but don't get access to the data themselves. Special rules apply to children's data - parental consent is mandatory, and AI-driven targeted ads for kids are not permitted. Also, companies must delete personal data once it's no longer needed, unless there's a legal reason to keep it.

Checklist

1. **Design clear and timely notices**
   - Display privacy notice at the earliest user interaction, i.e. during sign-up or first use.
   - Provide clear, concise, easily understandable notice providing all relevant information.
     - Ensure notice includes details of what personal data you hold, why you need it, how users can exercise their rights and complain to the authority.
2. **Collect and manage valid user consent**
   - Ensure consent is:
     - Freely given, specific, informed and unambiguous.
     - Expressed through clear affirmative action.
   - Maintain a verifiable record of all user consent.
3. **Review and improve your consent processes**
   - Establish a protocol to regularly review how consent is requested and recorded.
   - Make necessary updates to data/ consent collection practices and interface design.
4. **Additional considerations when processing children's data**
   - Have valid verifiable parental consent method, before processing children's data.
   - Ensure no tracking or behaviourally monitoring of children, unless covered by an exception.

Identifying a legal basis: Legitimate uses

As discussed above, a fiduciary must process data either on the basis of user consent or for one of the legitimate uses recognised in the Act. These are:

Voluntary provision

Data principals provide their personal data voluntarily for a specified purpose and do not object to its use.[126]

Governance

The state or its agencies process data to provide benefits or services, such as subsidies, licenses, or permits. This is permitted when:

- The data principal has previously consented to such processing.[127]
- The data is stored in a state-maintained database or document and has been digitized.[128]

State function

The State or its agencies process data to perform a function under law or in the interest of integrity or sovereignty of India.[129]

Legal obligations

A fiduciary processes data under a legal obligation to disclose data to a State agency.[130]

Legal compliance

A fiduciary processes data to comply with judgments or orders from any court, whether in India or abroad, related to contractual or civil matters.[131]

Medical emergencies

Data processing is allowed in cases that threaten life or health or during public health crises such as epidemics.[132]

Disaster management

Data can be processed to ensure safety or provide assistance during disasters or breakdowns of public order, as defined under the Disaster Management Act, 2005.[133]

Employment purposes

Data processing for employment purposes is allowed to prevent corporate espionage, safeguard confidentiality, and protect intellectual property.[134]

Of these, the one potentially most relevant for AI companies is the legitimate use of voluntary provision of personal data. Besides this, the law also recognises use of data for State functions, such as providing any benefit/ license/ subsidy or use of data in disaster management or in case of a health epidemic as legitimate uses. AI businesses developing products for government authorities/ public functions could evaluate whether their use-case falls within any of these legitimate uses.

Voluntary provision of data

Under the Act, if an individual voluntarily provides her data to the fiduciary for a specified purpose, and does not object to its use, the fiduciary need not separately take the individual's consent.[135] This provision is drawn from Singapore's law by merging two deemed consent provisions - deemed consent by conduct[136] and deemed consent by notification.[137] To rely on this legitimate use:

- The processing must be for the "specified purpose" for which the individual has voluntarily provided her data. "Specificed purpose" is defined elsewhere in the law as the purpose set out in the notice given to the individual.[138] The individual must "voluntarily" provide her data. This could mean two things: (a) Flowing from the illustration in the Act, the individual herself provides this information, in a way "volunteers" the information without it being solicited. However, this reading would limit the usage of this legitimate use in the context of digital applications, where data is almost always solicited by the platform; or (b) The individual provides the information without coercion.
- The individual does not indicate that she does not consent. Drawing from Singapore's guidance, this would mean the individual is given the ability to object to the use of her data, through an opt-out.

Example: A developer wishes to create a language model for a low resourced regional language. They create a platform for users to donate writing samples in that language. They make it clear on the platform that they do not want users to share personal data in their submissions. In case a user still submits writings which contain their personal data, processing this data can be justified under the ground of "voluntary provision".

Global regimes appear divided on the use of consent and other grounds of processing for using data to train AI models.

- Some regulators, such as the Dutch[139] and the German[140] data protection authorities, note that "legitimate interests" of a business could be relied on for using personal data to train AI models (in the context of the GDPR);
- Others, such as the Canadian Personal Information Protection and Electronic Documents Act (PIPEDA)[141] and the Norwegian Consumer Council[142], insist on consent.

Understanding fiduciary-processor relationships in the AI lifecycle

The DPDP Act recognises two entities:

- ***Data fiduciaries:*** who determine the purpose and means of the processing.[143] These are entities that control the data processing or call the shots. They make key decisions around the data required, the sources, what it will be used for, how long it will be retained.
- ***Data processors***: who process the data on behalf of the fiduciary.[144]

This distinction is recognised world over across data protection laws. For instance, the GDPR defines "controllers" and "processors" nearly identically to India's DPDP Act.[145] Singapore refers to "businesses" and "data intermediaries" or "service providers".[146] The classification as a data controller or data processor is usually a case-by-case assessment of which entity controls the data processing. In most regimes, consequently, the entity that calls the shots is also responsible for compliance.

On similar lines, under India's DPDP Act, fiduciaries are responsible for compliance, including for actions of their processors. Processors face no direct obligations under the law. They are only obliged to follow the instructions of the fiduciary – typically recorded in a contract between the two entities.

In the context of AI systems, often multiple entities are involved in different stages of model development, testing, validating, and then deployment, and re-training, enhancement, or improvement. Depending on the nature of the entity's involvement in the AI lifecycle, or the specific activities undertaken, it will either qualify as a data fiduciary or data processor. The assessment and identification of roles is complex since each of these may qualify as fiduciary or processor based on the level of control they exert over the data processing in the AI lifecycle. They may also make joint decisions around the data, in which case they could also be considered joint fiduciaries. The fiduciary-processor evaluation is a crucial assessment – since all obligations under the law extend to the fiduciary, and the fiduciary must then document its instructions to the processor and pass on certain obligations around data use to a processor through a contract.

An organisation is likely to be a fiduciary when it makes significant decisions about: source and nature of the data used to train the model, model parameters, feature selection, target output, algorithm, ongoing testing and updates, etc.[147] However, even if an entity makes certain technical decisions on the methods or "means" of processing, such as, decisions on formats for data storage, determining programming language, it is likely to be considered a processor, if other significant decisions are made by a different entity.

Let's consider a few examples:

EXAMPLE 1

A developer is developing a base model that could then be used by financial institutions for underwriting. In developing the model, it independently makes decisions around data sources, algorithms, model parameters, among others. It will be a fiduciary for this base training data since no other entity is in the picture at this stage.

A developer develops a bespoke model for a financial institution. The institution directs the developer on purpose of processing, desired output, source of data, etc. The developer takes certain decisions independently on the means of processing, such as the form in which the data will be sourced and stored. Here, the institution is likely to be understood as the fiduciary and the developer as a processor. This is because the institution exercises control over the data processing, even though the developer makes decisions on the method or "means", the institution makes significant decisions on "purpose".

EXAMPLE 2

A video streaming platform contracts a cloud service provider to host and manage its AI recommendation system infrastructure. The streaming platform (deployer) defines the purpose of training the AI system, specifying the use of a dataset originally collected for providing the service and determining the expected outcomes. The cloud provider manages data storage, training efficiency, and resource allocation without changing the system's intended use.

In this case, the streaming platform is the data fiduciary, as it decides the purpose and means of processing, while the cloud provider acts as the data processor, managing technical aspects.[148]

EXAMPLE 3

A hospital uses an AI-based diagnostic tool developed by an external company. The hospital provides patient data to the AI system for processing, and the AI tool helps generate medical diagnoses. The hospital decides how the tool is applied, the patient data being processed, and the purposes for which the results are used (e.g., diagnosing diseases). The developer, on the other hand, determines how the data is technically processed, such as the algorithms used, storage of interim results, and data management during processing.

In this case, the hospital would be the data fiduciary since it determines the purpose of processing (diagnosing patients using AI). The AI tool developer would be the data processor because it handles the technical means of processing patient data, such as how the algorithm works and the infrastructure supporting the processing, but not the primary purpose.

Even if it is a standardised service/ product, where the manner of processing is predetermined by the processor, the hospital in this case makes the choice as to whether to avail the service/ deploy the product. This fact means the hospital makes a determination of purpose and means.[149]

Therefore, the same entity could be a fiduciary for a data processing activity while being a processor for a different activity. The classification is activity specific. For instance, a developer creating a base model may be a fiduciary for the initial training dataset but may be only a processor once the product is deployed by its customer.

EXAMPLE 4

A legal tech platform partners with an AI model provider to offer an AI-powered legal research and drafting assistant. The platform collects user-submitted data (e.g., case facts, legal queries, client information) through its web interface, while the model provider processes these inputs via its API, generating legal responses using its large language model. Both entities jointly determine how user data is used—for instance, to generate outputs, retain logs for performance monitoring, and fine-tuning the underlying model. They also make shared decisions on input formatting, storage duration, and user consent mechanisms.

In this case, both the legal tech platform and the AI model provider may be considered joint data fiduciaries. They jointly determine the purpose (delivering AI-assisted legal services) and the means (collecting and processing user-submitted legal content via the API). As such, both parties are responsible for complying with applicable data protection obligations—such as issuing notices, obtaining valid consent, securing data, and facilitating user rights—within the scope of their respective roles and the shared processing purpose.

Where multiple entities are involved, they must execute a contract clarifying rights and liabilities. The contract between the entities should:

- Identify and document roles, i.e. who is the fiduciary or processor or if the entities are joint fiduciaries;
- Seek appropriate representations and warranties, for instance if a deployer provides its end-users' data for training the AI model, the developer must seek a representation that the deployer has appropriate consents from its end-users for using this data;
- Restrict the use of the data by the data processor to the documented purpose;
- Document responsibilities in case of a data breach;
- Document responsibilities in case an individual seeks to exercise their right to access, update, or erase their data;
- Include indemnities in case of breaches.

Chapter Summary

There is a distinction between data fiduciaries and data processors under the DPDP Act. Think of data fiduciaries as the ones calling the shots; they decide what data to collect, how to use it, and for what purpose. Data processors, on the other hand, just follow instructions and handle the technical side of things. In AI, this gets tricky because multiple players are involved and who's in control can change depending on the stage of development. Sometimes, two entities share control and become joint fiduciaries. Since the fiduciary is legally responsible, contracts between them and processors must clearly define roles, limit data use, and outline responsibilities in case of breaches or user requests.

Checklist

Identification of roles and responsibilities

- Identify third parties you work with where personal data sharing/ receipt is involved.
- Evaluate your role and the third party's role, i.e. whether a fiduciary, processor, or a joint fiduciary
    - For this assessment, assess who calls the shots and makes decisions as to source of data, how it will be used, target output, etc.
- Document roles and responsibilities in a contract with third parties, such as each party's obligations with respect to user rights, breach reporting, etc.

Individuals' rights

The DPDP Act grants individuals certain rights over their personal data, for as long as the organization holds and processes it. The rights extend only to personal data. Therefore, once anonymized to an extent that an individual cannot be identified from the dataset - the dataset no longer attracts the same obligations.

For instance, consider a developer who is creating a model for an AI-based agri-support tool that aims to identify appropriate interventions for farmers. The developer collects data directly from farmers across different regions of Northern India, which includes identifiers like names, locations or contact details. In this pre-processing stage of the AI model, the collected data is classified as personal data, since it can be used to link to a particular individual, and therefore, is subject to the rights conferred by the DPDP Act.

However, before the data is fed into the AI model for training, the developer anonymises the dataset, stripping it of identifiers that could link it back to individual farmers. This data has now been transformed into non-personal data. This is the processing stage for the AI model, where it processes only anonymized data to generate insights. Since the data no longer qualifies as personal data, individual rights under the DPDP Act will not apply to this dataset.

The fiduciary must have mechanisms in place to allow individuals to exercise their rights when the data is still personal. If multiple entities are involved in the data processing pipeline, the fiduciary must ensure that the processor supports compliance with individual rights. The fiduciary is ultimately responsible for ensuring that user rights can be exercised at the appropriate stages of data handling.

Each right is discussed below:

Access information:

An individual can request a summary of personal data being processed and the processing activity.[150] This would mean a description of the personal data that is collected, for instance, health data, purchase history, customer chats, etc. The reference to "processing activity" in the context of AI model training could mean describing how the data will be used, for instance, anonymised at pre-processing stage, and then used to train the model.

India's DPDP does not provide a right to explainability – found in certain global data protection laws.[151] However, explainability is a critical principle from an overall Responsible AI perspective and should be factored in while developing AI solutions. (Refer to Section II of this Handbook for more information)

Individuals can also request for identities of those with whom such data is being shared (along with details of what components of data are shared).[152] For instance, if you are a consumer facing platform and you engage a third-party service provider for data analytics - and you share customer data with this entity - you will be obligated to share details of this entity with the customer.

Seek correction/ erasure:

An individual can seek correction of inaccurate or misleading personal data, completion of incomplete data, and updating of data.[153] An individual can also seek erasure of her personal data, and fiduciary must delete her data, unless its retention is required under any law. Giving effect to rectification/ erasure rights is a tricky exercise and the scope of these rights in the context of AI systems is evolving across the globe.

The UK ICO sets out a helpful framing to understand these rights by breaking down the data into input data (training data used to train an AI model), output data (results), and personal data that is part of the model.

For instance, an individual's purchasing habits are a part of input for a model. The individual seeks rectification of an incorrect data field. At the pre-processing stage, before the data has been used to train the model- the organization must correct the data. For correcting such data, the organization should ask the individual to confirm her accurate details, and if required, ask her to submit supporting documentation to verify.[154]

Once deployed, the AI model shows results about an individual. For instance, from an individual's purchasing habits, she is shown relevant product listings on an e-commerce marketplace. The inferences/ insights about her purchasing habits are output. However, these are likely to be seen as subjective prediction scores rather than statements of fact[155], and the right to rectification may not apply to such output.

Where the model has "learnt" using an individual's data – which she later seeks to rectify or remove – organisations must consider if it is possible to make the model unlearn such data. There is some research ongoing on disgorgement, removal of offending data from a model[156], with the United States' Federal Trade Commission having directed disgorgement of the model on a handful of occasions when it was found to have been trained on illegally obtained data.[157]

There are different approaches to model disgorgement, which offer various trade-offs between utility, computational cost, and the strength of the guarantee provided.[158] For instance, retraining a model can be costly, and technically and economically infeasible, while adding noise to the weights of the AI model, through techniques like differential privacy, can also help minimize privacy concerns by disrupting the model's ability to retain specific data, but may end up affecting model utility.[159]

While acknowledging that machine unlearning techniques are still in their early stages, Commission Nationale Informatique & Libertés (CNIL) proposes alternative methods to address corrections / erasure requests. One such approach is fine-tuning models with new data to minimize the impact of outdated or incorrect information. Another method involves implementing output filters, which would prevent the generation of personal data related to individuals exercising their rights. Rather than creating a blacklist of data subjects who have exercised their rights, CNIL suggests using general rules to detect and pseudonymize personal data in model outputs, thereby reducing the risk of privacy breaches, such as inference-of-membership attacks.[160]

Given the practical difficulties of implementing technical solutions to modify trained models, CNIL emphasizes designing AI systems in a manner that makes it impossible to identify individuals in the training data. This can be achieved through anonymization techniques or methods that prevent the memorization or regurgitation of personal data. Such approaches will eliminate the need for individuals to exercise their rights over the model and reduce privacy risks [161]

Grievance redressal:

If individuals have concerns about how their data is being used, they have the right to raise a complaint.[162] The Rules provide that every organization must appoint a dedicated Point of Contact (POC) to handle these grievances and make their contact details easy to find - whether on their website, app, or in relevant communications.[163] For SDFs, this role must be filled by a Data Protection Officer.

The Rules also state that companies also need to be upfront about how long it will take to address complaints (which should be within 90 days) and put proper systems in place to resolve them within that timeframe.[164]

Nomination:

An individual can nominate another individual to exercise their rights on their behalf, in the event of their death or incapacity — ensuring that personal data rights are protected even in situations where the Data Principal cannot act personally.[165] In this respect, incapacity includes conditions such as mental unsoundness or physical infirmity that may prevent a Data Principal from exercising their rights.

To facilitate this, the Rules provide that individuals must follow the Data Fiduciary's prescribed process, which must be outlined in their terms of service and any applicable laws. The fiduciary must clearly publish the steps and requirements for making a nomination, to make it easier for users to exercise this right.[166]

Differences with other laws

Article 22 of the GDPR grants individuals the right not to be subject to decisions based solely on automated processing, including profiling, that produces legal or similarly significant effects on them.[167]

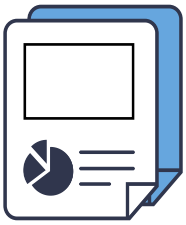
Simply put, when significant decisions, such as credit approvals, hiring decisions or insurance claims, are made by AI systems, individuals can request human intervention, express their views and contest the decision. Article 22 aims to safeguard individuals from potential harm due to AI systems making biased or inaccurate decisions in the absence of human oversight.

The DPDP Act does not include an equivalent to GDPR's Article 22, leaving individuals without a formal right to challenge or appeal fully automated decisions. As a result, organizations are not legally required to involve humans in automated decision-making or provide recourse for affected individuals - though this may be good practice both as a responsible AI measure as well as to avoid implications under future AI regulations in India

In the absence of these formal protections, organizations can adopt responsible AI principles like fairness, accountability and transparency to ensure ethical and fair use of AI. By ensuring that decisions are unbiased (fairness), involve human oversight where necessary (accountability), and are explainable (explainability), organizations can align with global standards and enhance trust. *(Refer to Section II of this Handbook for more information).*

Chapter Summary

- Individuals have key rights over their personal data under the DPDP Act: access, correction, erasure, grievance redressal, and nomination. They can request details on what data is collected, how it's used, and with whom it's shared provided that fulfilling such requests is technically feasible.
- Organizations must ensure clear processes for users to exercise their rights, including grievance redressal mechanisms and allowing nominated representatives to act on their behalf if needed. It is also to be noted that individuals have rights over their personal data for as long as it's processed, but once anonymized, those rights no longer apply.



Checklist

1. **Facilitate transparency and access for users**
   - Provide users with a mechanism to request a summary of their personal data being processed, including details on how the data is used and shared
     - Clearly communicate how personal data is used, including whether it is shared with third parties.
     - Disclose relevant details about processing activities, for example, if data is anonymized during pre-processing before being used for model training.
2. **Support data accuracy and correction pre-training**
   - Allow users to correct inaccurate or incomplete personal data.
   - Establish a process to verify and update user-submitted corrections.
3. **Enable data erasure and post-training remedies**
   - Allow users to correct inaccurate or incomplete personal data.
   - Define and implement strategies to address data correction or deletion requests where technically feasible.
4. **Offer grievance redressal and representation mechanisms**
   - Appoint a dedicated POC for user grievances.
     - Make the POC contact details easily accessible, for example, in privacy notices, on your website, or through customer support.
   - Provide a process for users to nominate a representative to act on their behalf in the event of death or incapacity.

Organisational Measures

Under the DPDP Act, the law casts a general responsibility on data fiduciaries to implement "appropriate technical and organisational measures".[168] Such measures are meant to further the objective of meeting legal obligations under the DPDP Act. Beyond this general obligation, the Rules call on data fiduciaries to implement appropriate technical and organizational measures particularly with regard to facilitating user grievance redressal,[169] obtaining verifiable parental consent when processing children's personal data,[170] and implementing reasonable security safeguards to protect personal data.[171]

Scope of application

While there is no clear definition or defined scope in the law, these terms are broadly understood as:

- ***Organisational measures***: build an overall internal culture of being committed to data protection. This includes implementation of data protection policies, yearly review of the processing activities, and training of employees and management.[172]
- ***Technical measures***: have a direct effect on the operation of technical processing of data.[173] This includes pseudonymization of personal data, encryption, access restrictions, using privacy enhancing or privacy preserving technologies.

Notably, the base requirement is that any of the measures you adopt should help in ensuring effective observance with the DPDP Act.[174]

Test for appropriateness and proportionality

'Appropriateness' of organizational and technical measures that you should adopt is purely contextual. A data fiduciary is not typically expected to implement every available organizational and technical measure. Instead, a comprehensive assessment of processing activities and analysis of underlying risks can be done to select and determine measures.

Taking cue from global regimes, this can be based on:[175]

- ***Nature***: Type of processing activity — like collection, recording, storage, organisation, etc.
- ***Scope***: How much data, whose data, for how long and the geographical or territorial scope of the data being processed.
- ***Context and purpose***: Why the processing is happening.

The measure adopted must be proportionate to the risk posed by the processing operation. For instance, an AI-based facial recognition software for use by law enforcement poses greater risks than a AI-based personalized feed on an e-commerce platform. The riskier a processing operation, the more comprehensive the accompanying evidence of an organisational or technical measure must be.[176]

How to ensure accountability from the AI development phase itself?

This is a tricky one. Accountability cannot be an afterthought, and will need to be built into the design process of AI models. Since larger models (like LLMs) are typically developed in a closed environment within corporations (where it may not be possible to share information of internal architecture), holding the right people accountable becomes difficult.[177] Proactively creating internal accountability structures can help demonstrate compliance and reduce likelihood of increased regulatory scrutiny.

To avoid this, it is important to put responsible practices in place, clarifying who is responsible for what. Building a strong culture of accountability from the beginning will help your team handle compliance challenges effectively and create an environment where people involved feel responsible for their actions.

A few principles that organisations can follow are:

- ***Placing responsibility***: Clearly demarcate roles, responsibilities and potential implications (including impact on individuals and communities) to different entities in the AI development cycle, especially developers.[178]
- ***Transparency***: Operate AI systems in a way that their decision-making processes are easily understandable to stakeholders (including end-users).
- ***Traceability***: Maintain thorough documentation of AI system development processes, decisions made, and data sources used, facilitating audits and assessments.
- ***Ethical standards***: Adhere to established ethical guidelines considering responsible AI principles (covered in detail in Section II of the Handbook).

Checklist

Here is an indicative list of organisational and technical measures for developers to consider. These may be adopted across different stages of the AI lifecycle, based on factors like the nature of the AI system, specific use-case, volume and sensitivity of personal data.

- Develop a data protection policy including guiding principles and define processes for identifying and managing data protection risks throughout the AI lifecycle.[179]
- Conduct Data Protection Impact Assessment (DPIAs) to help anticipate and mitigate risks. Include a description of data principal rights, the purpose of processing their data, and analysis of the risk to these rights.[180] Mandated for significant data fiduciaries under the DPDP Act,[181] they may be useful for organizations to embed into their processes to evaluate privacy risks/ harms at every stage of AI development and deployment.

- Apply techniques like privacy threat modelling (PTM) to map personal data collection, processing and usage across the AI system; and identify risks at each stage.[182]
- Use privacy-enhancing technologies (PETs) including application of anonymization and pseudonymization techniques like data masking and aggregation to reduce re-identification risks. Effectiveness of these techniques should be validated through rigorous testing and audits, assessing the likelihood of re-identification.
- Embed security from the beginning by incorporating safeguards. This is also mandated by the Rules. It includes encryption, data masking, virtual tokenization, and access controls (for instance, role-based access controls) to protect personal data held by businesses and their data processors.[183]
- Implement access logging mechanisms to detect unauthorized access, maintain data backups to ensure business continuity, and retain logs for at least one year for detection of unauthorized access and incident investigation.[184]
- Usage of ethical design frameworks such as fairness frameworks[185] and value-sensitive design[186] may help identify ethical dilemmas, evaluate potential impact of designs, and can prioritize user welfare in system design.
- Perform AI Impact Assessments to evaluate the privacy, security, and societal impact of the AI system prior to deployment.
- Ensure secure system configuration by following best practices such as applying security patches, disabling unnecessary services, and using encryption to protect both model training environments and downstream systems.
- Establish an AI Ethics Committee or involve experts to oversee AI project development and assess ethical implications.[187]
- Appointing key dedicated compliance officers, such as a Data Protection Officer (DPO) (also mandated for significant data fiduciaries under the DPDP Act).[188] Such officers would be responsible to ensure that an organization adheres to legal requirements and best practices, related to data privacy and security.
- Revamping internal processes through:
  - **Regular audits**: To evaluate compliance with the law and internal policies from data management to model performance.
  - **Stress tests**: Simulating extreme scenarios help in evaluating AI systems' response under pressure and help in identifying vulnerabilities and potential failure points.
  - **Scenario analysis**: Anticipates potential compliance challenges by exploring various hypothetical situations. By examining different use cases and their implications, developers can proactively identify risks and develop mitigation strategies. This practice encourages forward-thinking and prepares teams for a range of outcomes, ultimately fostering a culture of responsibility and diligence.
  - **Granular documentation**: Thorough documentation for the entire data pipeline to maintain comprehensive records including all processes, decisions, and compliance measures.
- Continuously monitor request and response patterns for anomalies indicative of active attacks or breaches.

Section II

Responsible AI

With AI's growing integration into everyday applications[189], governments across the world are actively considering how best to guide and govern its use.[190] Several jurisdictions have already proposed or implemented frameworks including European Union,[191] United States,[192] China,[193] United Arab Emirates,[194] among others.

In India, the conversation around AI governance is also gaining momentum. The government's flagship AI initiative, IndiaAI Mission, identifies "Trustworthy and Responsible AI" as a core pillar.[195] While a dedicated legal framework for AI is yet to be introduced, various government and industry-led initiatives signal increasing attention to ethical and responsible AI development. These include the NITI Aayog's approach papers,[196] MeitY's advisory frameworks on ethical AI, and the NASSCOM Responsible AI Playbook.[197] Notably, MeitY and the Office of the Principal Scientific Advisor have recently formulated the India AI Governance Guidelines,[197(A)] articulating high-level principles for responsible AI development and deployment in India and providing overarching direction for ministries, industry and researchers.

India's regulators are also actively participating in the conversation on AI governance. For instance, the Reserve Bank of India has issued the Framework for Responsible and Ethical Enablement of AI (FREE-AI) which outlines principles and safeguards for responsible AI adoption in the financial sector.[197(B)] The Securities and Exchange Board of India has also published a consultation paper proposing guidelines for the supervision and governance of AI and machine-learning tools used by market participants.[197(C)] Collectively, these efforts lay the groundwork for a principled approach to AI governance in India.

Responsible AI is not driven only through regulation. It is a practical necessity to create products that are trusted and therefore widely adopted.[198] It helps ensure that AI does not unintentionally amplify bias, cause harm, or undermine user autonomy. As AI systems become more embedded in critical sectors, there is growing recognition that responsibility must be shared.

The following sections outline the core principles of Responsible AI and how they can be applied in practice:

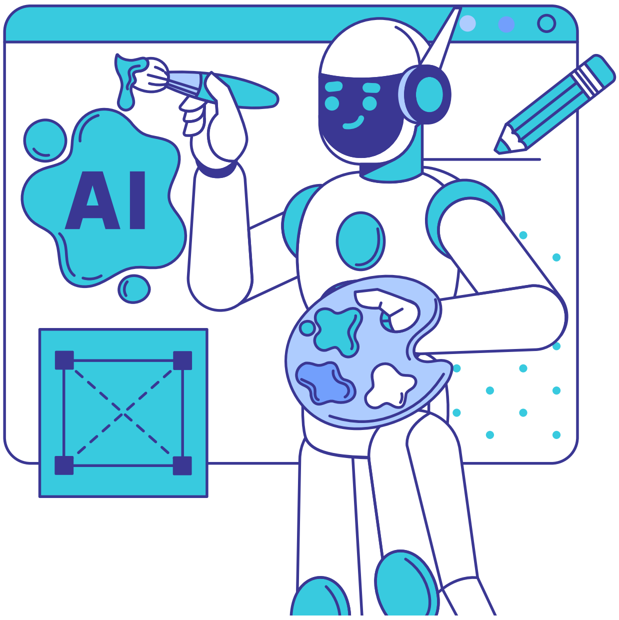
Fairness: Ensuring AI systems do not discriminate or exclude.

Transparency: Making AI decisions explainable and understandable.

Accountability: Clarifying who is responsible and how concerns can be addressed.

Security: Protecting systems from misuse, manipulation, and emerging risks.

These principles provide a practical framework to support the development of AI that is both innovative and aligned with user trust and societal values.



Fairness

What does Fairness mean?

Fairness in AI refers to principles and practices aimed at eliminating bias and discrimination in AI models. The goal of the fairness principle is to ensure equitable treatment across individuals and groups and promote inclusivity in automated decision-making.[199] It applies not only to discriminative AI systems (which classify, predict, or rank individuals), but also in generative AI systems (which generate content like image, text, or audio).[200] Addressing fairness requires attention to how AI models are developed, trained and aligned.[201]

It encompasses statistical fairness (which covers mathematical and computational methods that ensure AI models do not unfairly disadvantage a particular group) and social fairness (which covers broader and more qualitative impact of AI systems on individuals and society).[202, 203]

Different concepts of fairness in AI systems

Fairness in AI is a multifaceted concept, with overlapping but distinct types. This includes:[204]

- ***Mathematical side of fairness:*** Associated with quantitative definitions and measures which assess whether a model's predictions or decisions are fair — both at individual and group levels — to ensure fairness in a more 'objective' manner.
- ***Group fairness:*** Means treating members of different groups like caste, gender, or age, equally. It also focuses on making sure that outcomes are fairly balanced. For instance, a lending model deciding on loan approvals. If 40% of male applicants are approved and demographic parity is to be maintained, then 40% of female applicants should also be approved, irrespective of other factors like income or credit score.[205]
- ***Individual fairness:*** Asserts that similar individuals should be treated similarly. Decisions should be consistent and based on relevant attributes, not biased by irrelevant characteristics. For instance, in an AI system that predicts academic performance, individual fairness ensures that two students with similar academic histories and skills are treated equally, without bias based on irrelevant factors like their geographic location or family background.
- ***Counterfactual fairness:*** Aims to ensure that AI systems make the same decision for an individual — regardless of their group membership — even if their attributes were

different. For instance, in a loan approval AI system following counterfactual fairness, applicants' loan approval or rejection will be the same, if they are from different castes or religious groups — as long as their other attributes like income and credit score are the same.

- ***Procedural fairness:*** Emphasizes that the process used to make decisions should be fair and transparent. For instance, procedural fairness would ensure that an algorithm's decision-making process is transparent in an AI-driven decision-making system for healthcare treatment recommendations. This would include sharing how the algorithm arrived at its decision, a clear basis for its recommendation, and make the process open to review.
- ***Causal fairness:*** Aims to ensure that AI systems do not reinforce historical biases and inequalities. For instance, in an employment screening AI tool, causal fairness would ensure that the algorithm does not reinforce historical biases and patterns of discrimination, like favouring male candidates over female candidates for leadership positions. In fact, such systems should be designed to break these patterns, rather than perpetuate them.

All bias is not necessarily bad! A positive bias, created through curated data sets favouring marginalized social groups, can help AI systems make the right decisions.[206]

Sectoral examples of fairness in AI

Fintech: AI is used for credit scoring, loan approvals, and fraud detection. However, historical data used to train these models may reflect societal biases, leading to unfair outcomes. For example, a credit scoring model may unfairly penalize applicants from specific regions or socioeconomic backgrounds, denying them access to credit. Fairness in fintech AI systems is crucial to promoting financial inclusion and preventing discrimination.[207]

Healthcare: AI is used for disease diagnosis, drug discovery, and personalized treatment recommendations. However, AI models may make unfair decisions if the training data lacks diversity or reflects biases prevalent in the healthcare system.[208] For example, if an AI model is used to recommend heart disease treatments, the training data mainly reflects outcomes for male patients. The AI model may not be able to recommend equally effective treatments for women because it lacks sufficient data on their specific symptoms and responses to treatment, which can differ from men.

Agriculture: AI is used for precision farming, crop yield prediction, and pest management. However, if the AI models are trained on data from certain regions having specific farming practices, they may not perform well in other contexts.[209] For example, an AI model designed for crop yield prediction, trained on data from large-scale farms in the United States using advanced irrigation systems, may not perform well to small-scale farms in India relying on rain-fed agriculture. The AI model might overestimate yields because it assumes a consistent water supply, whereas India's local reality is of uncertain rainfall patterns.

Telecom: AI is used for network optimization, customer service, and fraud detection. However, if the AI models are trained on data from specific regions or customer segments, they may not perform well for other regions or segments.[210] For instance, an AI-powered chatbot trained primarily on customer service interactions in English or Hindi will not be able to respond effectively to customers who speak other languages like Tamil, Telugu, or Bengali. This limitation could lead to poor service quality in states where such languages are spoken more, such as Tamil Nadu, Andhra Pradesh, and West Bengal.

Why is ensuring Fairness important in India's context?

In a diverse country like India, ensuring fairness in AI systems is crucial for promoting socio-economic equity.[211] Given India's wide range of socio-economic backgrounds, castes, languages, cultures, and religions, AI systems must be designed to avoid discriminating against any particular social group.[212] Fair AI systems can help mitigate historical and societal biases, fostering inclusivity and reducing disparities. Therefore, it is essential to integrate fairness principles throughout the entire lifecycle of AI systems - from design and development to deployment and monitoring.[213]

Recognizing these risks, in 2023, the Telecommunication Engineering Centre (TEC), an arm of the Department of Telecommunications, Government of India, released the *Standard for Fairness Assessment and Rating of Artificial Intelligence Systems* to test for bias in AI-generated outcomes.[214] These standards provide a structured framework to assess and certify the fairness of AI systems through a three-step process: bias risk classification, fairness metric selection, and bias testing. It covers both group and individual fairness metrics, offering standard operating procedures (SOPs) for self-assessment or third-party auditing. The goal of these standards is to promote trustworthy AI by enabling transparency, risk-based evaluation, and comparability across systems.

Similarly, in 2023, Google announced its intention to study bias from an Indian societal context by focusing on cultural factors relevant to India, such as caste, religion, and language. Google emphasized that existing bias evaluation and mitigation measures must be recontextualized to the Indian context before application.

Challenges in ensuring Fairness

Ensuring fairness in AI systems is complex due to the nature of AI technologies and is exacerbated due to the Indian societal context. Fairness is not a one-size-fits-all concept — it rather depends on the situation, making it difficult to apply a universal standard across all AI applications.

- ***Conflicting definitions of fairness:*** A key challenge is of multiple, often conflicting, definitions of fairness.[215] For instance, equal opportunity aims to ensure that equally qualified individuals across groups have the same chance of a positive outcome. In contrast, demographic parity seeks equal outcomes across groups, regardless of group-specific qualifications.[216] This means if one group is underrepresented because there are fewer qualified individuals (who may have had historical disadvantage and limited access to opportunity by virtue of being in a particular demographic group), achieving demographic parity might require prioritizing these underrepresented groups — potentially at odds with equal opportunity. Developers must navigate these trade-offs, aligning their choices with the system's goals and its broader ethical and social context.
- **Biased training data:** A major source of unfairness comes from biased training

data.[217] AI systems learn patterns from the data they are given, and if that data reflects existing societal biases, the system may reinforce those biases in its decisions. Addressing this requires not only technical proficiency but also a nuanced understanding of the data's social implications.

- ***Lack of transparency:*** Many AI models, particularly those based on deep learning, often function as "black boxes," making it difficult to understand or explain how decisions are made. This lack of clarity makes it harder to detect and correct unfair outcomes.[218]
- ***Scale-related concerns:*** Solutions that work in controlled or small-scale environments may falter at scale. As AI systems grow in complexity and reach, ensuring fairness across diverse contexts becomes increasingly difficult.
- ***Performance trade-offs:*** Techniques designed to promote fairness like anonymization or algorithmic adjustments can sometimes reduce AI systems' accuracy or efficiency. Developers must often balance the trade-off between fairness and performance, which may limit the system's effectiveness.[219] For instance, in the prevention of adverse tuberculosis (TB) outcomes (PATO), the trade-off between accuracy and fairness became particularly evident when performance audits revealed that the AI system was more accurate in predicting adverse TB outcomes for male patients than for female patients. While the model demonstrated a high overall recall (~70%) - significantly better than rule-based baselines (~50%), this aggregate performance masked disparities across gender cohorts. Addressing these fairness gaps required post-hoc fairness-enhancing algorithms, which slightly adjusted the model's predictions to equalize accuracy across groups. Although this rebalancing led to a drop in the overall performance, it was a deliberate design choice to prioritize equity in public health outcomes. This ensures that no sub-group especially women (who may face different barriers to TB treatment)—was systematically under-served by the AI. This case underscores that fairness is not a static metric but an evolving design objective, especially in high-stakes settings where public service delivery is involved.
- ***Organisational resistance:*** Even when technical solutions are feasible, organizational resistance can pose implementation challenges. Some stakeholders may oppose fairness interventions if they threaten existing power structures or profit margins, making implementation difficult despite the availability of solutions.[220]

Fairness and mitigating model bias: the same concept?

Fairness and mitigating model bias are closely related; however, bias is more of a technical issue, while fairness is a social and ethical issue.[221]

Fairness in AI refers to the absence of discrimination or favouritism toward any individual or group based on protected characteristics such as caste, gender, age, or religion. It requires a conscious effort to ensure the algorithm does not discriminate against any group/ individual.

On the other hand, model bias refers to systematic errors in the AI lifecycle that skew outcomes in favour of or against certain individuals or groups.[222] These biases can emerge from training data, model design, algorithmic implementation, or deployment contexts. Even minor skews can amplify societal inequities - a phenomenon often called the Butterfly Effect in AI systems.[223]

Types of bias

- ***Sampling bias:*** Arises when training data does not represent its intended population. For instance, an AI diagnostic tool trained primarily on data from one ethnic group may not perform accurately for other ethnic groups.
- ***Algorithmic bias:*** Often stems from the design and implementation of the algorithm where certain attributes are unintentionally prioritized and may subsequently lead to unfair outcomes. For instance, historical arrest data from Oakland, California reflects patterns of over-policing in African American communities. If such data is used to train a predictive policing algorithm, it may reinforce and perpetuate those past biases, resulting in discriminatory outcomes.[224]
- ***Measurement bias:*** Occurs when data collection methods over/underrepresent groups. For instance, speech recognition systems trained on one gender may fail to recognize speech patterns from another gender.
- ***Representation bias:*** Appears when a dataset does not accurately represent the population it is meant to model, which leads to inaccurate predictions. For instance, a US-based hospital algorithm which predicted which patients need additional medical care was found to be biased against Black patients — because it used healthcare cost history, which did not account for different ways communities pay for healthcare.
- ***Generative bias:*** Occurs in generative AI models and emerges when the model's outputs disproportionately reflect specific attributes, perspectives, or patterns present in the training data, leading to skewed or unbalanced representations in the generated content. For instance, AI consistently depicting 'terrorists' as men with dark facial hair.
- ***Confirmation bias:*** Emerges when algorithms are designed to learn from user interactions, reinforcing the users' beliefs or biases. Such personalized content recommendations can create echo chambers by repeatedly reinforcing users' existing beliefs.

Addressing model bias is an important step towards achieving fairness in AI – it can help reduce the likelihood of unfair outcomes. However, achieving fairness may require additional efforts beyond bias mitigation.[225]

Developers (particularly with deep learning systems) face challenges in mitigating model bias. Deep learning systems often operate as black boxes, making bias difficult to trace or fix. Available datasets may be incomplete or historically biased.

How to mitigate model bias?

- ***Diversify (as far as practicable) AI development teams and datasets:*** To design and deploy AI systems, engage a range of stakeholders — across demographic groups, age, caste, gender and socioeconomic status — to prevent the system from developing skewed outputs.
- ***Ensure human-in-the-loop:*** Incorporating human reviewers or moderators in the AI lifecycle development can help mitigate risks and provide a checks-and-balance system to prevent propagation of biased or harmful content.
- ***Regular audits and tests for bias:*** Conducted by independent third-party reviewers or internal teams dedicated to fairness, audits allow examining input data and algorithm outputs to identify potential biases and their sources.
- ***Transparent and explainable AI systems:*** Documenting choices made during the algorithm's development, such as which features are used and how they are weighted in the decision-making process.
- ***Design self-learning AI systems:*** Capable of rectifying its outputs by incorporating feedback loops, which allow the system to adjust and refine its algorithms in response to identified biases.
- ***Utilise technical measures:*** Different technical measures like oversampling, synthetic data generation, regularization and ensemble models may help in achieving equalized odds. Care must be taken to ensure that 'de-biasing' measures suit the Indian context.

Evaluating fairness in AI: Why metrics are not enough

Evaluating fairness in AI systems involves more than just running an algorithm and reading off results. It requires active engagement with both performance metrics (like accuracy, precision, recall) and fairness-specific metrics (like demographic parity, equal opportunity, or disparate impact).

Why metrics matter?

Metrics are essential for identifying whether an AI system is treating individuals or groups unfairly. These metrics help quantify fairness issues, giving developers a way to assess whether the model might be favoring or disadvantaging certain populations. (a). Demographic parity checks if different demographic groups receive positive outcomes at equal rates. (b). Equalized odds measures whether the model's error rates (false positives and false negatives) are similar across groups.

Role of open-source tools

Tools (like IBM's AI Fairness 360, Microsoft's Fairlearn, and Google's What-If Tool) that help compute these metrics automatically. These tools save time by automating calculations, offer visualizations to better understand group-level performance and allow easy experimentation with different fairness interventions.

Tools aren't enough — they are facilitators

Despite their usefulness, these tools do not replace human judgment since:

- Fairness is context-specific: A model that satisfies one fairness metric may still be unfair in another context.
- Trade-offs exist: Improving fairness metrics might hurt model accuracy or operational efficiency. Choosing which trade-offs are acceptable requires ethical reasoning and domain understanding.
- Interpretation matters: Tools can flag issues, but only developers and domain experts can assess their real-world significance. For instance, is a 3% disparity in false positive rates across genders acceptable in a loan approval system?

Bottom Line

Metrics and tools are powerful, but evaluating fairness in AI is not a plug-and-play task. It requires thoughtful, case-specific interpretation—and sometimes, difficult judgment calls. Developers must go beyond surface-level statistics to understand the deeper impact of their models on individuals and society.

How to ensure Fairness?

- ***Identifying fairness goals, risks and stakeholder mapping[226]:*** Map and conduct potential risks and stakeholder analysis, considering those who could be affected by the AI system. Stakeholders include users, developers, impacted communities, regulators, and business partners. To better understand impact on identified stakeholders, conduct a risk assessment involving specific evaluation of potential risks of biased outcomes.
- ***Evaluating data sources and addressing bias[227]:*** Before any model is trained, fairness should be addressed in the primary data set itself. Biases embedded in historical data may amplify existing inequalities.[228] Techniques like resampling (to fix imbalances)[229], reweighting (to reduce influence of dominant data clusters/ biased samples)[230], or synthetic data generation (to fill gaps for underrepresented groups in the data)[231] can be used. Approaches like fairness-aware data clustering[232] can ensure that data patterns do not inadvertently encode societal assumptions and biases.
- ***Embedding fairness during model training:*** Integrate fairness into learning algorithms during the model training phase itself.[233] Fairness-aware algorithms should also take into account identified risks and stakeholder needs.[234] This can be done by modifying the objective function to

include fairness constraints,[235] or encoding techniques directly into the training process like adversarial debiasing,[236] equal opportunity[237] or demographic parity.[238] These methods can optimize both accuracy and fairness.

- ***Fairness assessment and mitigation:*** It is critical to implement a structured, ongoing process that spans the entire AI lifecycle — from system design and data collection, to model training, deployment, and updates.
- ***Model development workflow:*** Select fairness metrics early, aligned with real-world impact and system goals, and ensure the development process is inclusive and auditable.
- ***Mitigating bias:*** Effective bias mitigation must extend beyond model design and into practical resilience. It is essential to evaluate and stress-test fairness interventions under real-world conditions. This includes experimenting with multiple algorithmic techniques and examining how well these approaches generalize across varied datasets, including out-of-distribution (OOD) data. Where generalization is limited, risks should be transparently communicated to downstream users and decision-makers to ensure informed deployment.
- ***Continuous and iterative fairness testing:*** Fairness is not a one-time test—it must be monitored continuously. Regular testing throughout the model development cycle is necessary to detect emergent biases and disparities. These evaluations should include assessments across sensitive and intersectional sub-groups. In addition, systematic misclassification patterns (especially those disproportionately affecting specific groups) must be analysed and addressed as part of the fairness assurance strategy.
- ***Comprehensive documentation and transparency:*** Transparency is foundational to trustworthy AI. It facilitates fairness, allowing stakeholders to understand and trust the decision-making process,[239] and outlines how their feedback was incorporated into updates and refinements to the system. A well-documented record of fairness assessments, data decisions, model parameters, and mitigation measures enables accountability and external auditability. This documentation should also capture how stakeholder feedback is integrated into model refinement processes. Ensuring this level of visibility builds trust among users, regulators, and affected communities, and supports the explainability of complex AI systems.
- ***Fairness-oriented output adjustment:*** To adjust the model output in a manner that ensures fair treatment across different groups (especially stakeholders at higher risk of bias) post-processing methods can be applied, after a model is trained.[240] This can include calibration[241] and reject option classification.[242] Performance metrics such as accuracy, precision and recall, and fairness metrics such as demographic parity, equal opportunity can be evaluated to ensure that the model treats different demographic groups fairly.
- ***Continuous monitoring and regular evaluation:*** Monitoring and evaluation AI systems are required to ensure ongoing fairness. Tools like IBM's AI Fairness 360 or Google's Fairness Indicators offer structured methods for assessing fairness metrics across datasets and models, both during and after deployment. These tools support dynamic adjustment of models in response to observed bias, ensuring fairness is maintained in evolving real-world contexts.

Checklist

Conception and design

1. **Identify fairness goals, risks and stakeholders**
    - Identify goals of your AI system.
    - Identify potential risks with the AI system.
    - Map stakeholders (direct and indirect) impacted by the AI system.
    - Choose fairness metrics suited to the goals, risks and stakeholders.
    - Keep a plan to review and update fairness definitions as the system evolves.

2. **Evaluating data sources**
    - Review data sources for demographic diversity.
    - Identify and document sampling gaps during data collection.
    - Apply corrective methods to bridge sampling gaps.
    - In case using synthetic data, label and assess data for fairness.

Development

3. **Embed fairness during model training**
    - Use fairness-aware algorithms aligned with identified risks.
    - Evaluate and document trade-offs between fairness and accuracy.
    - If pre-trained model is being used, assess it for known biases.
    - Have an AI lifecycle-wide plan to assess and mitigate fairness risks.
    - Include regular testing and model monitoring in the plan.
    - Identify and define roles and responsibilities for ensuring fairness.
    - Ensure steps and decisions related to fairness are properly logged and versioned.

4. **Embed fairness during workflow**
    - Select fairness metrics (like demographic parity, equal opportunity, equalized odds) based on context and impact of AI system.
    - Have a process to analyze and visualize assessment results.

5. **Mitigate bias**
    - Experiment with different bias mitigation algorithms.
    - Validate the model's performance on out-of-distribution datasets.
    - Check if generalization is inadequate.
    - If it is, disclose residual risks clearly to users.

6. **Conduct continuous and iterative fairness testing**
    - Select fairness metrics (like demographic parity, equal opportunity, equalized odds) based on context and impact of AI system.
    - Test the AI system across sensitive and intersectional sub-groups.
    - Regularly assess error patterns and misclassifications for systemic bias.

7. **Comprehensive documentation**

- Document fairness assessments, mitigation strategies, and associated decisions systematically at each stage
- Keep a record of how stakeholder feedback informed model updates

Deployment

8. Monitor continuously

- Assess if fairness evaluations are based on metrics appropriate to the system's context
- Regularly monitor AI models post-deployment to check for evolving bias
- Integrate stakeholder feedback into fairness monitoring cycles
- Check if risks identified during stakeholder mapping been revisited and updated

9. Adjust output to ensure fair treatment

- Apply appropriate post-processing techniques to address bias
- Evaluate fairness improvements alongside core performance metrics

10. Incorporate feedback mechanisms and redressal:

- Define accessible channels for stakeholders to report concerns
- Include a clear, publicized process for contesting AI decisions
- Track user grievances and feed them back into model improvement cycles

Transparency

What does Transparency mean?

Transparency in AI refers to the clarity and openness with which AI systems are developed, deployed, and operated. Transparency demystifies AI processes, making AI's decision-making more understandable and ensuring its actions align with ethical standards.[243] It encompasses:

- Disclosure that AI is being used in a system or decision.
- Accessible and clear explanation of how the AI system operates by informing stakeholders how an AI system is developed, trained, operated, and deployed in a particular circumstance.[244, 245]
- Sharing details including possible use of open-source models,[246] which enhances trust in AI.[247]
- Traceability of decisions through detailed disclosure of algorithmic procedures, data handling, and decision-making frameworks, ensuring stakeholders can understand, scrutinize, and eventually trust the AI model.[248]

Many AI systems, particularly those based on deep learning, operate as "black boxes", meaning their internal workings are opaque and not easily understood by humans.[249] Third-party auditors lacking a contractual relationship with the audited system have faced limitations in access, which constrained some auditing techniques. However, in cases of complete independence with the freedom to ask more difficult questions about system outcomes, potential for more explicit assessments would get enhanced.[250]

Sectoral examples of transparency in AI

- **Healthcare**: AI models diagnosing diseases or recommending treatments must provide clear rationales for their decisions to be accepted by medical professionals and patients. For instance, in India, the Indian Council of Medical Research (ICMR) guidelines emphasize that AI algorithms must be transparent so that healthcare professionals understand the factors influencing treatment recommendations.[251]
- **Fintech**: AI systems employed in credit scoring must justify their decisions to meet regulatory standards and promote fair lending practices.[252]
- **Agriculture**: AI-powered systems that provide real-time insights on weather patterns, soil conditions, crop health, and

energy consumption must clarify the data collection methods and provide open access to algorithm updates.[253]

- **Smart Cities**: AI-powered transport systems should disclose the factors they employ to give real-time suggestions on traffic management. By disclosing this information, the system enhances public trust and allows users to understand the basis for traffic management decisions, thereby improving compliance with suggested routes and timings.[254]

Interrelatedness of transparency and explainability

Explainability in AI (XAI) systems refers to developing methods and tools to provide clear, understandable, and accessible explanations of their processes and decisions. This is essential for users and stakeholders to comprehend how AI models operate, how they make decisions, and on what basis those decisions are made.

It encompasses several key components:

- Transparency, which provides insight into the AI system's algorithms, data usage, and decision-making processes.
- Interpretability, which is the degree to which a human can understand the cause of an AI system's decision. Interpretability might involve simple, rule-based models with straightforward logic or complex models with mechanisms employed to elucidate decisions.
- Comprehensibility, which concerns the ability of different types of users, not just AI experts, to understand AI processes and outputs. Explanations must be tailored to the audience's level of technical expertise.
- Traceability, which is the ability to trace an AI system's decision-making process step-by-step. It helps audit AI systems and is crucial for validating the outcomes and ensuring that the AI adheres to the intended design.

Why is transparency important?

Transparency is important because it fosters trust among users, developers, and regulators, making processes and decisions more understandable[255] by showcasing how AI systems are not just "black boxes" but tools whose functionality and reasoning can be assessed, critiqued, and understood.[256] It can help demonstrate adherence to ethical standards and prevent misuse and biases that could lead to unfair outcomes. Moreover, it can drive innovation by encouraging the creation of common standards that allow different AI systems to work together.[257]

- **Building trust between AI systems and their users**: When users understand how decisions are made, they are more likely to trust and rely on AI systems.[258] This trust is crucial to facilitating the adoption of AI technologies, especially in sectors such as healthcare, finance, and legal where decisions may significantly impact individuals' lives.[259]
- **Facilitating accountability:** This is particularly important in scenarios where decisions need to be justified or where there may be disputes about the fairness or correctness of AI-generated outcomes.[260]
- **Promoting ethical decision-making[261]**: By ensuring that decisions are made transparently and can be evaluated against ethical standards. This is essential in ensuring that AI systems do not perpetuate or exacerbate existing biases.[262]
- **Enabling improvement of AI systems**: Through better diagnostics and innovations in AI system development, leading to more robust AI solutions. When AI developers and stakeholders understand how AI models arrive at their conclusions, they are better equipped to identify errors, biases, or areas of inefficiency.[263]

- **Catering to diverse stakeholder needs**: Different stakeholders, from developers to end-users and regulators, may require an understanding of AI systems relevant to the application's context. Explainability ensures that AI systems can be interpreted appropriately across this diverse spectrum, satisfying various informational needs and usage contexts.[264]

This means AI models can integrate more easily into existing technologies, making development faster and more efficient across platforms. By sharing how their systems operate, developers can ensure that their models are compatible with others, which helps advance the overall technological ecosystem. Operationalizing transparency also helps achieve the goals of other responsible AI principles, such as privacy, accountability and safety.[265]

Dark patterns

An important aspect of operationalizing transparency in AI systems is addressing dark patterns. These are manipulative UI/UX design choices that nudge or deceive users into actions they did not intend often by exploiting cognitive biases or hiding key information. The Indian Ministry of Consumer Affairs defines dark patterns as interface practices that mislead users, impair autonomy, or distort decision-making amounting to unfair trade practices or consumer rights violations. In the AI context, dark patterns can obscure data practices, complicate opt-outs, or present biased defaults, ultimately undermining user trust, fairness, and accountability.[266]

- **European Union**: The EU AI Act prohibits AI systems that use manipulative or deceptive techniques to distort user behavior, linking transparency directly with dark pattern prevention. It mandates clear disclosure for AI interactions (e.g., chatbots, deepfakes) to ensure users know when they're engaging with AI. Similarly, the Digital Services Act (DSA) bans dark patterns on online platforms, requiring interfaces to support informed, autonomous choices. Together, these laws reflect the EU's stance that transparency and fairness-by-design are key to preventing user manipulation.

- **United Kingdom**: UK links transparency with fair design across both privacy and consumer protection. The ICO's Age Appropriate Design Code bans manipulative UI practices targeting children, like nudges to weaken privacy, and holds that any consent gained through dark patterns may be invalid under UK GDPR. On the competition side, the Competition and Markets Authority (CMA) targets "harmful online choice architecture" (e.g., confirmshaming, sludge, biased defaults) as unfair commercial practices. The Digital Markets, Competition and Consumers Act 2023 empowers the CMA to directly penalize and impose "fairness by design" duties making deceptive interfaces a clear legal risk.

- **United States**: The Federal Trade Commission (FTC) has warned that manipulative use of generative AI may be illegal under existing consumer protection laws. Deceptive chatbot behaviour, unclear ad disclosures, or biased AI advice can all qualify as unfair practices. The FTC emphasizes that users must know when they're interacting with AI or ads and must not be misled by interface design. Emerging state laws (like California's Privacy Rights Act) are reinforcing this, defining dark patterns as practices that impair user autonomy.

- **Other global-level efforts**: Bodies like the OECD and G7 link transparency with preventing AI-driven manipulation. The OECD AI Principles urge clear disclosure when users interact with AI and warn that transparency alone is insufficient if interfaces remain misleading. Their research highlights how machine learning can amplify dark patterns by targeting user vulnerabilities. Similarly, the G7's AI Code of Conduct (Hiroshima Process) calls for transparency, content authenticity, and risk mitigation, reinforcing a global consensus that AI must empower users, not mislead them.

Transparency in practice

So, what does transparency mean in practice? Consider the healthcare sector – where AI models have several critical applications. For instance, they can assist in diagnosing diseases by analyzing medical images, predicting patient outcomes, and recommending personalized treatment plans.

Ensuring transparency here means that patients, their families, and healthcare providers should be: (a) informed that AI is being used, and (b) given a comprehensive explanation of broadly three aspects: (i) AI models' operational processes and outcomes; (ii) Benefits, potential drawbacks, and risks associated with employing the AI models in medical decision-making, along with the corrective measures taken to minimize any risks; and (iii) Details about data ownership who controls and accesses the patients' data.

Key challenges

Ensuring transparency can present several challenges – stemming from technical complexities, operational constraints, and broader ethical considerations.

- **Increasing complexity of AI models**: Especially those based on deep learning, which consists of millions of parameters and non-linear interactions that are inherently difficult to interpret.[267] This can make understanding how decisions are made difficult particularly since AI systems operate in dynamic environments and are continuously updated with new data.[268]
- **Trade-off between transparency and model performance**: Enhancing interpretability or transparency of AI models often involves simplifications or modifications that can reduce their performance.[269] For example, simpler models that are inherently more interpretable may not achieve the same level of accuracy as more complex models. Balancing transparency with performance can be a significant challenge, especially in applications where performance is critical.[270]
- **Intellectual Property (IP) concerns**: IP significantly hinders transparency, particularly for commercial AI applications. Companies may be reluctant to disclose detailed AI systems disclosures due to fears of exposing proprietary source codes to competitors. This tension between transparency and protecting IP rights adds another layer of complexity to the challenge.[271]
- **Lack of standardization**: Without a universal standard, approaches to transparency can vary widely, making it difficult for stakeholders to assess and compare the transparency of different systems.[272]
- **Conflict with privacy and security requirements**: Detailed explanations of how data influences AI decisions could inadvertently reveal sensitive or personal information, leading to privacy breaches.[273]
- **Technicalities involved**: Even when AI systems are designed to be transparent, the technical nature of AI and machine learning can make it difficult for non-experts to understand. Ensuring that explanations are accessible and meaningful to all users, regardless of their technical background, poses a substantial challenge in promoting user engagement and trust in AI systems.[274]

How to ensure transparency?

Developers can choose the most appropriate method for ensuring transparency depending on the system's context, purpose, and risk level, developers must choose the most appropriate methods from a broad toolkit. The following section offers a snapshot of key techniques ranging from problem formulation to post-deployment auditing to help understand the spectrum of available practices and guide further exploration.

- **Problem formulation to define purpose, boundaries and impact**: A transparent AI system starts with a well-defined purpose, aligned with user expectations and societal impact.[275] A well-scoped formulation must go beyond technical requirements. It should clarify who the system intends to serve, what success looks like, and which harms must be anticipated or avoided. Equally critical is identifying potential edge cases, limitations, and the system's interaction with human decision-makers. Early-stage transparency around these elements sets the tone for ethical design, informed deployment, and sustained accountability.
- **Transparency by design**: Embedding transparency features from the outset allows for systematic auditing, builds trust among stakeholders, and facilitates ethical oversight.[276] Whenever possible, developers should choose inherently interpretable models, such as decision trees,[277] linear regression,[278] or rule-based systems[279]. These models allow users to see how input variables are transformed into outputs, making the decision-making process transparent.[280]
- **Explainability and use of XAI techniques**: Interpretability is critical for demystifying AI systems, especially those deployed in sensitive domains. Developers may adopt either intrinsically interpretable models (e.g. decision trees or rule-based systems)[281] or post-hoc explanation techniques to clarify the behavior of more complex "black-box" models. There are tools which break down and explain the contributions of each feature to the model's output. Model-agnostic techniques include tools such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations) and ELI5 (Explain Like I'm Five) which all offer ways to.[282]
- **Data transparency and lifecycle documentation**: Comprehensive documentation is central to transparency. Throughout the AI lifecycle from data ingestion to model deployment developers must maintain detailed descriptions of the algorithms, AI training dataset,[283] data handling procedures, model training processes, and decision-making criteria. This documentation should be written in accessible language that serves both technical stakeholders (like fellow engineers, data scientists) and non-technical audiences (like end-users or policymakers).[284] Model Card toolkits can be a helpful resource for developers in preparing such documentation.[285]
- **Open-sourcing and reproducibility**: Open-source practices advance transparency by enabling independent verification, collaborative oversight, and public accountability. Making AI models and datasets openly accessible allows researchers, policymakers, and industry experts to examine, test, and improve systems, ensuring that their decision-making processes remain explainable and trustworthy.[286] This openness not only builds public trust but also aligns with regulatory expectations around auditability and ethical compliance. It enables third parties to replicate outcomes, scrutinize risk, and contribute to refining the system. In high-impact domains, open-sourcing acts as a safeguard against unaccountable or opaque AI decision-making.
- **Model architecture and training process**: Transparent AI development requires more than documenting model outcomes, it demands a full account of how models are selected, built, and trained. Developers should provide clear,

structured documentation covering model architecture, training configurations, and the rationale behind key design decisions.

- **Evaluation and validation transparency**: Developers must maintain robust version control and provenance tracking across codebases, datasets, and model artifacts. Additionally, methods for feature selection and importance should be disclosed, especially when different model types lead to varying interpretations of feature relevance such as differences between Random Forest and XGBoost, which are machine learning algorithms.[286(A)]
- **User-centric explanations and interfaces**: User-centric interfaces offering users insights into real-time AI decisions can enhance transparency. These interfaces should be designed to provide explanations tailored to the user's expertise level, ensuring that the system's operations are understandable.[287]
- **Monitoring, auditing, and model updates by incorporating visualisation techniques**: Visual aids enhance the comprehensibility of AI decisions. Visualization techniques can include plotting feature importance, decision trees, or the effects of different inputs on outputs. These visualization techniques are helpful in applications like medical imaging or autonomous driving, where understanding the model's focus areas can provide insights into its reliability and decision-making process. Examples include techniques such as heatmaps, saliency maps, and partial dependence plots, which can help visualize which parts of the data are most influential in making the model's predictions.[288]
- **Use content authentication and tracking**: With increasing concerns around verifying AI-generated content, using content authentication measures such as watermarking and content credentials help build trust and restore transparency.[289] Global initiatives such as the Coalition for Content Provenance and Authenticity (C2PA), an open technical standard, provides a secure and transparent framework by creating verifiable provenance information associated with digital content.[290]

Checklist

Conception and design

1. **Formulate problem statement and impact**
   - Document the AI system's purpose, objectives, and intended use case.
   - Map potential risks, limitations and stakeholders of the AI model.
2. **Embed transparency by design**
   - Prioritise transparency from the initial design phase.
   - Consider and select interpretable models where appropriate.
   - Embed decision logs and traceability mechanisms in the system.
3. **Include explainability in your model**
   - Select appropriate explainability methods.
   - Profile the training data for quality, representativeness to ensure there is no bias.
   - Well-document pre-processing, cleaning, and transformation steps.
4. **Ensure data Transparency through documentation**
   - Document dataset sources, structure, and pre-processing steps.

- Maintain accessible records of training processes, model architectures, and parameters.
- Centralize all lifecycle documentation in a single, maintained repository.
- Clearly articulate decision criteria and rationale for model selection.
- Use tools like Model Cards to structure documentation for diverse audiences.

5. **Encourage use of open-source and enable**
   - Use open-source AI components where appropriate.
   - Provide usage guidelines and licensing terms with open-source materials.
   - Enable reproducibility through independent validation and external collaboration.
   - Maintain version histories and changelogs for transparency and traceability.

Deployment

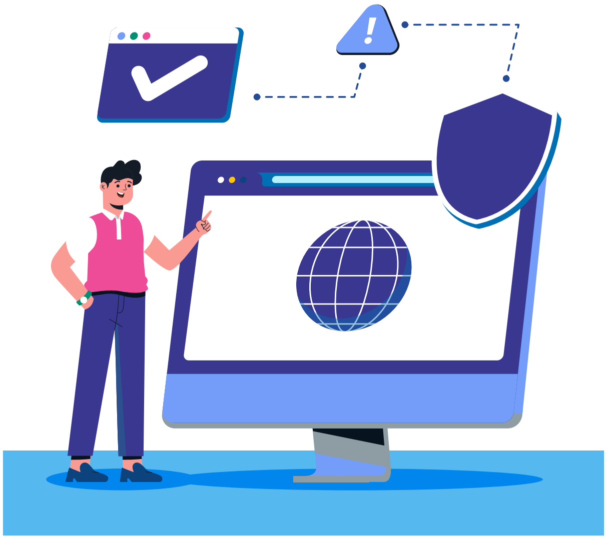
6. **Model architecture and training process should be in place**
   - Clearly document model type, architecture, and configuration.
   - Log model selection decisions with underlying rationale.
   - Record training procedures, including hardware, batch size, optimizer, and validation methods.

7. **Evaluate and validate transparency**
   - Report validation and test performance metrics in detail.
   - Disclose and compare feature selection or importance ranking methods across models.
   - Track provenance and version control for datasets, codebases, and model artifacts.

8. **Design user-centric interfaces**
   - Design user - facing explanations tailored to different technical expertise levels.
   - Establish clear response processes for user inquiries on AI decisions.
   - Provide users access to system documentation and channels for feedback or concerns.

Deployment

9. **Post-deployment transparency and interpretability tools**
   - Use visualization techniques (e.g. heatmaps, saliency maps, partial dependence plots) to enhance interpretability.
   - Publish regular transparency reports on system performance, fairness, and impacts.
   - Maintain detailed records of post-deployment changes, including model updates and data refreshes.



Accountability

What does Accountability mean?

Accountability in the context of AI refers to the clear identification of individuals and entities responsible for various stages of the AI system lifecycle, ensuring they can be held accountable for the outcomes produced by these systems. Essential to the principle of accountability is ensuring human oversight or audits to ensure responsible governance of AI systems.[291]

The following principles support effective accountability:

- **Awareness**, which means to cultivate a culture of ethical awareness and access to information among individuals involved in the design, development, and deployment of AI systems, empowering them to make responsible decisions.[292]
- **Clearly defined roles and responsibilities**, by delineating accountability for specific aspects of the AI system's development, operation, and impact.[293]
- **Audits**, to provide unbiased assessments (which must lead to appropriate consequences) of the AI system's performance, ensuring it aligns with its intended goals and identifying areas for improvement.[294]

Why is Accountability important?

Accountability in AI is critical, particularly as AI systems increasingly influence decisions across sectors with significant social, economic, and ethical implications such as healthcare, finance, and criminal justice. Without clear accountability, harmful outcomes stemming from AI-driven decisions can leave affected individuals without recourse or redress, unsure of exactly with whom the liability lies for a given problem, thereby eroding trust in the technology.[295]

Accountability is essential for fostering public trust and ensuring that AI aligns with societal values. By integrating responsible AI practices, developers not only demonstrate their commitment to ethical principles but also build long-term credibility.[296] This trust is crucial for user acceptance, as AI systems that are transparent and accountable are more likely to gain public approval and sustain market presence. Moreover, accountability helps position AI technologies within a framework that benefits both users and society, supporting human well-being and safeguarding human rights. Ethical deployment of AI reinforces integrity and ensures that technology serves humanity rather than undermining it.[297]

In the Indian context, accountability becomes even more critical given the country's diversity and complexity. India's societal fabric, with its varying socio-economic, cultural, and linguistic dimensions, amplifies the risks of algorithmic bias, discrimination, and misinformation.[298] AI accountability mechanisms help mitigate these risks, ensuring that AI systems are designed and deployed equitably, safeguarding users' rights and promoting societal trust in AI technologies.[299]

Moreover, adherence to accountability not only addresses local challenges but also enhances Indian developers' global competitiveness. As global frameworks such as the EU AI Act and other regulatory initiatives emphasize accountability, Indian developers who embed these practices will be better positioned to comply with international standards and expand into global markets.[300] Upholding accountability ensures that Indian AI systems are ethically sound and reliable, helping the country emerge as a responsible leader in the global AI landscape.

Sectoral examples of what AI Accountability means

- In the **Healthcare** sector, AI systems are increasingly used to diagnose diseases (e.g., detecting cancer in medical imaging). Accountability in this context means ensuring that these systems provide transparent and explainable diagnoses, and any decision made by the AI can be traced back to clear, ethical guidelines.[301] For example, if an AI misdiagnoses a patient, there should be a clear understanding of how that decision was made and with whom the liability lies. Thus, healthcare providers must have protocols in place for validating AI outputs.[302]
- In the **Financial services** sector, AI is used to assess creditworthiness and approve loans.[303] Accountability here means ensuring that the algorithms do not discriminate based on gender, race, or socioeconomic background. Developers need to ensure that models are regularly audited for biases and fairness, and provide clear explanations for why a particular loan was approved or denied.[304]
- In the **Education** sector, AI tools are being used to grade student exams and assignments. Accountability means ensuring that these systems are fair, transparent, and free from bias.[305] Developers need to ensure that AI grading systems can explain the rationale behind a score and allow students to contest the results.[306] Continuous monitoring is required to ensure that no group of students is unfairly advantaged or disadvantaged.
- In the **Agricultural** sector, AI models are used to predict crop yields and optimize irrigation schedules. Accountability means that these systems must provide transparent and reliable predictions, especially in regions where livelihoods depend on accurate forecasts. If an AI system predicts a wrong yield, it could result in significant economic loss for farmers, so systems need to be continuously audited for accuracy and provide clear explanations for their predictions.[307]

Key challenges that developers may face in ensuring Accountability in AI systems

Developers face several challenges in ensuring accountability as AI systems grow more complex and are used in critical sectors like healthcare, criminal justice, and finance. Standardized testing methods, such as Massive Multitask Language Understanding (MMLU)[308] and Bias Benchmark for QA (BBQ)[309], focus on accuracy but often fail to address broader accountability concerns. Generative AI models, in particular, struggle with context, leading to biased or inaccurate outputs. Techniques like transfer learning and domain adaptation can help, but the lack of transparency in many AI systems makes accountability difficult.

Ensuring accountability requires continuous monitoring, auditing, and updating to address evolving real-world data.[310] This is especially challenging for smaller companies and startups, as these processes are resource-intensive, requiring both technical expertise and financial investment. Many startups, especially in India, lack the resources to fully address accountability gaps, increasing the risk of errors or bias.

The absence of standardized accountability benchmarks complicates this further. While global regulations like the EU's AI Act stress the importance of accountability, developers often need to set their own criteria, leading to inconsistent practices.[311] This makes it harder to compare models or ensure uniform accountability across the industry.

Accountability in AI is not just a technical issue but a societal one, with real-world legal, ethical, and financial consequences.[312] To address these challenges, developers need comprehensive frameworks that cover the entire AI lifecycle, from data collection to deployment and monitoring.[313] Tools like model cards and explainability reports offer transparency and help stakeholders understand AI decision-making.[314] However, these tools must be scalable and accessible, especially for smaller organizations.

Collaboration across the AI industry is essential. Open-source platforms and efforts to standardize accountability metrics can provide smaller players with the resources they need to maintain responsible AI systems.[315]

How to ensure Accountability?

AI System level

- Leverage multiple metrics to balance error types and user experiences.[316]
- Regularly analyze raw data, addressing errors like missing values and ensuring diverse representation.[317]
- Clearly communicate model limitations and educate users on constraints for better feedback.[318]
- Conduct unit and integration tests for both ML and system components.[319] Continuously monitor real-world feedback and update models using the HEART framework[320] or blocklisting[321]. Allocate time for issue resolution, balancing immediate fixes with long-term strategies for lasting improvements.[322]
- Leverage open-source AI systems to enhance transparency and public trust.
- Open-sourcing model code, training datasets, and documentation enables external scrutiny, facilitates independent audits, and supports bias detection and error correction by a broader ecosystem of researchers, developers, and civil society stakeholders.
- Open systems also promote knowledge sharing, help benchmark best practices, and encourage the adoption of fairness-enhancing methods across the AI lifecycle.

Individual and Team level	● Provide employees, from executives to R&D teams, with ethical AI training. Use workshops, online courses, and expert insights to align skills with legal standards and values.[323] ● Define clear roles for AI governance to boost accountability. Involve the board, system owners, developers, and an AI governance committee to ensure ethical alignment.[324] ● Schedule regular external audits to evaluate AI systems' performance and ethical alignment. Share audit findings with stakeholders to drive improvements.[325] ● Promote ethical AI by rewarding responsible practices in performance reviews. Balance business goals with ethics and hold key decision-makers accountable for responsible AI use.[326]
Organizational level[327]	● Adopt AI models like LLMs and LVMs, recognizing their limitations in data quality and training methods. Implement A/B and stress testing to address shortcomings and build stakeholder trust.[328] ● Identify AI failure modes by examining data dependencies, prompt issues, and infrastructure limits. Enable user feedback on failures to boost transparency and accountability.[329] ● Use risk frameworks and bias detection tools to manage AI risks. Document residual risks and continuously optimize tools for safer, more effective AI solutions.[330] ● Ensure diverse representation, including individuals with disabilities, in data and design. Align with legal frameworks like the Persons with Disabilities Act to meet diverse user needs.[331] ● Clearly define responsible AI practices in End User Licensing Agreements and terms of use.[332] Establish reporting mechanisms for misuse and outline stakeholder responsibilities for transparency and legal compliance. ● Showcase responsible AI initiatives as a market differentiator. Emphasize ethical practices and risk mitigation to enhance brand reputation and align with emerging regulations.[333] ● Establish clear grievance redressal mechanisms to ensure accountability in AI-driven decision-making. Users may have the right to request human intervention if they believe an AI decision is incorrect, unfair, or biased, with a multi-tier review process enabling escalation to human reviewers when necessary.[334] AI platforms should provide accessible grievance portals where users can lodge complaints, track their status, and receive timely resolutions. For high-risk AI applications, regulatory bodies may mandate third-party AI audits or establish AI ombudsman bodies to oversee complaints related to bias, discrimination, or unfair outcomes.[335]

Checklist

Conception, design and development

1. **Evaluate data sources and problem formulation**

   - Document all data sources and specify their influence on model decisions.
   - Implement oversight mechanisms for data handling (e.g., access and change logs).
   - Appoint designated personnel responsible for oversight and governance.
   - Define accountability frameworks based on levels of user intervention.
   - Classify the AI system as autonomous, human-in-the-loop, or hybrid.

2. **Model suitability and contextual factors**

   - Ensure model selection aligns with the use case in terms of transparency, efficiency, and accuracy.
   - Incorporate contextual decision-making factors into algorithm design.
   - Remove redundant features to improve processing efficiency.
   - Choose appropriate data normalization or alternative pre-processing techniques.

3. **Performance and error detection**

   - Address assumptions related to hardware, calibration, and environmental variability.
   - Use tools to detect and mitigate bias and performance errors.

4. **Error minimisation and risk management**

   - Develop processes to mitigate harm from erroneous AI behaviour.

Deployment

5. **Monitoring and ethical oversight**

   - Implement systems to track ethical decisions during deployment.
   - Ensure documentation is accessible for future teams and external audits.
   - Establish communication channels for stakeholders regarding AI decisions and limitations.
   - Tailor user-facing explanations based on the technical proficiency of end users.
   - Embed feedback loops for continuous system monitoring and improvement.
   - Set up grievance redressal mechanisms to allow users and impacted individuals to report issues, contest decisions, and seek remediation where appropriate.

Security

What does Security mean?

Security in the context of AI refers to protecting AI systems and the data they process from potential threats and vulnerabilities.[336] It involves securing the development and deployment of AI technologies to safeguard sensitive information and ensure the integrity and confidentiality of AI-driven processes.[337] As AI becomes integral to industries, securing these systems against cyberattacks[338] like adversarial attacks, data poisoning, and other model-specific threats are crucial.[339] Unlike traditional software security, which focuses on protecting code and data,[340] AI security must also address AI models' unique risks, requiring specialized techniques and tools.

So, why should developers care about Security?

AI security is essential to maintaining the reliability and trustworthiness of AI systems, which are increasingly being deployed in critical sectors.[341] Developers must recognize that any breach or compromise in AI security can lead to serious consequences, such as unauthorized access to sensitive data, disruption of services, or manipulation of AI-driven decisions.[342] These risks are not hypothetical; they are real and growing as AI becomes more embedded in high-stakes environments such as defense, healthcare, and finance.[343]

For Indian developers, the stakes are particularly high. India is positioning itself as a global leader in AI development, with significant investments in sectors like defense, where AI is used for surveillance, autonomous systems, and cybersecurity.[344] A failure in AI security could compromise national security and erode India's credibility as a reliable AI development hub. This has implications for India's ambitions to lead in AI innovation on the global stage.[345]

Additionally, the Indian government's focus on AI for governance, defense, and economic growth means that developers working in this space must be mindful of the geopolitical and national security implications.[346] As AI is adopted for critical infrastructure and defense projects, even small vulnerabilities can have disproportionate impacts, leading to significant operational and security risks.[347]

In the global context, Indian developers must also be aware of evolving international standards for AI security.[348] For instance, the EU's AI Act imposes strict security and transparency standards for high-risk AI applications.[349] Developers must comply with evolving regulations and standards like ISO/IEC 27001 to ensure market access and trust with international partners.[350]

What are the common risks associated with AI Security?

Identifying and mitigating various risks and attacks that can compromise security is crucial to ensuring the integrity and reliability of AI models and systems. Each of the risks mentioned below highlights the need for robust security measures to protect AI systems from evolving threats.

Data Security Risks	Vulnerabilities exist throughout the AI pipeline from data collection to storage and transfer. Attackers can exploit these points to gain unauthorized access, alter data, or inject malicious inputs.[351] For example: A healthcare AI system is trained on patient records to predict disease outbreaks. However, weak encryption during data transfer allows attackers to intercept and alter the patient records. The modified data skews the model's analysis, resulting in incorrect outbreak predictions. This leads to delayed responses, worsening public health outcomes and undermining trust in the healthcare system.[352]
Data Poisoning	Attackers manipulate input data, such as images or text, to deceive AI models into making incorrect predictions. This undermines the trustworthiness of AI systems.[353] For example: A facial recognition system at an airport is designed to identify potential security threats. Attackers subtly manipulate images by adding imperceptible noise, causing the AI to misclassify certain individuals as "safe" even though they pose a threat. This manipulation could allow dangerous individuals to bypass security checks, posing a serious risk to public safety.[354]
Input Manipulation	By altering real-time inputs like sensor readings or user data, attackers can influence AI outputs, potentially leading to system failures or incorrect decisions.[355] For example: An autonomous vehicle uses AI to navigate based on real-time sensor data. An attacker spoofs the vehicle's sensors to make it believe that there is an obstruction ahead when there isn't one. This manipulation causes the car to unexpectedly stop in the middle of a busy highway, leading to a traffic accident and potential loss of life.[356]
Model Inversion Attacks	Attackers may reverse-engineer AI models to infer sensitive training data, posing significant privacy risks.[357] For example: A fitness app uses AI to recommend personalized health plans based on users' biometric data. Through model inversion, attackers reverse-engineer the AI system to infer sensitive information about individual users, such as health conditions or physical traits. This invasion of privacy could lead to targeted scams or discrimination based on the inferred data.[358]

Membership Inference Attacks	Adversaries can determine if specific data points were included in a model's training dataset, potentially revealing private information.[359] For example: An e-commerce company uses AI to personalize product recommendations for users. An adversary conducts a membership inference attack and discovers that certain individuals' purchasing data was included in the training dataset. This could lead to the revelation of private shopping habits, such as medical supplies or personal products, violating user privacy.[360]
Model Poisoning	This is when an adversary manipulates a trained model's parameters/ weights to cause it to behave in some undesirable fashion.[361] For example: A machine learning model is trained to distinguish between images of cats and dogs. An adversary computes the gradient of the loss function to slightly adjust the pixels of a correctly classified cat image. This subtle change causes the model to misclassify the altered image as a dog, even though it still looks like a cat to the human eye. This manipulation exploits the model's vulnerabilities and leads to incorrect predictions.[362]
Supply Chain Attacks	These target the software and hardware used in AI systems, potentially introducing malicious code or compromising third-party services.[363] For example: A national security agency uses AI for intelligence analysis. However, an attacker compromises a third-party AI library used in the system, embedding malicious code. This code exfiltrates sensitive intelligence data once deployed, resulting in the exposure of critical national security information.[364]
Other exploratory attacks	Attackers probe AI systems to uncover vulnerabilities or proprietary information, which may be used in future attacks.[365] For example: A financial AI system is used to detect fraudulent transactions. An attacker continuously probes the system with different transaction patterns to learn its decision-making process. Over time, the attacker identifies weaknesses in the system, eventually crafting fraudulent transactions that bypass the detection mechanisms, leading to financial loss for the institution.[366]

Key challenges that developers may face in ensuring Security in AI systems

Developers face significant challenges when securing AI systems, particularly with the growing complexity of modern AI models, such as those based on deep learning or generative techniques.[367] These systems are often opaque, making it difficult to fully understand or predict how they will respond to malicious inputs.[368] This "black box" nature presents a challenge for security because vulnerabilities can go unnoticed, exposing systems to attacks like adversarial manipulation or data poisoning.[369]

One key issue is the trade-off between security and system performance.[370] More powerful AI models, such as those used in generative AI or for complex decision-making tasks, require substantial computational resources.[371] This makes them more prone to attacks that exploit weaknesses in model architecture or computational constraints, such as resource exhaustion or denial-of-service attacks.[372] Ensuring robust security while maintaining system efficiency becomes increasingly difficult as AI systems scale.[373]

Interoperability with existing systems poses another major challenge.[374] Developers must integrate AI-driven security tools with legacy systems, which may not have been designed to handle the complexities of modern AI security threats.[375] This requires careful consideration of compatibility and seamless integration to avoid creating new vulnerabilities.[376] As AI systems are incorporated into more critical sectors, ensuring security without disrupting established workflows is essential.[377]

The scalability of AI security is also critical.[378] Maintaining consistent security across all environments becomes challenging as AI systems process larger and more diverse datasets.[379] AI models must be resilient to cyberattacks at scale, including the risk of data poisoning, where malicious inputs corrupt the training data, leading to faulty outcomes or decisions in production. Ensuring that AI security systems can scale while protecting sensitive data is a pressing issue, particularly in industries like healthcare and finance.[380]

From a regulatory standpoint, complying with national and international security standards adds complexity. While addressing AI-specific risks, developers must ensure their AI systems meet frameworks like the DPDP Act, SOC 2, ISO, or GDPR.[381] The challenge lies in navigating these evolving regulations, which often lag behind the rapid development of AI technologies, making it difficult to ensure compliance without hindering innovation.

A particularly difficult challenge is managing data privacy in AI systems. Generative AI models, for example, may inadvertently expose sensitive data or re-identify individuals from anonymized datasets.[382] Ensuring that these models handle data securely while still producing accurate results is a constant balancing act. The increasing sophistication of re-identification techniques has made traditional anonymization methods less effective, requiring more advanced privacy-preserving methods to protect personal information.[383]

Finally, overreliance on AI security systems can create blind spots. Organizations may place too much trust in AI-driven security solutions, neglecting crucial human elements like employee training or incident response planning.[384] This can leave gaps that attackers can exploit, particularly in fast-moving threat environments where AI systems alone may not be agile enough to detect or respond to new attack vectors.[385]

Sectoral examples of what Security in AI means

- In the **Healthcare** sector, AI detects early signs of Alzheimer's disease by analyzing brain scans. Machine learning models analyze medical images to identify subtle patterns indicative of cognitive decline, helping doctors make earlier diagnoses.[386] Ensuring security in these AI systems is crucial to prevent unauthorized access to sensitive patient data and maintain the integrity of diagnostic processes.[387]
- In the **Financial services** sector, AI detects credit card fraud by analyzing transaction patterns.[388] The system flags anomalies that indicate potential fraudulent activities. Securing these AI systems is critical to prevent hackers from manipulating data or bypassing fraud detection mechanisms, which could result in financial losses for customers and institutions.[389]
- In the **Education** sector, AI is used in adaptive testing platforms that adjust the difficulty of questions based on student performance. These systems rely on secure algorithms to ensure that student data is

protected, and that the testing environment remains fair and free from manipulation or cheating.[390]

- In the **Agriculture** sector, AI can help farmers optimize irrigation by analyzing weather data and soil moisture levels. The AI models recommend optimal watering schedules to improve crop yield and conserve resources. Security is essential to protect these systems from cyberattacks that could lead to incorrect farming decisions, causing potential crop damage.[391]

How to ensure Security?

To ensure security in AI systems, developers may integrate various practices throughout the design, development, and deployment phases.[392] These practices help enhance the protection of AI models, ensuring their resilience against attacks and adherence to security standards.

- **Adopt secure coding practices** to identify and eliminate vulnerabilities that cyber attackers could exploit. Regular code reviews are essential for detecting known, unknown, and unexpected vulnerabilities, including security exploits and data leaks. Secure coding is critical in safeguarding sensitive data and ensuring the overall security of applications.

- **Implement access control measures** by integrating advanced authentication and authorization techniques, such as Multi-Factor Authentication (MFA), Role-Based Access Control (RBAC), and Attribute-Based Access Control (ABAC). These measures provide additional layers of security, ensuring that only authorized users have access to sensitive data and systems.

- **AI Security tools, techniques and frameworks**: AI-specific security tools, like adversarial robustness toolkits, help detect vulnerabilities and provide defences against attacks, such as evasion or poisoning, enhancing the overall security of AI systems.

    - Utilize techniques such as adversarial training to defend against adversarial attacks. Adversarial training involves retraining models with adversarial examples, teaching them to ignore noise and focus on unperturbed features.[493]

    - By minimizing a model's privileges, AI developers can prevent it from autonomously taking actions that may lead to errors or security breaches, such as connecting to email facilities that could inadvertently send sensitive information.

    - Some useful open-source tools include NB Defense[394] (which helps integrate security measures early in the development lifecycle), Adversarial Robustness Toolbox[395] (which provides a range of pre-built attacks and defenses to protect models from adversarial threats like evasion and poisoning), Garak[396] (which scans LLMs for vulnerabilities such as hallucinations and prompt injection), and Google's Secure AI Framework[397] (which helps safeguard algorithms and environments through encryption, anomaly detection, and ongoing assessments).

    - Developers can also perform comprehensive privacy and security risk analyses for every AI initiative. These analyses should inform the development of security and privacy controls based on protection goals such as Confidentiality, Integrity, and Availability (CIA), as well as privacy goals like Unlinkability, Transparency, and Intervenability, referencing standards like ISO/IEC TR 27562:2023 for detailed guidance.[398]

    - Developers can enlist international standards that can help foster global interoperability while ensuring security such as the ISO/IEC 42001, a management system standard that provides guidelines for managing AI systems within organizations. Verification by an independent assessor ensures customers using models/ products/ services of the responsible development, deployment and operation of AI models.

- **Adopt a secure development program** by integrating security practices into the AI software development lifecycle. Leverage existing secure software development methodologies to encompass AI-specific considerations, including secure development training, code review processes, security requirements, secure coding guidelines, threat modeling for AI-specific threats, static analysis tooling, dynamic analysis tooling, and penetration testing.[399]
- **Implement organizational measures** to protect sensitive information, assign accountability, and conduct regular risk assessments.[400]
  - Establish robust data security protocols, including encryption and regular audits, to ensure compliance with relevant regulations. These measures promote transparency and accountability, encouraging the implementation of rigorous data handling practices and thorough documentation of AI model development processes. Regular audits provide insights into adherence to security standards and highlight areas for improvement.[401]
  - Integrate AI tools with existing security infrastructure (e.g., SIEM, IDS), and include AI-specific security measures throughout the lifecycle, such as model parameters, data, and third-party assets.
  - Regularly train employees on security protocols and incident response planning to ensure preparedness for evolving threats, including adversarial attacks and data breaches.
- **Continuous monitoring and validation**: Monitor AI systems for performance metrics, compliance with relevant regulations, and output accuracy. Regularly test AI behavior against varied datasets to detect performance issues and security vulnerabilities. This ongoing validation ensures that models remain resilient against changes in real-world conditions or potential attacks.
- **Ensuring human oversight and implementing guardrails in the form of rules** can help detect unwanted model behavior, allowing for the correction of or halting the model's decision-making process. However, defining the exact properties of wanted versus unwanted behavior can be challenging, limiting the effectiveness of guardrails and human oversight. Further, adding red-teaming activities can help find flaws in a systemic fashion.[411]
- **Adopt open-source AI frameworks strategically to enhance security through transparency**, peer review, and collaborative threat detection. While open-source AI can expose models to potential misuse if not properly managed, it also allows developers to identify vulnerabilities early, implement shared security standards, and strengthen defenses against attacks like adversarial manipulation and data poisoning.[402] By leveraging open-source security tools and best practices, organizations can improve response times to emerging threats, ensure compliance with evolving security protocols, and build more resilient AI systems.

Checklist

Conception, design and development

1. **Data collection and storage**
   - Encrypt all training data to ensure confidentiality and integrity.
   - Perform integrity checks to verify data authenticity.
   - Collect data only from trusted and verified sources.
2. **Secure data storage and transfer**
   - Store and manage data using secure, access-controlled systems.
   - Implement secure transfer protocols to protect data in transit.
3. **Data privacy compliance**
   - Ensure data collection complies with applicable privacy laws (e.g., DPDPA).
   - Apply data minimization principles during collection and storage.
4. **Model development, security testing, and secure coding**
   - Follow secure coding practices (e.g., input validation, secure API usage).
   - Conduct regular code reviews and vulnerability scans.
   - Use regularization or other techniques to prevent overfitting.
   - Avoid overly complex model architectures to enhance security.
5. **Adversarial robustness**
   - Simulate adversarial attacks (e.g., evasion, input manipulation).
   - Train models using adversarial and out-of-distribution examples.
6. **Data handling during training stage**
   - Implement access controls to protect training datasets.
   - Secure training environments (e.g., hardware, network) to reflect deployment conditions.
7. **Open-source AI security integration**
   - Strategically adopt open-source AI frameworks to enhance transparency, peer review, and threat detection.
   - Use community-reviewed tools and libraries to detect vulnerabilities and enforce secure standards.
   - Monitor for potential misuse or tampering in open-source components.
   - Leverage open-source security tools to improve threat response times and model resilience.

Deployment

8. **Monitoring, auditing, and incident response**
   - Deploy monitoring tools to track performance and detect threats.
   - Enable real-time alerts for anomalies or adversarial behavior.
   - Establish a documented incident response plan for AI-specific threats.
   - Prepare recovery strategies for swift remediation post-breach.
9. **Regular security audits**
   - Conduct periodic security audits against standards and best practices.
   - Review access logs to detect unauthorized or suspicious activity.
10. **Organizational measures and human oversight**
    - Update internal security policies to reflect evolving AI risks.
    - Implement behavioral guardrails to flag abnormal model behavior.
    - Set override mechanisms for human intervention during critical risks.
11. **Stakeholder communication and transparency**
    - Set up communication channels to inform stakeholders of security events
    - Educate users on security protocols and their roles in AI system safety.
12. **Model re-training**

- Retrain or fine-tune models based on real-world feedback and emerging threats.

Annexure – Case Studies

Case Study 1 : Cough Against TB tool

Introduction

Tuberculosis (TB) is a major public health problem, especially in countries like India where access to healthcare can be limited in rural and low-income areas. Traditional diagnostic tools, such as chest X-rays and laboratory tests, are effective but require specialized equipment and trained personnel. However, these are often unavailable in remote locations.

WadhwaniAI, a non-profit organization that develops AI solutions for social good, created an innovative tool called Cough Against TB. This tool utilizes AI to assist in screening individuals for Pulmonary TB based on the sound of their coughs and their self-reported symptoms. It works on all Android smartphones (in both online and offline settings) and is designed for use by healthcare workers in the field.

This case study explains how the tool works, the challenges faced by the developers, and the techniques they used to ensure that the tool is fair, accurate, scalable, and respectful of user privacy.

What is the Cough Against TB tool?

Cough Against TB is a mobile-based application that uses a three part AI architecture:

1. **Cough detector model** – This model identifies and isolates cough signal from audio recordings.
2. **TB inference, ensemble model** – This model analyzes the cough sound, along with information about symptoms (such as fever or weight loss), to provide an inference on the likelihood of TB .

If the tool predicts that the person is Presumptive for Pulmonary TB, they are referred to a hospital or clinic for further testing. The tool is not intended to diagnose TB but rather to help healthcare workers identify individuals who require more detailed testing, further enhancing the screening capacity in the field.

Key challenges and how they were solved

1. **Ensuring fairness**: AI models can sometimes work better for some groups of people than others. For example, a model trained primarily on data from adult men may not perform well for children or women. This is called bias.

   To reduce bias and make the model fair for all users, WadhwaniAI used the following methods:

   - **Balanced data collection**: Cough samples were collected from individuals of diverse ages, genders, and locations to ensure the model had a diverse training dataset.
   - **Cohort-wise evaluation**: The model's performance was tested separately for different groups to check if it was equally accurate for all.
   - **Adversarial training**: The model was trained to focus solely on features related to TB, while ignoring those related to gender or age. This was achieved using a second model, referred to as an "adversary", which attempted to infer a person's group affiliation from the main model's outputs. If the adversary was successful, the main model was penalized during training.

- **Domain adaptation**: This technique enabled the model to perform well across various locations and recording conditions, such as quiet clinics and noisy outdoor settings.
- **Facility vs. community distribution shift**: Most data was collected from health facilities where TB prevalence is higher, but deployment is intended for community settings where prevalence is lower and conditions differ. This mismatch can cause performance issues. To mitigate this, algorithmic interventions were implemented to ensure the model works effectively across both settings.

2. **Improving accuracy**: AI models can sometimes make mistakes. In healthcare, it is important to avoid both false positives (wrongly saying someone might have TB) and false negatives (missing a real case of TB). To improve accuracy, WadhwaniAI added a human-in-the-loop system. This means that a trained healthcare worker reviews the AI's output and makes the final decision. This increased the system's accuracy by approximately 9%.

   In addition to reviewing model output, human oversight is also used during data collection. For instance, if a cough recording is poor in quality, the healthcare worker may ask the individual to cough again. This ensures high-quality inputs are fed into the model. WadhwaniAI also followed defined criteria to decide which samples could be included in the dataset. This helped reduce noise and prevent poor-quality or non-representative data from affecting model performance.

3. **Making the tool work on simple phones**: Most AI models are large and need powerful computers to run. But in rural areas, healthcare workers often use basic smartphones. To solve this, WadhwaniAI used a method called model pruning. This involves removing non-essential parts of the model. As a result:

   - The Cough Detector Model was reduced from 43 megabytes to 1.2 megabytes.
   - The TB Inference Model was reduced from 43 megabytes to 6 megabytes.

   These smaller models could now run on low-cost Android phones, even without internet access. This enables the deployment of the solution in diverse settings and environments (edge deployment).

4. **Protecting user privacy**: WadhwaniAI ensured that users gave informed consent before recording their coughs. The consent process was intentionally designed to be user-friendly and accessible. Legal language was avoided to ensure that individuals could understand what they were agreeing to. The system also removes personal or identifiable information from the collected data as part of standard privacy practices.

5. **Monitoring the model in real-world use**: AI models can sometimes perform well during testing but exhibit different behavior after deployment. To manage this, WadhwaniAI created a centralized dashboard. This system enables the team to monitor how the models perform in various locations and whether they remain fair and accurate over time. If problems are identified, the models can be retrained or adjusted accordingly. Post-deployment tracking is also part of WadhwaniAI's broader commitment to responsible AI. It ensures that issues such as performance drift, rising error rates, or emergent biases can be identified and corrected promptly.

Key takeaways

Cough Against TB is a good example of how AI can be used responsibly in public health. The tool has been tested in various field settings and adapted to real-world challenges.

1. **Fairness matters:** AI models should work equally well for all groups. This requires careful data collection and testing.
2. **Human oversight enhances reliability**: Allowing healthcare workers to verify AI outputs makes the system safer and more effective.
3. **Lightweight models are important**: In low-resource settings, models must run on basic phones without internet.
4. **Privacy must be respected**: People should be informed about how their data is being used and provide clear consent.
5. **Ongoing monitoring is essential**: AI models should be tracked and updated after deployment to ensure they continue to perform well.

Case Study 2: Prevention of Adverse TB Outcomes (PATO)

The Problem

India sees approximately 2.5 million TB cases annually, with 7% resulting in adverse treatment outcomes—either mortality or loss to follow-up (LTFU). LTFU refers to patients discontinuing treatment, often due to stigma, medication side effects, travel burdens, or costs. These patients are at risk of developing drug-resistant TB, posing a threat to both individual and public health.

While several government interventions exist to improve treatment outcomes (e.g., direct benefit transfers (DBT), these adverse outcomes persist. The challenge is to identify high-risk patients at the time of treatment initiation to enable targeted, timely interventions.

The AI Solution

To address this, an AI model—referred to as PATO (Prediction of Adverse Treatment Outcomes)—was developed and deployed across 16 Indian States/UTs. The model uses historical patient data from Ni-kshay, the Government of India's national TB database, to predict which patients are at high risk of adverse outcomes at the start of their treatment.

Key features

- **Training data**: The model is trained on past TB treatment records using over 40 structured variables, including demographic data, clinical indicators, comorbidities (e.g., HIV, diabetes), and whether the patient is enrolled in the direct benefit transfer scheme.
- **Binary classification**: The model predicts a binary outcome (high-risk or not) combining LTFU and death. This formulation simplifies deployment since both events require similar interventions, such as intensified home visits and phone follow-ups.
- **Integrated deployment**: Health workers upload patient data weekly. The model processes this data and generates risk lists sent directly to relevant field staff via an app, guiding real-time interventions.
- **Evaluation Metric**: Recall while targeting a fraction that can be tuned based on health-worker availability (currently 35%) of total patients is used to evaluate model effectiveness, reflecting performance in low-resource, real-world settings where only a limited number of patients can be flagged for follow-up.
- **Privacy protocols**: Patient data is de-identified, encrypted, and stored on secure servers within India. Access is restricted and gated with strict upload protocols that enhance both privacy and data quality. The system rejects incomplete forms and mandates critical fields to run predictions, thereby strengthening the data quality within the ecosystem.
- **Baselines & Benchmarking**: The team designed rule-based models simulating the best checklist-based government guidelines, and hybrid models combining those with insights from the literature and the data. While these baselines achieved ~50% recall, the AI model reached ~70%, significantly outperforming traditional approaches.
- **Transparency Tools**: The model highlights feature importance to support interpretability and policy feedback, helping identify which variables are most predictive of adverse outcomes.

- **Fairness Checks**: Extensive cohort-wise performance evaluations revealed strong performance across most cohorts and some disparities, particularly better performance on male patients compared to female patients. Post-hoc fairness interventions have been tested and planned for implementation to improve equity across gender and geography.
- **Temporal robustness & Drift Handling**: The model uses time-based data splits, and is evaluated for robustness across different timeframes. Quarterly retraining to remain responsive to shifts in patient behavior and healthcare practices has been tested and planned for implementation.
- **Telemetry**: The team has set up dashboards measuring key aspects of deployment - live patient numbers across high-risk, low risk, outcomes and interventions, as well as running confusion matrices and data drift. These can further be filtered using categories such as notification time and location, enabling granular tracking of the deployed solution.

Trade-offs and Challenges

- **Accuracy vs. Fairness**: While the model performs well overall, addressing performance disparities across sensitive cohorts is an ongoing process. Fairness-enhancing algorithms now balance model accuracy across groups, reducing gender-based gaps, however, they may somewhat reduce overall performance.
- **Scalability vs. Complexity**: A binary classification system enables easy deployment but may lose granularity in risk assessment. The simplicity was a strategic choice to prioritize scale and actionability on the field.
- **Privacy vs. Interpretability**: Although patient data is de-identified and securely stored, further enhancements like differential privacy or federated learning are worth exploring, though these may cause challenges to model interpretability.
- **Data Quality as a Feature**: The privacy-preserving upload process unintentionally improved data completeness and consistency, making better predictions possible and contributing to stronger government data systems.

Key takeaways

The PATO initiative represents a scalable and responsible application of AI in public health. Key takeaways for developers include:

1. **AI needs thoughtful problem formulation**: Grouping death and non-adherence into one outcome class and choosing a metric aligned with public health needs was a result of stakeholder consultations and reflects real-world intervention design.
2. **Fairness is a dynamic process**: Fairness audits are only the beginning; meaningful equity requires iterative improvement and model adjustments.
3. **Simple designs enable impact at scale**: Deployable, actionable models often outperform theoretically optimal but complex alternatives.
4. **Data systems improve with AI**: Enforcing input quality and secure protocols strengthens not just the model but the entire data ecosystem.
5. **Model monitoring builds trust**: Regular retraining and forward-looking validation ensure that AI systems evolve alongside real-world conditions.

Case Study 3: Shishu Maapan

The Problem

Accurate anthropometric measurements in the first 42 days of a newborn's life are essential for early detection of growth issues and developmental risks. However, frontline health workers in India often face major challenges:

- Broken equipment such as faulty spring balances.
- Inconsistent measurement techniques due to limited training.
- Cultural taboos and community hesitancy around measuring newborns.
- Incentive-based misreporting, such as avoiding low birth weight labels that demand extra follow-ups.

Errors in measurement are common, up to 180 grams on average, a significant margin given that 2.5 kg is the clinical threshold for low birth weight. Traditional systems lack both the precision and the checks needed to address these problems effectively.

The AI Solution

To tackle this, an AI-based video measurement application called Shishu Maapan was introduced. Built for use on low-end Android smartphones, the solution allows frontline health workers to record short videos of newborns, which are then analyzed to estimate weight, length, head circumference, chest circumference, and middle upper arm circumference.

This tool is integrated into India's Home-Based Newborn Care (HBNC) framework and has been designed specifically for low-resource, high-need settings.

Key features

- **Video-Based Measurement**: Health workers capture a short video (15-20 seconds) of the newborn capturing multiple angles. The AI model processes the video to estimate parameters like weight, length, chest circumference, and head circumference.
- **Offline Capability**: The model, compressed from 120 MB to 32 MB, can run without internet access, enabling use in rural and remote areas.
- **Tamper-Proof Digital Records**: All data is entered digitally, with no opportunity for alteration. This helps prevent underreporting of low-birth-weight cases.
- **Privacy by Design**:
    - Data is collected only by trusted community workers.
    - Informed consent is obtained in local languages.
    - No video is stored locally on phones.
    - Video data is encrypted during transmission and storage, then deleted.
- **Human-in-the-Loop Monitoring**: Medical officers oversee the deployment and usage of the app, ensuring alignment with community needs and health system priorities.
- **Fairness through Contextualization**:
    - The model is fine-tuned using 1,000 videos per geography to adapt to local body proportions.
    - Performance is assessed across weight bins and gender cohorts.
    - A third-party evaluation is being conducted to ensure objectivity and fairness.

- **User-Centric Design**:
    - The app is made for non-tech-savvy users.
    - Features included in app dashboard, growth chart, HBNC visit scheduler and real time feedback support to check cropping and ensuring correct placement of baby in the video frame while capturing the video.

Challenges and Trade-offs

- **Privacy vs. Usability**: Avoiding local storage and encrypting all data improves privacy but limits local processing, requiring cloud-based operations which may not always be reliable in low-connectivity regions.
- **Fairness vs. Generalization**: Tailoring the model to specific geographies improves fairness, but requires localized data collection and validation.
- **Accuracy vs. Compression**: Reducing model size is critical for offline use, but can reduce precision. Real-time feedback helps counteract these limitations.
- **Adoption vs. Complexity**: Many FLHws face app fatigue due to the number of digital tools they're expected to use. Shishu Maapan's simplified interface is a deliberate response to this challenge.

Key takeaways

Shishu Maapan is a textbook example of applying responsible AI in maternal and child health. It balances competing demands of privacy, fairness, usability, and cost-effectiveness, while fitting seamlessly into India's public health infrastructure.

1. **Responsible AI starts with design**: Building privacy, consent, and interpretability into the system ensures it is accepted and trusted by communities.
2. **Fairness requires contextualization**: A one-size-fits-all AI model won't work across India's diverse geographies. Localization improves accuracy and trust.
3. **Compression unlocks scale**: Reducing model size without compromising utility is key to scaling in low-resource settings.
4. **Monitoring matters**: Human oversight from trained officers ensures that AI use remains aligned with broader health goals.
5. **AI should adapt to users, and not the other way around**: Tools must meet health workers where they are, not where engineers want them to be.

Case Study 4: Krishi Saathi

The Problem

Indian farmers frequently face critical knowledge gaps on weather, pest management and practices, mandi prices, and crop insurance that impact yield and income. Although such information exists, it is often dispersed, complex, or not updated in real-time. Farm Tele Advisors (FTAs) at Kisan Call Centers (KCCs) try to bridge this gap, but typically rely on manual searches and general internet queries, leading to:

- Long call wait times, often exceeding 4 minutes per farmer.
- Inconsistent or incomplete answers.
- Delayed decision-making for time-sensitive agricultural activities like sowing or harvesting.

The system needed a transformation to speed up responses, improve accuracy, and maintain regulatory compliance.

The AI Solution

To address this, Krishi Saathi, an AI-powered agriculture conversational chatbot, was developed. It helps FTAs provide reliable, timely responses to farmer queries. The solution is now operational in 17 centers across 14 Indian states and Union Territories (UTs).

Key features

- **Multilingual text input support:**
  - Farmers speak their queries in their local languages when interacting with the FTAs over the IVR call.
  - FTAs then enter the query based on the interaction with the farmers by typing it down, the chatbot then generates the contextually relevant answer in English, and translates it into farmers' local language, and summarises it if required.
- **Public and vetted data sources:** All information comes from government-vetted databases and sources, including:
  - Weather forecasts from the Indian Meteorological Department (IMD) and current weather information from Google Weather.
  - Market pricing from eNAM and pest advisory from national and state-level sources.
- **Modular and agentic architecture:** The chatbot uses various data, models and API services:
  - OpenAI's gpt-4o-mini LLM for RAG-QA and agentic framework (English-only).
  - Language translation through Bhashini model to provide the advisories in farmers' local (Indian) language.

  This separation helps trace and correct errors, enhancing modularity and debugging.
- **On-Premises deployment for compliance:** The chatbot application is hosted on the government's secured server.
- **Strict prompting and domain guardrails:**
  - The application is configured to respond only to the agriculture-related queries.
  - It refuses to answer non-agriculture domain irrelevant questions. It responds with "I don't know" to all such queries.
- **Real-time SMS integration:** Five-day weather forecasts are sent directly to farmers via SMS in their local language, enabling proactive farm planning.

Human-in-the-Loop evaluation

- **Component-level evaluation:** Every major component such as language translation, weather and market price data, LLM and agent's accuracy is individually reviewed by expert evaluators. This granular review ensures that each part of the system is aligned with what farmers actually need.
- **Manual and automated checks:** Custom metrics to balance the factual accuracy with coverage (e.g., all important events included, no false positives (i.e., hallucination) information). LLM outputs are reviewed with tailored prompts and tested for edge cases to identify potential failure modes.
- **Evaluation design as a core task:** Unlike traditional ML models, LLMs require specialized prompt design and scenario crafting to be properly evaluated. Significant efforts are put into making this rigorous and human-guided.

Challenges and trade-offs

- **Accuracy vs. Accessibility:** Responses are generated in English, which simplifies quality control. Translations are layered afterward, which can introduce errors but allows better root cause analysis.
- **Speed vs. Trust:** AI speeds up responses dramatically, but only human validation ensures quality and clarity.
- **Coverage vs. Compliance:** Limiting the chatbot to vetted agricultural topics ensures reliability but restricts flexibility to address off-topic queries.

Key Takeaways

Krishi Saathi bridges the critical information gap between Indian farmers and timely agricultural advice. By combining multilingual support, curated government data, and modular AI components, it delivers accurate and actionable information quickly and securely.

1. **Domain-specific AI needs curation:** Using only public, government-approved data boosts reliability and mitigates misinformation risks.
2. **Modularity enables debugging:** Separating language translation, various data service integrations via agentic architecture from core QA functions helps isolate and fix issues faster.
3. **Security and compliance must be native:** Application hosting on the secured servers, strict prompting with guardrails, and logging build trust and ensure regulatory alignment.
4. **Human evaluation is non-negotiable:** Evaluating LLM generated responses requires careful scenario design and human effort across each system component.
5. **AI needs to fit the infrastructure:** In the future, by using voice input, modular translation, and SMS-based outputs, Krishi Saathi is designed to meet all kinds of farmers' requirements.

Case Study 5: Digital Green

The Problem

Smallholder farmers in India often lack timely, local, and actionable information on agriculture practices, crop planning, and risk management. Government and NGO extension systems struggle to deliver personalized advice at scale, especially in low-resource settings with limited digital access and language diversity.

The AI Solution

Digital Green is developing an AI assistant that provides personalized agricultural advisory to farmers. It uses generative AI, powered by a Retrieval-Augmented Generation (RAG) architecture to ground LLM responses in trusted, domain-specific knowledge, which here is validated content from local partners, government advisory systems, and prior interactions with farmers.

How It Works

Initially built around peer-to-peer video-based learning, Digital Green transitioned during the pandemic to smartphone and chatbot-based tools. Their current solution, Farmer Chat, a mobile app, enables farmers or other intermediaries to ask agricultural queries via voice, text and/or image. Before launching the app, Digital Green tested its AI integration through proof-of-concept pilots with government extension workers on messaging platforms like Telegram and whatsapp. These pilots laid the foundation for Farmer Chat, which now has over 250,000 downloads and 2.8 Mn conversations within a year of launch.

This generative AI approach is central to Digital Green's business model: the use of LLMs has enabled the organization to scale rapidly, delivering nuanced, regionally appropriate answers across diverse cropping systems. To ensure factual accuracy, responses are constrained to a carefully curated knowledge base, avoiding obscure or hallucinated outputs that could harm farmer decision-making. This knowledge base includes government verified content as well as content from partners like Food and Agriculture Organization of the UN (FAO) and International Crops Research Institute for the Semi-Arid Tropics (ICRISAT).

Key features

- **Query intake:** Users submit questions through a conversational interface. Community intermediaries often facilitate this for voice inputs in vernacular languages.
- **Query filter:** There are modules to detect personal information in the text and/or image which detects and removes personal information before processing it further.
- **Query orchestrator agent:** Identifies the intent of the query, extracts agriculture entities like crop, concern to be able to route it to relevant pipelines to generate appropriate responses. If not able to extract the relevant entities to satisfactorily respond, prompts back users with options.
- **Tool calling agent:** identifies and call appropriate tools like weather, market prices, RAG endpoint (documents/ videos) etc to generate the context specific response.
- **RAG pipeline:** A search is conducted over a curated knowledge base of verified content from government, NGOs, and public content. Relevant documents and/or videos are then retrieved and passed for response generation using LLM. The LLM generates a response grounded in the retrieved material. If insufficient data is found, the

model makes sure to say "I don't know" or polite variation instead of hallucinating. System prompts are embedded in the LLM calls to prevent unintended disclosures or leakage of sensitive information. These prompts act as a safeguard against overly permissive or "friendly" model behavior that could inadvertently surface personal data.

- **Data minimization:** The system does not pass user identities or PII. Training and refinement use only anonymized, aggregate data from past interactions. To put it in GDPR terms, in deployments with government partners, Digital Green acts as a data processor, rather than a controller. This role limits their data obligations and aligns with their strategy of only using query-specific context for response generation, without associating queries with individual identities.
- **Farm Stack integration:** In government deployments, the system plugs into Farm Stack, which is an open-source infrastructure enabling secure data access without centralized storage of personal farmer data. Farm Stack was originally created for other use cases and has since been repurposed by Digital Green as a data integration and management tool. It allows government partners to transfer data securely while avoiding data centralization, mining, or persistent tracking of farmers.

Challenges and Tradeoffs

- **Compliance vs. Usage:** The more tightly Digital Green adheres to privacy-by-design principles (such as not storing queries or identities), the more they restrict their ability to build personalized, or optimized AI services. This tradeoff is especially acute in deployments with government partners, where Digital Green operates purely as a data processor and is bound by stricter compliance constraints.
- **Privacy vs. Personalization:** While personalization improves advice quality, storing and passing individual farmer profiles and related personal information raises data protection concerns. Digital Green follows data minimisation and stores any PII through encryption in partitioned data tables. For the downstream process of analysis and training, all the conversation logs are passed through PI removal service to mask/ remove personal data and anonymize all training inputs. But this can limit the ability to build persistent user histories.
- **Accuracy vs. Transparency/Explainability:** The RAG architecture enhances factual accuracy but introduces complexity. It is difficult to transparently explain how responses are generated to end-users, especially in low-literacy contexts. The focus on factual correctness can lead to questions not covered by content to not be responded to.
- **Language and Context Limitations:** Handling diverse regional languages, especially through voice-based inputs, presents significant challenges. Variations in phrasing, dialect, and the way farmers frame questions can affect both retrieval and generation. This requires continuous tuning of the system to ensure relevance and clarity in responses.

Key Takeaways

1. **Ground models in trusted knowledge:** Grounding generative AI in domain-specific knowledge is essential for safety and trust. Digital Green RAG to constrain LLM outputs to vetted sources like government advisories and research institutions. This mitigates hallucinations and ensures relevance, especially in high-stakes use cases.

2. **Design for privacy by default:** Privacy-preserving design can coexist with high-quality advisory, especially if you're willing to have appropriate filters and minimise the collection of information to what is required for that specific query. By working only with anonymized, aggregate data for further review, analytics and fine tuning, Digital Green limits privacy risks. However, this comes at the cost of long-term user profiles, requiring alternate strategies to improve relevance.

3. **Compliance as a constraint:** Compliance constraints are not just legal; they shape system capabilities. Operating as a data processor in government deployments means Digital Green cannot define how data is stored or reused. Developers must factor in these legal roles early, as they can limit optimization, training, and personalization potential.

4. **Responsible AI needs infrastructure, not just model tuning:** Open-source tools like Farm Stack help enforce data minimization and secure sharing, especially in public-sector settings. Developers should invest as much in systems design as they do in model performance.

5. **Design for linguistic diversity:** Language diversity is not just a translation problem and is rather a design challenge. Variations in how farmers phrase questions across dialects and regions affect both retrieval accuracy and generation quality. Ongoing fine-tuning, community input, and localized evaluation are crucial for sustained performance.

6. **Refusal as a safety feature:** Refusal to respond can be a feature, not a failure. Explicitly training the model to acknowledge gaps in its knowledge reinforces safety and user trust. This is especially important in contexts where incorrect advice can have economic or health consequences.

References

1. NASSCOM, ‘The Developer’s Playbook for Responsible AI in India’ (November 2024) https://nasscom.in/ai/pdf/the-developer’s-playbook-for-responsible-ai-in-india.pdf
2. Justice K S Puttaswamy (Retd.) and Anr. v. Union of India [2017] 10 SCC 1
3. Justice K S Puttaswamy (Retd.) and Anr. v. Union of India [2017] 10 SCC 1 https://cdnbbsr.s3waas.gov.in/s3ec0490f1f4972d133619a60c30f3559e/documents/aor_notice_circular/43.pdf
4. Digital Personal Data Protection Act [2023] https://www.meity.gov.in/static/uploads/2024/06/2bf1f0e9f04e6fb4f8fef35e82c42aa5.pdf
5. Digital Personal Data Protection Act [2023], Schedule
6. Digital Personal Data Protection Rules [2025] https://www.meity.gov.in/static/uploads/2025/11/53450e6e5dc0bfa85ebd78686cadad39.pdf.
7. UK ICO, ‘Guidance on AI and Data Protection’ (15 March 2023) https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/
8. Dutch Data Protection Authority, ‘Second (interim) advice on Supervisory Structure’ (12 June 12 2024) https://www.autoriteitpersoonsgegevens.nl/documenten/ai-verordening-tweede-tussenadvies-ap-rdi
9. DSK, Guidance on AI and Data Protection (6 May 2024) https://www.datenschutzkonferenz-online.de/media/oh/20240506_DSK_Orientierungshilfe_KI_und_Datenschutz.pdf
10. Personal Data Protection Commission Singapore, ‘Advisory Guidelines on Use of Personal Data in AI Recommendation and Decision Systems’ (1 March 2024) https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/advisory-guidelines-on-the-use-of-personal-data-in-ai-recommendation-and-decision-systems.pdf
11. CNIL, ‘Recommendations on the Development of AI systems’ (7 June 2024) https://www.cnil.fr/en/ai-cnil-publishes-its-first-recommendations-development-artificial-intelligence-systems
12. DPDP Act, Section 3
13. DPDP Act, Section 3(c)(ii)
14. Ministry of Electronics and Information Technology, India AI Governance Guidelines (IndiaAI, 5 November 2025) https://indiaai.gov.in/article/india-ai-governance-guidelines-empowering-ethical-and-responsible-ai
15. DPDP Act, Section 2(x)
16. DPDP Act, Section 3(b)
17. DPDP Act, Section 2(i)
18. DPDP Act, Section 2(k)
19. DPDP Act, Section 4(1)

20. DPDP Act, Section 5(3)
21. DPDP Rules, Rule 3
22. DPDP Act, Section 6(1)
23. DPDP Act, Section 6(4)
24. DPDP Act, Section 5(2)
25. DPDP Act, Section 7
26. DPDP Act, Section 17
27. DPDP Act, Section 17
28. DPDP Act, Section 8(4)
29. DPDP Act, Section 8(5)
30. DPDP Act, Section 8(6); DPDP Rules, Rule 7
31. DPDP Act, Section 8(7)
32. DPDP Act, Section 8(10)
33. DPDP Rules, Rule 14
34. DPDP Rules, Rule 9
35. DPDP Rules, Rule 14
36. DPDP Act, Section 2(z) read with Section 10(1)
37. DPDP Act, Section 10(2)(a)
38. DPDP Act, Section 10(2)(b)
39. DPDP Act, Section 10(2)(c)(i)
40. DPDP Rules, R ule 13(3)
41. DPDP Rules, R ule 13(2)
42. DPDP Act, Section 9(1)
43. DPDP Act, Section 9(3)
44. DPDP Act, Section 8(2)
45. DPDP Act, Section 16(1)
46. DPDP Rules, Rule 15.
47. DPDP Rules, Rule 13(4).
48. The Reserve Bank of India, ‘Storage of Payment System Data’ (2018; 2019) https://www.rbi.org.in/Scripts/NotificationUser.aspx?Id=11244&Mode=0; https://www.rbi.org.in/commonman/english/scripts/FAQs.aspx?Id=2995
49. DPDP Act, Section 11(1)
50. DPDP Act, Section 12
51. DPDP Act, Section 14

52. DPDP Act, Section 13

53. DPDP Act, Section 15

54. DPDP Act, Section 18

55. DPDP Act, Section 27(1)(b)

56. DPDP Act, Schedule

57. DPDP Act, Section 33(2)(e)

58. DPDP Act, Section 37(1)(b)

59. DPDP Act, Section 31

60. DPDP Act, Section 32

61. Tal Roded and Peter Slattery, 'What drives progress in AI? Trends in Data' (FutureTech, 19 March 2024) https://futuretech.mit.edu/news/what-drives-progress-in-ai-trends-in-data-#:~:text=AI%20models%20need%20data,changed%20between%202010%20and%202023

62. Scott Hatfield 'The Bigger the Better: Why does more data make AIs better?' (Medium, 26 February 2023) https://medium.com/@Toglefritz/the-bigger-the-better-why-does-more-data-make-ais-better-eb0aafa2f725

63. 'AI in the Banking Sector', Infosys BPM https://www.infosysbpm.com/blogs/bpm-analytics/fraud-detection-with-ai-in-banking-sector.html

64. 'What is generative AI?', Accenture https://www.accenture.com/in-en/insights/generative-ai

65. Preeti S Chauhan and Nir Kshetri, 'The Role of Data and Artificial Intelligence in Driving Diversity, Equity, and Inclusion' (ResearchGate, April 2022) https://www.researchgate.net/publication/359882110_The_Role_of_Data_and_Artificial_Intelligence_in_Driving_Diversity_Equity_and_Inclusion#:~:text=Chauhan%20and%20Kshetri%20%5B62%5D%20emphasize,%5B62%5D.%20…

66. 'What is an AI model?', IBM https://www.ibm.com/topics/ai-model

67. DPDP Act, Section 3 (a)

68. 'What is personal data?', European Commission https://commission.europa.eu/law/law-topic/data-protection/reform/what-personal-data_en#:~:text=Personal%20data%20is%20any%20information,person%2C%20also%20constitute%20personal%20data

69. 'What is considered personal data under the EU GDPR?', GDPR.EU <https://gdpr.eu/eu-gdpr-personal-data/#:~:text=There%20are%20more%20factors%20to,reasonably%20access%20from%20another%20source

70. 'What are the identifiers and related factors?', UK Information Commissioner's Office https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/what-are-identifiers-and-related-factors/ ; 'Article 29 Working Party documents', European Commission https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pd

71. 'Data Aggregation', IBM (2021) https://www.ibm.com/docs/en/tnpm/1.4.2?topic=data-aggregation

72. Example adapted from PDPC Singapore, ‘Advisory Guidelines on Key Concepts in the Personal Data Protection Act’, (16 May 2022) https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/ag-on-key-concepts/advisory-guidelines-on-key-concepts-in-the-pdpa-17-may-2022.pd

73. Information Commissioner’s Office ‘What are the identifiers and related factors? ’, UK Information Commissioner’s Office https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/what-are-identifiers-and-related-factors/ ; European Commission, Article 29 Working Party, ‘Opinion 4/2007 on the concept of personal data’ https://www.pdp.ie/docs/1030.pdf

74. Ministry of Electronics and Information Technology, ‘Report by the Committee of Experts on Non-Personal Data Governance Framework’ (, 2020) https://ourgovdotin.wordpress.com/wp-content/uploads/2020/07/kris-gopalakrishnan-committee-report-on-non-personal-data-governance-framework.pdf

75. Ministry of Electronics and Information Technology, ‘Report by Committee of Experts on Non-Personal Data Governance Framework’ (2020) <https://ourgovdotin.wordpress.com/wp-content/uploads/2020/07/kris-gopalakrishnan-committee-report-on-non-personal-data-governance-framework.pdf

76. Ministry of Electronics and Information Technology, ‘Report by Committee of Experts on Non-Personal Data Governance Framework’ (2020) https://ourgovdotin.wordpress.com/wp-content/uploads/2020/07/kris-gopalakrishnan-committee-report-on-non-personal-data-governance-framework.pdf

77. The Personal Data Protection Bill 2018, Section 3(3) https://www.meity.gov.in/writereaddata/files/Personal_Data_Protection_Bill,2018.pdf

78. PDPC Singapore, ‘Guide to Basic Anonymisation’ (2024) https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/guide-to-basic-anonymisation-(updated-24-july-2024).pdf

79. ‘Anonymisation’, ICO UK https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/

80. Office of the Victorian Commissioner, ‘The Limitations of De-Identification’ (May 2018) <https://ovic.vic.gov.au/privacy/resources-for-organisations/the-limitations-of-de-identification-protecting-unit-record-level-personal-information/

81. Agencia Espanola Proteccion Datos, ‘Anonymization III: The risk of re-identification’ (February, 2023) https://www.aepd.es/en/prensa-y-comunicacion/blog/anonymization-iii-risk-re-identification

82. James Bennett and Stan Lanning, ‘The Netflix Prize’, (University of Illinois Chicago) https://www.cs.uic.edu/~liub/KDD-cup-2007/proceedings/The-Netflix-Prize-Bennett.pdf

83. Dan Jackson, ‘The Netflix Prize’ (Thrillist, 8 July 2017) https://www.thrillist.com/entertainment/nation/the-netflix-prize

84. ‘Cautionary Tales: Learning from the Frontlines of Data Privacy and Security’, Subsalt https://www.getsubsalt.com/post/cautionary-tales-learning-from-the-frontlines-of-data-privacy-and-security---part-3-of-3?utm_campaign=security-funded-138-total-eclipse-of-the-ipo-spark&utm_medium=newsletter&utm_source=ReturnOnSecurity#:~:text=Following%20the%20release%20of%20the,Internet%20Movie%20Database%20(IMDb)

85. PDPC Singapore, ‘Guide to Basic Anonymisation’ (2022) https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/guide-to-basic-anonymisation-(updated-24-july-2024).pdf

86. Introduction to anonymisation’, UK ICO https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/introduction-to-anonymisation/

87. Single Resolution Board (SRB) vs. European Data Protection Supervisor (EDPS), C-413/23 P https://curia.europa.eu/juris/document/document.jsf?text=&docid=303863&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=16803864

88. Government of United Kingdom, Department of Science, Innovation and Technology, ‘Cyber Security Risks to Artificial Intelligence’ (15 May 2024) https://www.gov.uk/government/publications/research-on-the-cyber-security-of-ai/cyber-security-risks-to-artificial-intelligence

89. Oxford Internet Institute, ‘On Personal Data , Forgiveness, and the Right to Be Forgotten’ (10 March 2015) https://www.oii.ox.ac.uk/news-events/videos/on-personal-data-forgiveness-and-the-right-to-be-forgotten/

90. PDPC Singapore, ‘Advisory Guidelines on Use of Personal Data in AI Recommendation and Decision Systems’ (1 March 2024) https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/advisory-guidelines-on-the-use-of-personal-data-in-ai-recommendation-and-decision-systems.pdf

91. European Commission, ‘Guidance on the Regulation on a Framework for the Free Flow of Non-Personal Data in the European Union’ (29 May 2019) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:250:FIN

92. PDPC Singapore, ‘Guide to Basic Anonymisation’ (2022) https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/guide-to-basic-anonymisation-(updated-24-july-2024).pdf

93. Dmytro Romanchenko, ‘ Creating AI Software: The Critical Role of Data Collection’ (Synd/Code, 19 March 2024) https://syndicode.com/blog/importance-of-data-ai-development/> accessed on 16 September 2024

94. DPDP Act, Section 3(c)(ii)

95. Singapore Personal Data Protection Act, Section 17 read with Part 2(1) of the First Schedule https://sso.agc.gov.sg/Act/PDPA2012

96. ibid

97. Ministry of Electronics and Information Technology, India AI Governance Guidelines (IndiaAI, 5 November 2025) https://indiaai.gov.in/article/india-ai-governance-guidelines-empowering-ethical-and-responsible-ai

98. PDPC Singapore, ‘Advisory Guidelines on Key Concepts in the Personal Data Protection Act’ (16 May 2022) https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/advisory-guidelines/ag-on-key-concepts/advisory-guidelines-on-key-concepts-in-the-pdpa-17-may-2022.pdf

99. DPDP Act, Sections 5, 6 and 7

100. DPDP Act, Section 5(1)

101. DPDP Act, Section 4(1)(a)

102. DPDP Act, Section 5(1)

103. DPDP Rules, Rule 3

104. DPDP Rules, Rule 3

105. DPDP Rules, Rule 3

106. DPDP Rules, Rule 3

107. Garante per la protezione dei dati personali, ‘ChatGPT: Italian SA to lift temporary limitation if OpenAI implements measures’ (12 April 2023) https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9874751#english

108. DPDP Act, Section 6

109. DPDP Act, Section 5(1)(i)

110. DPDP Act, Section 8(7)

111. DPDP Rules, Third Schedule

112. DPDP Rules, R ule 8(2)

113. DPDP Rules, Third Schedule

114. DPDP Act, Section 6(4)

115. DPDP Act, Section 6(4)

116. DPDP Act, Section 6(5)

117. Lauren Merk and Bailey Sanchez, ‘FTC requires algorithmic disgorgement as a COPPA remedy for the first time’ (Future of Privacy Forum, 14 March 2024) https://fpf.org/blog/ftc-requires-algorithmic-disgorgement-as-a-coppa-remedy-for-first-time/

118. DPDP Act, Section 9(1)

119. ‘Children’s Online Privacy Protection Rule: A Six-Step Compliance Plan for Your Business’, Federal Trade Commission https://www.ftc.gov/business-guidance/resources/childrens-online-privacy-protection-rule-six-step-compliance-plan-your-business#step4

120. DPDP Act, Section 9(3), DPDP Rules, R ule 12 and Fourth Schedule

121. DPDP Rules, Fourth Schedule

122. DPDP Rules, Rule 4

123. DPDP Rules, First Schedule

124. DPDP Rules, First Schedule

125. Ministry of Electronics and Information Technology, ‘Electronic Consent Framework — Technology Specifications’ https://dla.gov.in/sites/default/files/pdf/MeitY-Consent-Tech-Framework%20v1.1.pdf

126. DPDP Act, Section 7(a)

127. DPDP Act, Section 7(b)(i)

128. DPDP Act, Section 7(b)(ii)

129. DPDP Act, Section 7(c)

130. DPDP Act, Section 7(d)

131. DPDP Act, Section 7(e)

132. DPDP Act, Section 7(f)

133. DPDP Act, Section 7(h)

134. DPDP Act, Section 7(i)

135. DPDP Act, Section 7(a)

136. Singapore’s Personal Data Protection Act (PDPA), 2012, Section 15

137. Singapore’s Personal Data Protection Act (PDPA), 2012, Section 15 (a)

138. DPDP Act, Section 2(za)

139. Autoriteit Persoonsgegevens, ‘Dutch DPA Scraping is Almost Always Illegal’ (2024) https://www.autoriteitpersoonsgegevens.nl/themas/algoritmes-ai/algoritmes-ai-en-de-avg/regels-bij-gebruik-van-ai-algoritmes

140. DSK, ‘DSK Guidance on AI and Data Protection’ (2024) https://www.datenschutzkonferenz-online.de/media/oh/20240506_DSK_Orientierungshilfe_KI_und_Datenschutz.pdf

141. Mitaksh Jain, ‘Why did Canadian privacy authorities expand their investigation into OpenAI’s ChatGPT?’ (Medianama, 26 May 2023) https://www.medianama.com/2023/05/223-why-canadian-authorities-investigation-into-chatgpt-2/

142. Øyvind H. Kaldestad, ‘Legal Complaint against Meta’s use of personal content for AI training’ (Forbrukerrådet, 6 June 2024) https://www.forbrukerradet.no/siste-nytt/digital/klager-inn-metas-bruk-av-personlig-innhold-til-ki-trening/legal-complaint-against-metas-use-of-personal-content-for-ai-training

143. DPDP Act, Section 2(i)

144. DPDP Act, Section 2(k)

145. European Union’s GDPR, Articles 4(7) and 4(8)

146. Singapore’s Personal Data Protection Act (PDPA), 2012, Section 2

147. ‘What are the accountability and governance implications of AI?’, UK ICO https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/what-are-the-accountability-and-governance-implications-of-ai/#how-shouldweunderstand

148. CNIL, ‘Determining the legal qualification of AI system providers’ (7 June 2024) https://www.cnil.fr/en/determining-legal-qualification-ai-system-providers#:~:text=The%20qualification%20of%20the%20AI,and%20means%20of%20the%20processing

149. European Data Protection Board, ‘Guidelines on the concepts of data controller and processor in the GDPR’, para 30 https://www.edpb.europa.eu/system/files/2023-10/EDPB_guidelines_202007_controllerprocessor_final_en.pdf

150. DPDP Act, Section 11(1)(a)

151. Bryan Casey, Ashkon Farhangi and Roland Vogl, ‘Rethinking Explainable Machines: The GDPR’s Right to Explanation Debate and the Rise of Algorithmic Audits in Enterprise’ (2019) https://btlj.org/data/articles2019/34_1/04_Casey_Web.pdf

152. DPDP Act, Section 11(1)(b)

153. DPDP Act, Section 12

154. 'How do we ensure individual rights in our AI systems?', UK ICO https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-individual-rights-in-our-ai-systems/#howdoindividualrightsapply

155. ibid

156. Brandon LaLonde, 'Explaining model disgorgement' (IAPP, 13 December 2023) https://iapp.org/news/a/explaining-model-disgorgement

157. Joshua A. Goland, ' Algorithmic Disgorgement: Destruction of Artificial Intelligence Models as the FTC's Newest Enforcement Tool for Bad Data' (Richmond Journal of Law and Technology, 2023) https://jolt.richmond.edu/files/2023/03/Goland-Final.pdf

158. Allessandro Achille, Michael Kearns and Carson Klingenberg, 'AI model disgorgement: Methods and choices' (2024) https://www.pnas.org/doi/10.1073/pnas.2307304121

159. Fabian Pedregosa and Eleni Triantafillou, 'Announcing the first Machine Unlearning Challenge' (Google Research, 2023) https://research.google/blog/announcing-the-first-machine-unlearning-challenge/

160. CNIL, 'Respect and facilitate the exercise of data subjects' rights' (2 July 2024) https://www.cnil.fr/en/respect-and-facilitate-exercise-data-subjects-rights

161. CNIL, 'Respect and facilitate the exercise of data subjects' rights' (2 July 2024) https://www.cnil.fr/en/respect-and-facilitate-exercise-data-subjects-rights

162. DPDP Act, Section 8(10).

163. DPDP Rules, Rule 14(3).

164. DPDP Rules, Rule 14(3).

165. DPDP Act, Section 14

166. DPDP Rules, Rule 14(4).

167. GDPR, Article 22

168. DPDP Act, Section 8(4)

169. DPDP Rules, Rule 14(3)

170. DPDP Rules, Rule 10

171. DPDP Rules, Rule 6

172. GDPR , Article 24 https://gdpr-info.eu/art-24-gdpr/

173. ibid

174. DPDP Act, Section 8(4)

175. GDPR , Article 24 https://gdpr-info.eu/art-24-gdpr/

176. ibid

177. 'AI Accountability', Carnegie Council for Ethics in International Affairs https://www.carnegiecouncil.org/explore-engage/key-terms/ai-accountability#:~:text=AI%2Denabled%20technology%20often%20implicates,the%20tech's%20dynamic%20learning%20potential

178. DSK, 'Position Paper on Organisational and Technical Measures' (6 November 2019) https://www.datenschutzkonferenz-online.de/media/en/20191106_positionspapier_kuenstliche_intelligenz.pdf

179. GDPR, Article 24(2)

180. DPDP Act, Section 10(2)(c)(i)

181. DPDP Act, Section 10(2)(c)(i)

182. Kim Wuyts, 'LINDDUN GO: A Lightweight Approach to Privacy Threat Modeling' (IEEE, 22 October 2020) https://ieeexplore.ieee.org/abstract/document/9229757 .

183. DPDP Rules, Rule 6.

184. DPDP Rules, Rule 6.

185. Murat Durmus, ' An overview of some available Fairness Frameworks and Packages' (Medium, 24 May 2021) https://murat-durmus.medium.com/an-overview-of-some-available-fairness-frameworks-packages-ff22fde9d2f4

186. Steven Umbrello and Ibo Van de Poel, ' Mapping value sensitive design onto AI for social good principles' (Springer Nature, 1 February 2021) https://link.springer.com/article/10.1007/s43681-021-00038-3

187. Reid Blackman, ' Why you need an AI Ethics Committee' (Harvard Business Review, 2022) https://hbr.org/2022/07/why-you-need-an-ai-ethics-committee

188. DPDP Act, Section 10(2)(a)

189. Veda C. Storey, Wei Thoo Yue, J. Leon Zhao and Roman Lukyanenko, 'Generative Artificial Intelligence: Evolving Technology, Growing Societal Impact, and Opportunities for Information Systems Research' (Information Systems Frontiers, 25 February 2025) https://link.springer.com/article/10.1007/s10796-025-10581-7

190. International Association of Privacy Professionals, 'Global AI Law and Policy Tracker' (6 September 2023) https://iapp.org/media/pdf/resource_center/global_ai_law_policy_tracker.pdf

191. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, 'Artificial Intelligence Act' (Official Journal of the European Union, 12 July 2024) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

192. The White House, 'White House Unveils America's AI Action Plan' (23 July 2025) https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence ; https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan/

193. Embassy of the People's Republic of China in India, 'China Announces AI Cooperation Initiative with India' (29 July 2025) https://in.china-embassy.gov.cn/eng/zgxw/202507/t20250729_11679232.htm

194. UAE Artificial Intelligence Office at the Prime Minister’s Office, ‘Towards a Future of Responsible AI’ (February 2024) https://ai.gov.ae/wp-content/uploads/2025/01/Towards-a-Future-of-Responsible-AI-EN-White-Paper.pdf

195. India AI, ‘India AI’ https://indiaai.gov.in/

196. NITI Aayog, ‘Adopting the Framework: A Use Case Approach on Facial Recognition Technolog ’ (November 2022) https://www.niti.gov.in/sites/default/files/2022-11/Ai_for_All_2022_02112022_0.pdf ; NITI Aayog, ‘National Strategy For Artificial Intelligence’ (June 2018) https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf

197. NASSCOM, ‘The Developer’s Playbook for Responsible AI in India’ (November 2024) https://nasscom.in/ai/pdf/the-developer’s-playbook-for-responsible-ai-in-india.pdf

197(A) Ministry of Electronics and Information Technology, ‘India AI Governance Guidelines’ (IndiaAI, 5 November 2025) https://indiaai.gov.in/article/india-ai-governance-guidelines-empowering-ethical-and-responsible-ai

197(B) Reserve Bank of India, ‘Framework for Responsible and Ethical Enablement of AI’ (August 2025) https://rbidocs.rbi.org.in/rdocs/PublicationReport/Pdfs/FREEAIR130820250A24FF2D4578453F824C72ED9F5D5851.PDF

197(C) Securities and Exchange Board of India, ‘Consultation Paper on guidelines for responsible usage of AI/ML In Indian Securities Markets’ (June 2025) https://www.sebi.gov.in/reports-and-statistics/reports/jun-2025/consultation-paper-on-guidelines-for-responsible-usage-of-ai-ml-in-indian-securities-markets_94687.html

198. UNESCO, ‘Recommendation on the Ethics of Artificial Intelligence’ (26 November 2021) https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

199. Emilio Ferrara, ‘Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies’ (MDPI, 26 December 2023)

200. Leonardo Banh and Gero Strobel, ‘Generative artificial intelligence’ (Electronic Markets Volume 33, December 2023) https://scholar.google.com/citations?view_op=view_citation&hl=en&user=yHrEt4wAAAAJ&citation_for_view=yHrEt4wAAAAJ:aqlVkmm33-oC

201. Stanford University Human-Centered Artificial Intelligence, ‘AI’s Fairness Problem :When treating everyone the same is the wrong approach’ (6 February 2025) https://hai.stanford.edu/news/ais-fairness-problem-when-treating-everyone-the-same-is-the-wrong-approach

202. Alycia N. Carey, Xintao Wu, ‘The statistical fairness field guide: perspectives from social and formal sciences’ (AI and Ethics, 2023) https://link.springer.com/article/10.1007/s43681-022-00183-3

203. Campolo, A and others ‘AI Now 2017 Symposium Report; AI Now’ (AI Now Institute at NYU, 2017)

204. Sahil Verma, Julia Rubin, ‘Fairness Definitions Explained’ (ACM/IEEE International Workshop on Software Fairness, 2018) https://fairware.cs.umass.edu/papers/Verma.pdf

205. Najeeb Jebreel, ‘Group Fairness in Machine Learning’ (LinkedIn, April 2024) https://www.linkedin.com/pulse/group-fairness-machine-learning-najeeb-jebreel-ck3jf/

206. Esther Shittu, ‘Differentiating between good and bad AI bias’ (Techtarget, 11 February, 2022) https://www.techtarget.com/searchenterpriseai/feature/Differentiating-between-good-and-bad-AI-bias

207. ‘Credit & Loan Processing: Is AI Biased When Assessing Credit Worthiness?’ (IT Magination, 7 September 2024) https://www.itmagination.com/blog/credit-loan-processing-ai-biased-when-assessing-credit-worthiness

208. Norori N and others, ‘Addressing Bias in Big Data and AI for Health Care: A Call for Open Science’ (Cell Press Patterns , 8 October 2021) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8515002/

209. Mohd Javaid and others , ‘Understanding the Potential Applications of Artificial Intelligence in Agriculture Sector’ (Advanced Agrochrem, 2022)

210. Divya, ‘Harnessing the Power of AI in the Telecom Industry’ (IndiaAI, 11 July 2023) https://indiaai.gov.in/article/harnessing-the-power-of-ai-in-the-telecom-industry

211. Ferdinando Fioretto,’Building Fairness into AI is Crucial – and Hard to Get Right’ (The Conversation, 19 March 2024) https://theconversation.com/building-fairness-into-ai-is-crucial-and-hard-to-get-right-220271

212. N Bhalla, L Brooks and T Leach, ‘Ensuring a ‘Responsible’ AI Future in India: RRI as an Approach for Identifying the Ethical Challenges from an Indian Perspective’ ( ) AI Ethics, 2023) DOI: 10.1007/s43681-023-00370-w

213. A Singhal, N Neveditsin, H Tanveer and V Mago, ‘Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review’ (JMIR Medical Informatics, 3 April 2024) https://pubmed.ncbi.nlm.nih.gov/38568737/

214. ‘Fairness Assessment of Artificial Intelligence Systems’, (Telecommunication Engineering Centre, July 2023) thttps://tec.gov.in/pdf/SDs/TEC%20Standard%20for%20fairness%20assessment%20and%20rating%20of%20AI%20systems%20Final%20v5%202023_07_04.pdf

215. ‘Fairness in AI: Its not One-Size-Fits-All’ (Innodata) https://innodata.com/fairness-in-ai-its-not-one-size-fits-all/

216. ‘Machine Learning Glossary; Responsible AI’ (Machine Learning) https://developers.google.com/machine-learning/glossary/responsible-ai#:~:text=demographic%20parity&text=Contrast%20with%20equalized%20odds%20and,to%20depend%20on%20sensitive%20attributes

217. Jake Silberg and James Manyika, ‘tackling bias in artificial intelligence (and in humans)’ (Mc Kinsey Global Institute, 6 June 2019) https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/tackling-bias-in-artificial-intelligence-and-in-humans

218. Dino Pedreschi and others, ‘Meaningful Explanation of Black Box AI Decision Systems’, (Proceedings of the AAAI Conference on Artificial Intelligence, 17 July 2019) https://ojs.aaai.org/index.php/AAAI/article/view/5050

219. Carmine Ferrar and others, ‘Fairness-aware machine learning engineering: how far are we?’ (Empirical Software Engineering, 24 November 2023) https://link.springer.com/article/10.1007/s10664-023-10402-y

220. Nabeel Rehman and others, ‘The Psychology of Resistance to Change: The Antidotal Effect of Organizational Justice, Support and Leader-Member Exchange’ (Frontiers in Psychol-

ogy, 2 August 2021) https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.678952/full

221. D Rao, ‘Fairness in AI Systems – Everything You Need to Know!’ (Persistent, 21 May 2021) https://www.persistent.com/blogs/fairness-in-ai-systems/

222. James Holdsworth, ‘What is AI Bias?’ (IBM, 22 December 2023) https://www.ibm.com/topics/ai-bias

223. Emilio Ferrara, ‘The Butterfly Effect in Artificial Intelligence Systems: Implications for AI Bias and Fairness’ (SSRN, 26 October 2023) https://doi.org/10.2139/ssrn.4614234]

224. Alexandra Jonker and Julie Rogers, ‘What is Algorithmic Bias’ (IBM) https://www.ibm.com/think/topics/algorithmic-bias

225. J M Alvarez, A B Colmenarejo, A Elobaid et al, ‘Policy Advice and Best Practices on Bias and Fairness in AI’ (2024) 26 Ethics and Information Technology 31 https://doi.org/10.1007/s10676-024-09746-w accessed 6 September 2024.

226. Bogina, V., Hartman, A., Kuflik, T. and others, . ‘Educating Software and AI Stakeholders About Algorithmic Fairness, Accountability, Transparency and Ethics’ (International Journal of Artificial Intelligence in Education, April 2021) https://doi.org/10.1007/s40593-021-00248-0

227. Charfaoui Y, ‘Resampling to Properly Handle Imbalanced Datasets in Machine Learning’ (Comet, 22 September 2023) https://www.comet.com/site/blog/resampling-to-properly-handle-imbalanced-datasets-in-machine-learning/#:~:text=Resampling%20changes%20the%20dataset%20into,main%20methods%3A%20Oversampling%20and%20Undersampling.

228. National Institute of Standards and Technology, , ‘Towards a Standard for Identifying and Managing Bias in Artificial Intelligence’ (US Department of Commerce, 2022) https://www.dwt.com/-/media/files/blogs/artificial-intelligence-law-advisor/2022/03/nist-sp-1270--identifying-and-managing-bias-in-ai.pdf

229. Resampling means adjusting the data so that all categories (or classes) have equal examples. When training an AI model, if one class has many more examples than another, the model might focus on that larger class and ignore the smaller one.

230. Reweighting helps an AI model focus on underrepresented groups in a biased dataset. By giving more weight to examples from smaller groups and less to those from larger ones, the model learns more fairly and avoids favoring the majority group.

231. Synthetic data generation creates new, artificial examples to boost underrepresented groups in a dataset.

232. Fairness-aware data clustering is a method of grouping data while ensuring that different groups (like gender, race, or other attributes) are treated fairly. It aims to create clusters that are not biased or dominated by one group, helping ensure that the results are balanced and equitable for all groups involved.

233. Mingyang Wan, Daochen Zha, Ninghao Liu and Na Zou, ‘In-Processing Modeling Techniques for Machine Learning Fairness: A Survey’ (Association for Computing Machinery, 2023) https://doi.org/10.1145/3551390

234. Ferrara, C., Sellitto, G., Ferrucci, F. and others . ‘Fairness-aware machine learning engineering: how far are we?’ (Empirical Software Engineering Volume 29, 2024) https://doi.org/10.1007/s10664-023-10402-y

235. Modifying the objective function to include fairness constraints means adjusting the goal of an AI model to not only focus on accuracy but also to treat different groups fairly. It adds rules to ensure the model doesn’t favor one group over another, making the results more balanced and fairer.

236. Adversarial debiasing reduces bias in machine learning by using two models: the main model makes predictions, while the adversary detects bias. The main model is trained to improve accuracy while minimizing bias, leading to fairer outcomes.

237. Equal opportunity in machine learning ensures that all groups have the same true positive rate, meaning qualified individuals from any group are equally likely to receive a positive outcome.

238. Demographic parity ensures that a model’s positive outcomes are equally distributed across different demographic groups, regardless of sensitive attributes like race or gender.

239. ‘Why You Should Know and Care About Algorithmic Transparency’ (Oxford Insights, 6 October 2022) https://oxfordinsights.com/insights/why-you-should-know-and-care-about-algorithmic-transparency/

240. T D Jui and P Rivas, ‘Fairness Issues, Current Approaches, and Challenges in Machine Learning Models’ (International Journal of Machine Learning and Cybernetics 31 January 2024) https://doi.org/10.1007/s13042-023-02083-2

241. Calibration means adjusting the model so individuals with the same score have an equal chance of a positive outcome, regardless of their group, making decisions fairer across all groups.

242. Reject option classification promotes fairness in AI by allowing the model to avoid making decisions when it’s unsure, reducing the chances of biased or inaccurate results.

243. ‘ Transparency’ (INDIAai) https://indiaai.gov.in/ai-standards/transparency

244. ‘Transparency and Explainability’ (OECD.AI) https://oecd.ai/en/dashboards/ai-principles/P7

245. ‘What Is Explainable AI (XAI)?’ (IBM, 29 March 2023) https://www.ibm.com/topics/explainable-ai

246. ‘ The Open Source AI Definition 1.0 RC1’ (Open Source Initiative, 2023) https://opensource.org/ai/drafts/the-open-source-ai-definition-1-0-rc1#398e6e4c-5d98-4796-8bda-5bf97dc04a76

247. ‘ Data Transparency in Open Source AI: Protecting Sensitive Datasets’ (Open Source Initiative, 2023) https://opensource.org/blog/data-transparency-in-open-source-ai-protecting-sensitive-datasets; Digital Public Goods Alliance, ‘DPGA 5 Year Strategy’ (15 November 2023) https://www.digitalpublicgoods.net/dpga-strategy2023-2028.pdf

248. Ramak Molavi Vasse’i and Jesse McCrosky, ‘ AI Transparency in Practice ’ (Mozilla Foundation, March 15, 2023) https://foundation.mozilla.org/en/research/library/ai-transparency-in-practice/ai-transparency-in-practice/

249. Vikas Hassija and others, ‘ Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence’ (Cognitive Computation, 24 August 2023) https://link.springer.com/article/10.1007/s12559-023-10179-8

250. Missing Links in AI Governance” (UNESCO, 2023) https://unesdoc.unesco.org/ark:/48223/pf0000384787

251. DHR-ICMR Artificial Intelligence Cell, ‘Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare’ (Indian Council of Medical Research, 2023) https://main.icmr.nic.in/sites/default/files/upload_documents/Ethical_Guidelines_AI_Healthcare_2023.pdf

252. Sauradeep Bag, ‘AI and Credit Scoring: The Algorithmic Advantage and Precaution’ (ORF, 31 May 2024) https://www.orfonline.org/expert-speak/ai-and-credit-scoring-the-algorithmic-advantage-and-precaution

253. AA Mana and others , ‘Sustainable AI-Based Production Agriculture: Exploring AI Applications and Implications in Agricultural Practices’ (ScienceDirect, 2024) Agricultural Technology

254. Vishal Jain and Archan Mitra, ‘Integrative Hybrid Information Systems for Enhanced Traffic Maintenance and Control in Bangalore: A Synchronized Approach’ (Hybrid Information Systems: Non-Linear Optimization Strategies with Artificial Intelligence 2024)

255. Reid Blackman and Beena Ammanath,‘Building Transparency into AI Projects’ (Harvard Business Review, 20 June 2022) https://hbr.org/2022/06/building-transparency-into-ai-projects

256. S Ali and others, ‘Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence’ (Information Fusion, 2023) https://www.sciencedirect.com/science/article/pii/S1566253523001148?via%3Dihub

257. ‘ Transparency’ (INDIAai) https://indiaai.gov.in/ai-standards/transparency

258. Andrea Ferrario and Michele Loi, ‘How Explainability Contributes to Trust in AI’ (28 January 2022) https://ssrn.com/abstract=4020557 or http://dx.doi.org/10.2139/ssrn.4020557

259. N Omrani and others, ‘To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts’ (Technological Forecasting and Social Change August, 2022) https://doi.org/10.1016/j.techfore.2022.121763

260. J Rueda, JD Rodríguez, IP Jounou and others, ‘“Just” accuracy? Procedural fairness demands explainability in AI-based medical resource allocations’ (AI & Soc, 2024) https://doi.org/10.1007/s00146-022-01614-9

261. Femi Osasona and others , ‘Reviewing the Ethical Implications of AI in Decision Making Processes’ (International Journal of Management & Entrepreneurship Research February 2024) https://doi.org/10.51594/ijmer.v6i2.773

262. ibid.

263. ibid.

264. Luca Longo and others, ‘Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions’ (Information Fusion, June 2024) https://www.sciencedirect.com/science/article/pii/S1566253524000794

265. F. Quaglia, A. Santoro and B. Ciciani, “Conditional checkpoint abort:an alternative semantic for re-synchronization in ccl,” (Proceedings 16th Workshop on Parallel and Distributed Simulation, 2002) pp. 130-137, https://ieeexplore.ieee.org/document/10042109

266. ‘What are dark patterns’, Ministry of Consumer Affairshttps://doca.gov.in/DarkPatternsBusterHackathon/about-us.php#:~:text=Dark%20patterns%20have%20been%20defined,consumer%20autonomy%2C%20decision%20making%20or

267. Vikas Hassija and others, ‘ Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence ’ (Springer International Publishing, 2023) https://link.springer.com/article/10.1007/s12559-023-10179-8

268. Soori M and others , ‘AI-Based Decision Support Systems in Industry 4.0, A Review’ Journal of Engineering and Technology, 2024) https://doi.org/10.1016/j.ject.2024.08.005

269. Wanner J and others, ‘ Stop Ordering Machine Learning Algorithms by Their Explainability! An Empirical Investigation of the Tradeoff Between Performance and Explainability’ (Springer International Publishing, January , 2021) https://link.springer.com/chapter/10.1007/978-3-030-85447-8_22

270. Saeed W and Omlin C, ‘Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities’ (Knowledge-Based Systems Volume 263, March 2023) https://doi.org/10.1016/j.knosys.2023.110273

271. Mitchell AD, Let D and Tang L, ‘ AI Regulation and the Protection of Source Code’ International Journal of Law and Information Technology, 2023) https://academic.oup.com/ijlit/article/31/4/283/7475778

272. Schmitz A and others, ‘ A Global Scale Comparison of Risk Aggregation in AI Assessment Frameworks’ (AI and Ethics Volume 5, 2024) https://link.springer.com/article/10.1007/s43681-024-00479-6

273. Felzmann H and others, ‘ Transparency You Can Trust: Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns’ (Big Data & Society, 2019) https://journals.sagepub.com/doi/full/10.1177/2053951719860542

274. Balasubramaniam N and others, ‘ Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements’ (Information and Software Technology, 2023) https://www.sciencedirect.com/science/article/pii/S0950584923000514

275. Inioluwa Deborah Raji, Andrew Smart and others, ‘Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing’ (FAT* ‘20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2024)

276. Heike Felzmann and others, ‘ Towards Transparency by Design for Artificial Intelligence’ (Science and Engineering Ethics, 2022) https://link.springer.com/article/10.1007/s11948-020-00276-4

277. Athira Nambiar and others , ‘Model-Agnostic Explainable Artificial Intelligence Tools for Severity Prediction and Symptom Analysis on Indian COVID-19 Data’ (Frontiers in Artificial Intelligence, 2023) https://doi.org/10.3389/frai.2023.1272506

278. Shruti Misra, ‘Interpretable AI: Linear Regression’ (Medium, 6 June 2023) https://medium.com/@shrutimisra/interpretable-ai-linear-regression-dbeafbf04db7

279. Dawid Macha, Michał Kozielski, Łukasz Wróbel, Marek Sikora, ‘RuleXAI—A package for rule-based explanations of machine learning model ’ (SoftwareX, 2022) https://doi.org/10.1016/j.softx.2022.101209

280. L-V Herm and others, ‘Stop Ordering Machine Learning Algorithms by Their Explainability! A User-Centered Investigation of Performance and Explainability’ (International Journal of Information Management, April 2022) https://doi.org/10.1016/j.ijinfomgt.2022.102538

281. Sean Tull and others, ‘ Towards Compositional Interpretability for XAI ’ (arXiv.org, June 25, 2024) https://arxiv.org/abs/2406.17583

282. Pantelis Linardatos, Vasilis Papastefanopoulos and Sotiris Kotsiantis, ‘ Explainable AI: A Review of Machine Learning Interpretability Methods’ (Entropy, 2020) https://www.mdpi.com/1099-4300/23/1/18

283. Open Data Institute, ‘Policy Intervention 1: Increase Transparency Around the Data Used to Train AI Models’ (2023) https://theodi.org/news-and-events/blog/policy-intervention-1-increase-transparency-around-the-data-used-to-train-ai-models/; Digital Public Goods Alliance, ‘Core Considerations for Exploring AI Systems as Digital Public Goods ’ (2023) https://www.digitalpublicgoods.net/AI-CoP-Discussion-Paper.pdf .

284. Marina Micheli and others, ‘ The Landscape of Data and AI Documentation Approaches in the European Policy Context’ (Ethics and Information Technology, 2023) https://link.springer.com/article/10.1007/s10676-023-09725-7

285. Margaret Mitchell, Simone Wu and others, ‘Model Cards for Model Reporting’ (Cornell University, 2019) https://arxiv.org/abs/1810.03993

286. ‘The Open Source AI Definition 1.0 RC1’, Open Source Initiative https://opensource.org/ai/drafts/the-open-source-ai-definition-1-0-rc1#398e6e4c-5d98-4796-8bda-5bf97dc04a76 ; ‘Data Transparency in Open Source AI: Protecting Sensitive Datasets’, Open Source Initiative https://opensource.org/blog/data-transparency-in-open-source-ai-protecting-sensitive-datasets

286(A) ‘Difference Between Random Forest and XGBoost’, GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/difference-between-random-forest-vs-xgboost/

287. Sinan Kaplan, Hannu Uusitalo and Lasse Lensu, ‘ A Unified and Practical User-Centric Framework for Explainable Artificial Intelligence’ (Knowledge-Based Systems Volume 283, 2024) https://www.sciencedirect.com/science/article/pii/S0950705123008572

288. Elhassan Mohamed, Konstantinos Sirlantzis and Gareth Howells, ‘ A Review of Visualisation-as-Explanation Techniques for Convolutional Neural Networks and Their Evaluation’ (Displays, 2022) https://www.sciencedirect.com/science/article/pii/S014193822200066X

289. AWS ‘Responsible use of AI guide’ (December, 2024) https://d1.awsstatic.com/products/generative-ai/responsbile-ai/AWS-Responsible-Use-of-AI-Guide-Final.pdf

290. ‘Advancing digital content transparency and authenticity ’, Coalition for Content Provenance and Authenticity https://c2pa.org/

291. Reid A, O’Callaghan S and Lu Y, ‘Implementing Australia’s AI Ethics Principles: A Selection of Responsible AI Practices and Resources’ (Australian Policy Online, 21 June 2023) https://apo.org.au/node/323321

292. ‘Building a Responsible AI: How to Manage the AI Ethics Debate’ , International Organisation for Standardisation https://www.iso.org/artificial-intelligence/responsible-ai-ethics#toc4

293. ‘AI Accountability: Who’s Responsible When AI Goes Wrong?’, Emerge Digital) https://emerge.digital/resources/ai-accountability-whos-responsible-when-ai-goes-wrong/

294. Siddhant Chatterjee, Airlie Hilliard, Ayesha Gulley , ‘The argument for holistic AI Audits’ (OECD.AI Policy Observatory, 19 December 2023) https://oecd.ai/en/wonk/holistic-ai-audits

295. Vikas Hassija and others, ‘Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence’ (Cognitive Computation, 2024) https://link.springer.com/article/10.1007/s12559-023-10179-8

296. Verma R, ‘Global Perspectives on AI Accountability: A Comparative Analysis and India’s Regulatory Landscape’ (Outlook Business, 22 February 2024) https://www.outlookbusiness.com/news/global-perspectives-on-ai-accountability-a-comparative-analysis-and-indias-regulatory-landscape

297. Chatterjee S, ‘Leadership with AI’ (Emeritus India, 28 February 2024) https://emeritus.org/in/learn/responsible-ai/>

298. Kathuria R, Kedia M and Kapilavai S, ‘Implications of AI on the Indian Economy, (NASSCOM, Google and CRIER, July 2020) https://icrier.org/pdf/Implications_of_AI_on_the_Indian_Economy.pdf

299. Chi Y, ‘Safeguarding the Future: Nurturing Safe, Secure, and Trustworthy Artificial Intelligence Ecosystems And the Role of Legal Frameworks ’ (ResearchGate, April 2024) https://www.researchgate.net/publication/379078942_Safeguarding_the_Future_Nurturing_Safe_Secure_and_Trustworthy_Artificial_Intelligence_Ecosystems_and_the_Role_of_Legal_Frameworks

300. Ananya Raj Kakoti and Singh G, ‘The EU AI Act: Implications and Lessons for India’ (Hindustan Times, 29 May 2024) https://www.hindustantimes.com/ht-insight/international-affairs/the-eu-ai-act-implications-and-lessons-for-india-101716968919076.html

301. DHR-ICMR Artificial Intelligence Cell, ‘Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare’ (ICMR, 2023) https://main.icmr.nic.in/sites/default/files/upload_documents/Ethical_Guidelines_AI_Healthcare_2023.pdf

302. Babushkina D, ‘Are We Justified Attributing a Mistake in Diagnosis to an AI Diagnostic System?’ (AI and Ethics Volume 3, 2022) https://link.springer.com/article/10.1007/s43681-022-00189-x

303. Dan Faggella, ‘Artificial Intelligence Applications for Lending and Loan Management’ (Emerj, December 2017) https://emerj.com/ai-sector-overviews/artificial-intelligence-applications-lending-loan-management/

304. Ferrara E, ‘Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies’ (MDPI, 2023) https://www.mdpi.com/2413-4155/6/1/3

305. Hwang G-J and others, ‘Vision, Challenges, Roles and Research Issues of Artificial Intelligence in Education’ 1 (Computers and Education: Artificial Intelligence, 2020) https://www.sciencedirect.com/science/article/pii/S2666920X20300011

306. LaJuan Perronoski Fuller and Bixby C, 'The Theoretical and Practical Implications of OpenAI System Rubric Assessment and Feedback on Higher Education Written Assignments' (American Journal of Educational Research, 2024) https://pubs.sciepub.com/education/12/4/4/index.html

307. Talaviya T. and others, 'Implementation of Artificial Intelligence in Agriculture for Optimisation of Irrigation and Application of Pesticides and Herbicides' (Artificial Intelligence in Agriculture, 2020) https://www.sciencedirect.com/science/article/pii/S258972172030012X

308. 'Challenges in Evaluating AI Systems' (Anthropic, 4 October 2023) https://www.anthropic.com/news/evaluating-ai-systems

309. ibid

310. Reid A, O'Callaghan S and Lu Y, 'Implementing Australia's AI Ethics Principles: A Selection of Responsible AI Practices and Resources' (Australian Policy Online , 21 June 2023) https://apo.org.au/node/323321

311. EU Artificial Intelligence Act- Up-To-Date Developments and Analyses of the EU AI Act https://artificialintelligenceact.eu/

312. Collina L, Mostafa Sayyadi and Provitera M, 'Critical Issues about A.I. Accountability Answered' (California Management Review November 2023) https://cmr.berkeley.edu/2023/11/critical-issues-about-a-i-accountability-answered/

313. De Silva D and Alahakoon D, 'An Artificial Intelligence Life Cycle: From Conception to Production' (Patterns 2022) https://www.sciencedirect.com/science/article/pii/S2666389922000745

314. 'Model Card Toolkit', TensorFlow https://www.tensorflow.org/responsible_ai/model_card_toolkit/guide

315. 'Scaling Responsible AI Solutions Challenges and Opportunities', (OECD.AI Policy Observatory) https://gpai.ai/projects/responsible-ai/RAI05%20-%20Scaling%20Responsible%20AI%20Solutions%20-%20Challenges%20and%20Opportunities.pdf

316. 'Google Responsible AI Practices – Google AI', (Google AI) https://ai.google/responsibility/responsible-ai-practices/?src_trk=em66db367f74d1b3.409348581555557277

317. ibid.

318. Omar Ali and others, 'The Effects of Artificial Intelligence Applications in Educational Settings: Challenges and Strategies' (Technological Forecasting and Social Change, February 2024) https://www.sciencedirect.com/science/article/pii/S0040162523007618

319. ISE M, 'Integration Testing- Engineering Fundamentals Playbook' (Github.io, 2024) https://microsoft.github.io/code-with-engineering-playbook/automated-testing/integration-testing/

320. A user-centered approach for measuring product success based on Happiness, Engagement, Adoption, Retention, and Task success

321. The practice of preventing specific users, domains, or content from accessing a service or platform to enhance security or maintain quality control

322. Oladele S, ‘A Comprehensive Guide on How to Monitor Your Models in Production’ (neptune.ai, 11 August 2022) https://neptune.ai/blog/how-to-monitor-your-models-in-production-guide

323. ‘Responsible AI at Microsoft, Microsoft) https://www.microsoft.com/en-us/ai/responsible-ai

324. ‘AI Governance’, IBM https://www.ibm.com/topics/ai-governance

325. ‘Auditing Artificial Intelligence’, ISACA https://ec.europa.eu/futurium/en/system/files/ged/auditing-artificial-intelligence.pdf

326. ‘Google Responsible AI Practices – Google AI’ (Google AI) https://ai.google/responsibility/responsible-ai-practices/

327. Reid A, O’Callaghan S and Lu Y, ‘Implementing Australia’s AI Ethics Principles: A Selection of Responsible AI Practices and Resources’ (Australian Policy Online 21 June 2023) https://apo.org.au/node/323321

328. ‘Nasscom AI 10 Value Driven Actions’ (Nasscom , 2024) https://nasscom.in/ai/10-value-driven-actions-to-launch-your-responsible-ai-journey/

329. TerryLanfear, ‘Failure Modes in Machine Learning’ (Microsoft , November 2022) https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

330. Yazdi M and others, ‘Navigating the Power of Artificial Intelligence in Risk Management: A Comparative Analysis’ (MDPI , 2024) https://www.mdpi.com/2313-576X/10/2/42

331. Eduard Fosch Villaronga and Poulsen A, ‘Diversity and Inclusion in Artificial Intelligence’ (ResearchGate, July 2022) https://www.researchgate.net/publication/361783049_Diversity_and_Inclusion_in_Artificial_Intelligence

332. Sundaraparipurnan Narayanan, ‘Nasscom AI, 10 Value Driven Actions’ (Nasscom) https://nasscom.in/ai/blog/10-value-driven-actions-to-launch-your-responsible-ai-journey.html

333. Paul B. de Laat, ‘Companies Committed to Responsible AI: From Principles towards Implementation and Regulation?’ (Springer Nature Link, 2021) https://link.springer.com/article/10.1007/s13347-021-00474-3

334. ‘ The Crucial Role of Humans in AI Oversight’, Cornerstone OnDemand <https://www.cornerstoneondemand.com/resources/article/the-crucial-role-of-humans-in-ai-oversight/

335. ‘How Do We Ensure Individual Rights in Our AI Systems?’ , UK ICO https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-individual-rights-in-our-ai-systems/

336. Akash Takyar, ‘Security in AI Development: An Overview’ (LeewayHertz) https://www.leewayhertz.com/security-in-ai-development/

337. ‘Automation’, IBM https://www.ibm.com/topics/automation

338. ‘What is a Cyberattack’, IBM https://www.ibm.com/topics/cyber-attack

339. ‘What is Data Breach’, IBM https://www.ibm.com/topics/data-breach

340. ‘What Is AI Security’ (Hpe https://www.hpe.com/in/en/what-is/ai-security.html

341. Nineta Polemi and others, ‘Challenges and Efforts in Managing AI Trustworthiness Risks: A State of Knowledge’ (Frontiers in Big Data, 2024) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11119750/

342. ‘AI and Privacy: The Privacy Concerns Surrounding AI, Its Potential Impact on Personal Data’ (The Economic Times, 25 April 2023) https://economictimes.indiatimes.com/news/how-to/ai-and-privacy-the-privacy-concerns-surrounding-ai-its-potential-impact-on-personal-data/articleshow/99738234.cms?from=mdr

343. Dr. Nivash Jeevanandam, ‘Impact of AI in the Indian Army’ (INDIAai, 15 January 2024) https://indiaai.gov.in/article/impact-of-ai-in-the-indian-army>

344. Antoine Levesques, ‘Early Steps in India’s Use of AI for Defence’ (IISS, 18 January, 2024) < https://www.iiss.org/online-analysis/online-analysis/2024/01/early-steps-in-indias-use-of-ai-for-defence/>

345. Amlan Mohanty and Shatakratu Sahu, ‘India’s AI Strategy: Balancing Risk and Opportunity’ (Carnegie Endowment for International Peace, 22 Februrary 2024) https://carnegieendowment.org/posts/2024/02/indias-ai-strategy-balancing-risk-and-opportunity?lang=en

346. Sukanya Thapliyal, ‘A Discourse on AI Governance That India Must Shape’ (The Hindu , 4 September 2024) https://www.thehindu.com/opinion/lead/a-discourse-on-ai-governance-that-india-must-shape/article68602063.ece

347. ‘An overview of catastrophic AI risks’ (Centre for AI Safety) https://www.safe.ai/ai-risk

348. Jyoti Panday and Mila T Samdub , ‘ Promises and Pitfalls of India’s AI Industrial Policy’ (AI Now Institute, 12 March 2024) https://ainowinstitute.org/publication/analyzing-indias-ai-industrial-policy

349. Danny Tobey and others, ‘EU Publishes Its AI Act: Key Steps for Organizations ’ (DLA Piper, 12 July 2024) https://www.dlapiper.com/en/insights/publications/ai-outlook/2024/eu-publishes-its-ai-act-key-considerations-for-organizations

350. International Organisation for Standardization ISO/IEC 27001:2022 (2022) https://www.iso.org/standard/27001

351. Kaur R, Dušan Gabrijelčič and Tomaž Klobučar, ‘Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions’ (Information Fusion, September 2023) https://www.sciencedirect.com/science/article/pii/S1566253523001136

352. Yadav N and others, ‘Data Privacy in Healthcare: In the Era of Artificial Intelligence’ (National Library of Medicine, 27 October 2023) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10718098/

353. Blessing Guembe and others, ‘The Emerging Threat of Ai-Driven Cyber Attacks: A Review’ (Applied Artificial Intelligence, 4 March 2022) https://www.tandfonline.com/doi/full/10.1080/08839514.2022.2037254

354. Sun M and others, ‘MakeupAttack: Feature Space Black-Box Backdoor Attack on Face Recognition via Makeup Transfer’ (Cornell University, 22 August 2024) https://arxiv.org/abs/2408.12312

355. Quincozes SE and others, ‘A Survey on Intrusion Detection and Prevention Systems in Digital Substations’ (Computer Networks, 15 January 2021) https://www.sciencedirect.com/topics/computer-science/data-injection-attack

356. Anastasios Giannaros and others, ‘Autonomous Vehicles: Sophisticated Attacks, Safety Issues, Challenges, Open Topics, Blockchain, and Future Directions’ (Journal of Cybersecurity and Privacy, 5 August 2023) https://www.mdpi.com/2624-800X/3/3/25

357. Columbus L, ‘Adversarial Attacks on AI Models Are Rising: What Should You Do Now?’ (VentureBeat, 21 September 2024) https://venturebeat.com/security/adversarial-attacks-on-ai-models-are-rising-what-should-you-do-now/

358. Ismail Dergaa and others, ‘Using Artificial Intelligence for Exercise Prescription in Personalised Health Promotion: A Critical Evaluation of OpenAI’s GPT-4 Model’ (Biology of Sport, 13 December 2023) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10955739/

359. El Mestari SZ, Lenzini G and Demirci H, ‘Preserving Data Privacy in Machine Learning Systems’ (Computers & Security, February 2024) https://www.sciencedirect.com/science/article/pii/S0167404823005151

360. Raji MA and others, ‘E-Commerce and Consumer Behavior: A Review of AI-Powered Personalization and Market Trends’ (GSC Advanced Research and Reviews, 6 February 2024) https://gsconlinepress.com/journals/gscarr/sites/default/files/GSCARR-2024-0090.pdf

361. Huang H and others, ‘Data Poisoning Attacks to Deep Learning Based Recommender Systems’ (Cornell University, 8 January 2021) https://arxiv.org/abs/2101.02644

362. Goodfellow I, Shlens J and Szegedy C, ‘Explaining and Harnessing Adversarial examples’ (ICLR, 2015) https://arxiv.org/pdf/1412.6572

363. Kaur R, Dušan Gabrijelčič and Tomaž Klobučar, ‘Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions’ (Information Fusion, September 2023) https://www.sciencedirect.com/science/article/pii/S1566253523001136

364. Bowen E and others, ‘Cyber AI: Real Defense’ (Deloitte Insights, 6 December 2021) https://www2.deloitte.com/us/en/insights/focus/tech-trends/2022/future-of-cybersecurity-and-ai.html

365. Abdullahi M and others, ‘Detecting Cybersecurity Attacks in Internet of Things Using Artificial Intelligence Methods: A Systematic Literature Review’ MDPI, 10 J anuary 2022) https://www.mdpi.com/2079-9292/11/2/198

366. Aras MT and Guvensan MA, ‘A Multi-Modal Profiling Fraud-Detection System for Capturing Suspicious Airline Ticket Activities’ (Applied Sciences, 9 December 2023) https://www.mdpi.com/2076-3417/13/24/13121

367. Bill Scherlis, ‘Weaknesses and Vulnerabilities in Modern AI: AI Risk, Cyber Risk, and Planning for Test and Evaluation’ (SEI Blog, 12 August 2024) https://insights.sei.cmu.edu/blog/weaknesses-and-vulnerabilities-in-modern-ai-ai-risk-cyber-risk-and-planning-for-test-and-evaluation/

368. Longo L and others, ‘Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions’ (Information Fusion, June 2024) https://www.sciencedirect.com/science/article/pii/S1566253524000794

369. Berghoff C, Neu M and Arndt von Twickel, ‘Vulnerabilities of Connectionist AI Applications: Evaluation and Defense’ (Frontiers in Big Data, 22 July 2020) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7931957/

370. Wolter K and Reinecke P, ‘Performance and Security Tradeoff’ (Lecture Notes in Computer Science 2010) https://link.springer.com/chapter/10.1007/978-3-642-13678-8_4

371. Reddy S, ‘Generative AI in Healthcare: An Implementation Science Informed Translational Path on Application, Integration and Governance’ (Implementation Science, 2024) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10941464/

372. Singh C and Ankit Kumar Jain, ‘A Comprehensive Survey on DDoS Attacks Detection & Mitigation in SDN-IoT Network’ (e-Prime- Advances in Electrical Engineering Electronics and Energy, J une 2024) https://www.sciencedirect.com/science/article/pii/S2772671124001256

373. Habbal A, Mohamed Khalif Ali and Mustafa Ali Abuzaraida, ‘Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, Applications, Challenges and Future Research Directions’ (Expert Systems with Applications, 15 April 2024) https://www.sciencedirect.com/science/article/abs/pii/S0957417423029445

374. Walker DM and others, ‘Perspectives on Challenges and Opportunities for Interoperability: Findings from Key Informant Interviews with Stakeholders in Ohio’ (JMIR Medical Informatics, 2023) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10007006/

375. Jada I and Mayayise TO, ‘The Impact of Artificial Intelligence on Organisational Cyber Security: An Outcome of a Systematic Literature Review’ (Data and Information Management, June 2024) https://www.sciencedirect.com/science/article/pii/S2543925123000372

376. ‘Securing the Public Safety IoT Ecosystem with Blockchain- IEEE Public Safety Technology Initiative’ (Ieee.org) https://publicsafety.ieee.org/topics/securing-the-public-safety-iot-ecosystem-with-blockchain

377. Segil J, ‘Basics of SPM December 2024’ (Security Industry Association, 19 March 2024) https://www.securityindustry.org/2024/03/19/how-ai-can-transform-integrated-security/

378. Gerry Aue and others, ‘Scaling AI for Success: Four Technical Enablers for Sustained Impact’ (McKinsey & Company, 27 September 2023) https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/tech-forward/scaling-ai-for-success-four-technical-enablers-for-sustained-impact

379. Jada I and Mayayise TO, ‘The Impact of Artificial Intelligence on Organisational Cyber Security: An Outcome of a Systematic Literature Review’ (Data and Information Management, June 2024) https://www.sciencedirect.com/science/article/pii/S2543925123000372

380. Miles Brundage and others, ‘The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation’ (OpenAI, University of Oxford and more, 2018) https://arxiv.org/pdf/1802.07228

381. ‘The Three Biggest Security Challenges Facing AI and Data Initiatives’ (Databricks, 31 October 2018) https://www.databricks.com/blog/2018/10/31/the-three-biggest-security-challenges-facing-ai-and-data-initiatives.html

382. ‘Generative AI: The Data Protection Implications’ (CEDPO) https://cedpo.eu/wp-content/uploads/generative-ai-the-data-protection-implications-16-10-2023.pdf

383. Gadotti A and others, ‘Anonymization: The Imperfect Science of Using Data While Preserving Privacy’ (Science Advances, 2024) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC466941/

384. Dwivedi YK and others, 'Opinion Paper: "So What If ChatGPT Wrote It?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy' (International Journal of Information Management, August 2023) https://www.sciencedirect.com/science/article/pii/S0268401223000233

385. 'AI and Ethics: Balancing Progress and Protection' (Dataconomy, January 2023) https://dataconomy.com/2023/01/16/artificial-intelligence-security-issues/

386. Ul Rehman S and others, 'AI-Based Tool for Early Detection of Alzheimer's Disease' (Heliyon, 10 April 2024) https://www.sciencedirect.com/science/article/pii/S2405844024054069

387. Mirkin S and Albensi BC, 'Should Artificial Intelligence Be Used in Conjunction with Neuroimaging in the Diagnosis of Alzheimer's Disease?' (Frontiers in Aging Neuroscience, 18 April 2023) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10177660/

388. Levitt K, 'How Is AI Used in Fraud Detection?' (NVIDIA Blog, 13 December 2023) https://blogs.nvidia.com/blog/ai-fraud-detection-rapids-triton-tensorrt-nemo/

389. Bello O and others, 'AI-Driven Approaches for Real-Time Fraud Detection in US Financial Transactions: Challenges and Opportunities' (European Journal of Computer Science and Information Technology, 2023) https://tudr.org/id/eprint/3078/1/AI-Driven%20Approaches.pdf

390. Nigam A and others, 'A Systematic Review on AI-Based Proctoring Systems: Past, Present and Future' (Springer Nature, 2021) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8220875/

391. Mrutyunjay Padhiary and others, 'Enhancing Precision Agriculture: A Comprehensive Review of Machine Learning and AI Vision Applications in All-Terrain Vehicle for Farm Automation' (Smart Agricultural Technology, 2024) https://www.sciencedirect.com/science/article/pii/S2772375524000881

392. 'Ensuring the Security of an AI System's Development', (Cnil.fr) https://www.cnil.fr/en/ensuring-security-ai-systems-development

393. Aleksander Madry and others, 'Towards Deep Learning Models Resistant to Adversarial Attacks' (Cornell University, 2017) https://arxiv.org/abs/1706.06083

394. 'NB Defense- OECD.AI', Oecd.ai https://oecd.ai/en/catalogue/tools/nb-defense

395. 'Welcome to the Adversarial Robustness Toolbox — Adversarial Robustness Toolbox 1.17.0 Documentation' (Readthedocs.io, 2018) https://adversarial-robustness-toolbox.readthedocs.io/en/latest/

396. 'Welcome to Garak!', Garak.ai2 https://docs.garak.ai/garak

397. 'Google's Secure AI Framework- Google Safety Center', Safety.google) https://safety.google/cybersecurity-advancements/saif/

398. General Controls – AI Exchange', Owaspai.org) https://owaspai.org/docs/1_general_controls/#secprogram

399. 'Microsoft Security Development Lifecycle Practices', Microsoft.com) https://www.microsoft.com/en-us/securityengineering/sdl/practices

400. ‘AI Risk Management Framework’ (National Institute of Standards and Technology, January 2023) https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

401. Akash Takyar, ‘Data Security in AI Systems’ (LeewayHertz- AI Development Company) https://www.leewayhertz.com/data-security-in-ai-systems/

402. Charles Jackson, ‘Open source, open risks: The growing dangers of unregulated generative AI’ (IBM, 2024) https://securityintelligence.com/articles/unregulated-generative-ai-dangers-open-source/

NOTES

german cooperation
DEUTSCHE ZUSAMMENARBEIT

Implemented by
giz

FAIR FORWARD
Artificial intelligence for all.

IKIGAI LAW

DSCI
PROMOTING DATA PROTECTION

nasscom