


Downstreaming Responsible Artificial Intelligence in India: A Primer





The research was conducted by **Aapti Institute**, a public research institute that works on the intersection of technology and society. Aapti examines the ways in which people interact and negotiate with technology both offline and online.

Author: Vaishnavi Patil | **Project Guidance:** Kunal Raj Barua | **External Technical Editor:** Aindriya Barua

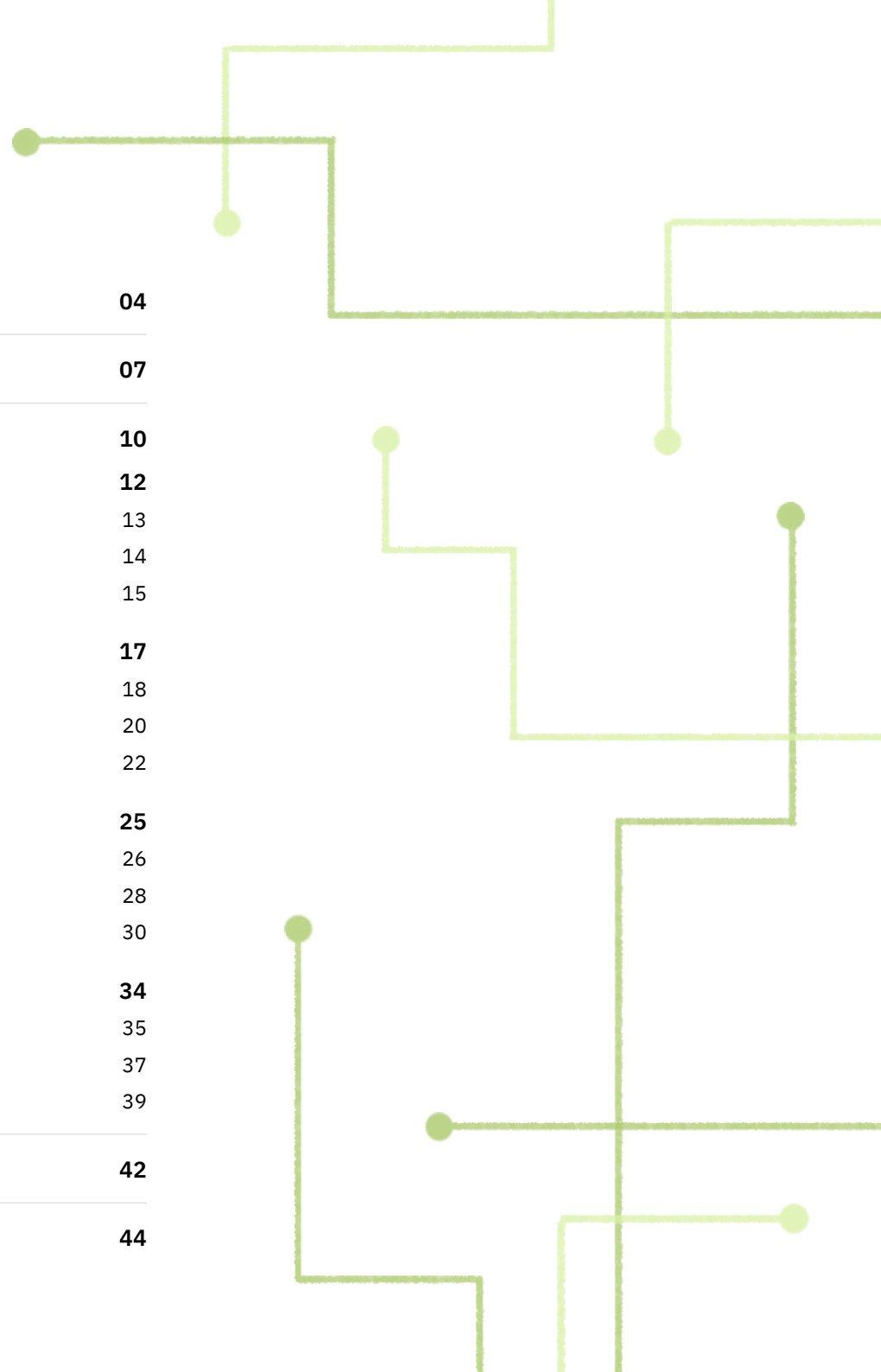
Report design: Meher Rajpal | **Cover illustration:** Monjira Sen

ACKNOWLEDGEMENTS

In addition to contributions from the wider Aapti team, this report also draws upon the expertise of numerous academic and industry experts, practitioners. We are grateful for their input during interviews and feedback sessions.

Table of Contents

Responsible AI	04
The Case for Responsible AI	07
Introduction to the Primers	10
Fairness	12
Preproduction – Fairness	13
Development – Fairness	14
Adaptation – Fairness	15
Privacy	17
Preproduction – Privacy	18
Development – Privacy	20
Adaptation – Privacy	22
Transparency & Accountability	25
Preproduction – Privacy	26
Development – Privacy	28
Adaptation – Privacy	30
Safety & Security	34
Preproduction – Safety & Security	35
Development – Safety & Security	37
Adaptation – Safety & Security	39
Macro Recommendations	42
A Short Note to Developers	44



SECTION 1

Responsible AI

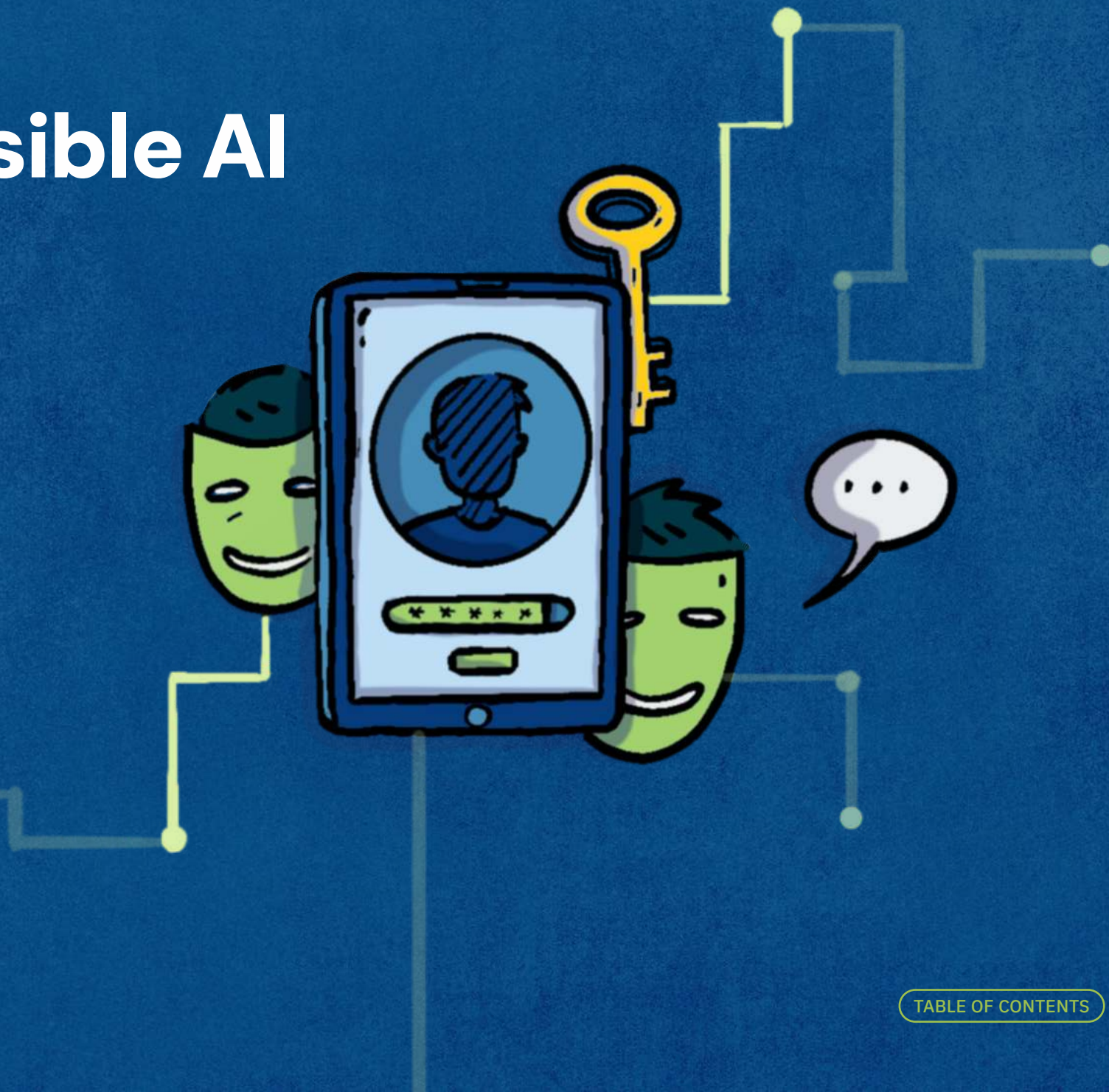


TABLE OF CONTENTS

Responsible AI (RAI)

Introduction

Artificial Intelligence (AI) has become an unprecedented node for growth, enhancing productivity, optimising decision-making, and providing solutions that deliver value. An entire ecosystem of developers sustain the creation and maintenance of such critical tools — making AI the cornerstone of modern innovation.

With AI becoming deeply embedded in high-stake domains like healthcare and finance, its processes and usage has invited global scrutiny. People have raised concerns regarding unintended biases, privacy violations, lack of accountability, and potential of severe harm if systems fail. Regulatory frameworks like the EU AI Act, the General Data Protection Regulation (GDPR), and the Artificial Intelligence and Data Act (AIDA) are some of the prominent responses to these

concerns. India, too, is adopting the Digital Personal Data Protection (DPDP) Act, 2023 to provide for the processing of the digital personal data.

As the stakes rise, so does the risk for reactive, “shut-down” regulations that could stifle innovation. As such, there is now a call for taking up a proactive approach — embedding ethical oversight, technical best practices, and accountability measures within the workflows of AI systems. The principles and practices of Responsible AI promise the actualisation of this approach, while invoking the importance of the developer who, being at the heart of the development of AI systems, become the centerpiece in this conversation around AI systems.

What is Responsible AI?

Industry leaders like Microsoft, Google, and IBM have been pioneering this field, setting standards, values, best practices in place. IBM describes Responsible AI as *“a set of principles that help guide the design, development and use of AI – building trust in AI solutions that have the potential to empower organizations and their stakeholders. Responsible AI involves the consideration of a broader societal impact of AI systems and the measures required to align these technologies with stakeholder values, legal standards and ethical principles.”*

Available Resources in the Ecosystem

These are some resources that meet industry standards, to foster the adoption of Responsible AI.



FRAMEWORKS

These frameworks are a set of guidelines to follow for the effective adoption of Responsible AI.

- [Responsible AI Standard v2, Microsoft](#)
- [Guidelines for Participatory and Inclusive AI, PAI](#)
- [Handbook on Data Protection and Privacy for Developers of Artificial Intelligence \(AI\) in India, GIZ](#)
- [Part 2 – Operationalising Principles for Responsible AI, NITI Aayog](#)
- [Guidelines for Secure AI System Development, NCSC](#)
- [Responsible Use of Machine Learning, AWS](#)
- [Responsible AI Architect's Guide, Nasscom](#)
- [Model AI Governance Framework for Gen-AI, AI Verify Foundation](#)
- [Responsible Use Guide, Meta Llama](#)
- [People+AI Guidebook, People+AI](#)
- [Guidance for Safe Foundation Model Deployment, PAI](#)
- [Responsible AI Guidelines for Generative AI, Nasscom](#)
- [Artificial Intelligence Risk Management Framework, NIST](#)
- [AI Risk-Management Standards Profile for General-Purpose AI Systems \(GPAIS\) and Foundation Models, UC Berkeley](#)
- [The Developer's Playbook for Responsible AI in India, Nasscom](#)



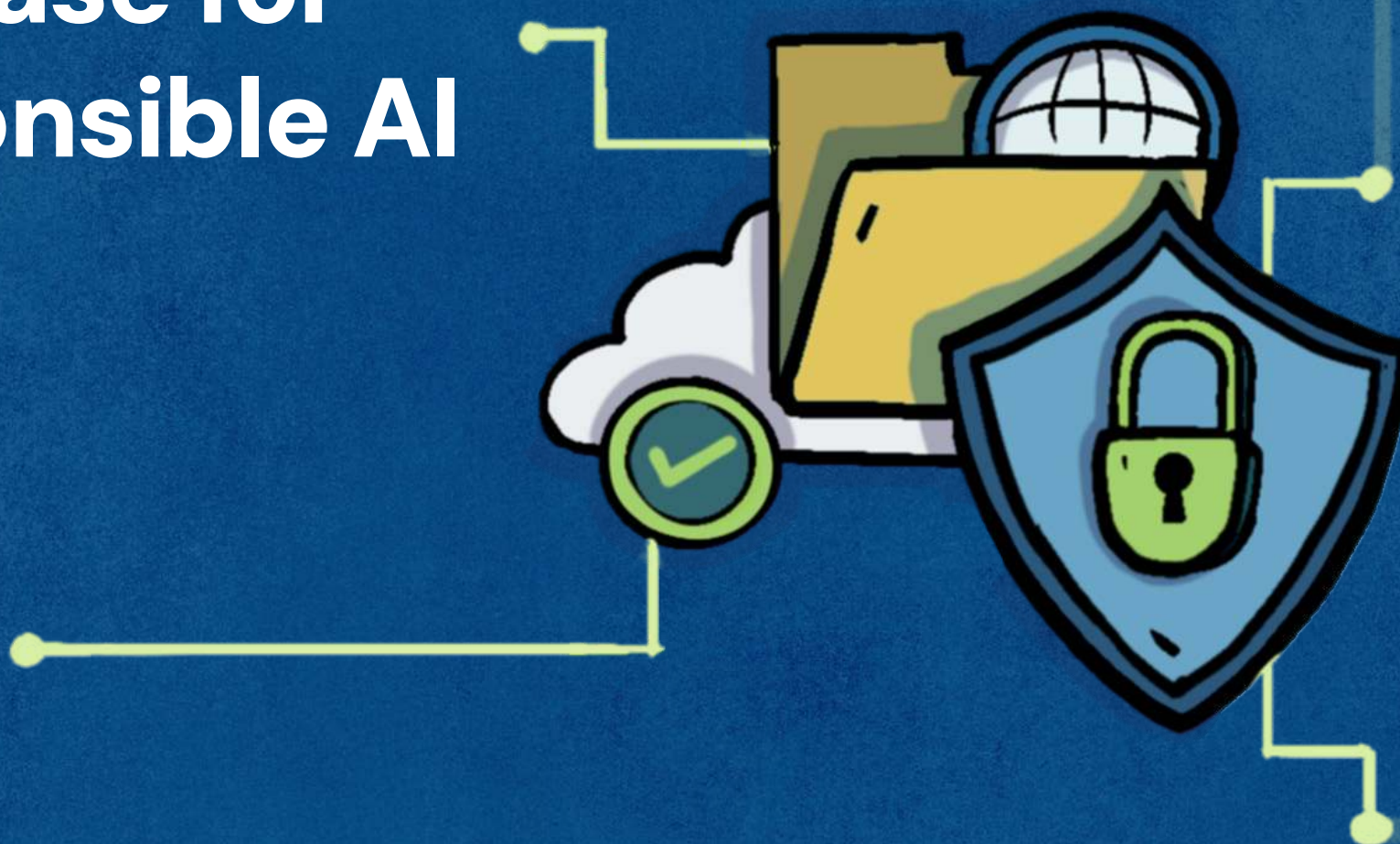
TOOLS

These are some important tools to help enable the implementation of Responsible AI.

- [IBM AI Fairness 360](#)
- [IBM AI Explainability 360](#)
- [Microsoft Fairlearn](#)
- [Google What-If Tool](#)
- [MITRE ATLAS](#)
- [Model Cards, Hugging Face](#)
- [Adversarial Robustness Toolbox \(ART\)](#)
- [AxCrypt](#)
- [watsonx.ai](#)
- [Ray RLlib](#)
- [OpenAI Spinning Up](#)
- [Azure AD](#)
- [Cobalt Strike](#)
- [Cuckoo Sandbox](#)
- [Kubernetes](#)
- [Grafana](#)
- [CISCO ASA](#)
- [Ethical Impact Assessment, UNESCO](#)
- [AI and Data Protection Risk Toolkit, ICO](#)

SECTION 2

The Case for Responsible AI







[TABLE OF CONTENTS](#)

The Case for Embedding RAI




FOR BUSINESS

	Competitive advantage	90% executive respondents in the EIU survey consider long-term benefits and cost savings outweigh initial investment in RAI, enhancing product quality and competitiveness.
	Talent acquisition & retention	Unethical practices discourage diversity of talent, compromising quality and competitive advantage. RAI mitigates staff attrition costs and boost productivity.
	Enhanced market reach	Geographies and demographics offer differing expectations, norms and regulations. RAI helps foresee such challenges and access these markets.
	Better risk management	Upfront investment in RAI reduces downstream risks and their costs like business performance, reputational damage, sunk costs, lost sales, cancelled contracts, non-compliance fees, etc.
	Broader revenue streams	Product differentiation due to RAI helps increase market share. 2019 Ethisphere analysis shows World's Most Ethical Companies outperform Large Cap Index companies over 5 years by 14.4% and over 3 years by 10.5%.
	Procurement advantages	In bidding processes, RAI gives procurement advantages (VCs, public grants, government projects, etc.). >90% of company respondents in the EIU survey include ethical consideration in these processes.
	Increased pricing power	RAI can increase pricing power through better branding and reputation – which leads to price premiums of 26% on average. 2015 Nielsen survey showed 66% consumers are willing to pay more for ethical goods.

FOR THE ECOSYSTEM

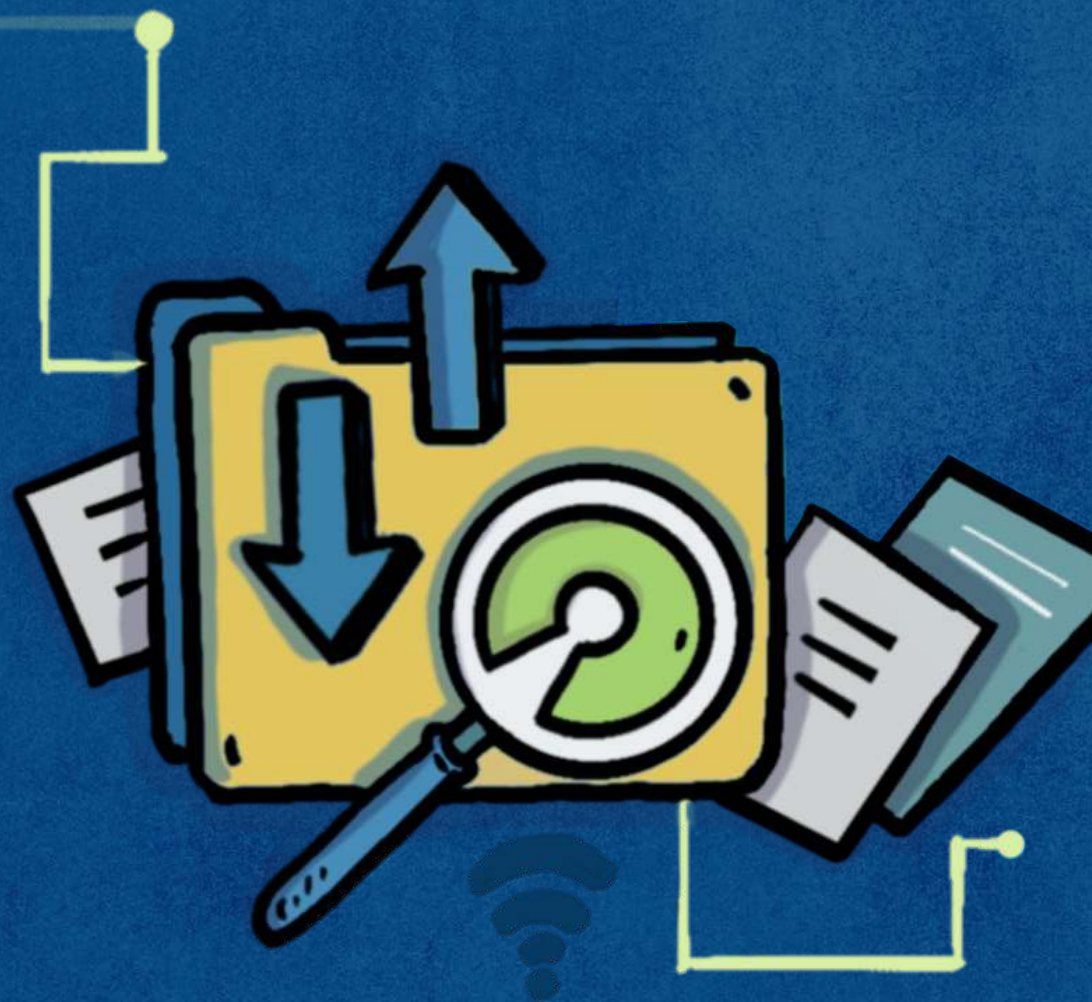
	Increased public trust	RAI <u>reduces</u> the risk of public scandal and distrust by prioritising users, leading to strong brand development, engagement and public relations.
	Increased collaboration	Multistakeholder collaborations that enhance RAI <u>allow</u> tech companies to become thought leaders.
	Sustainable investing	Conscious investors have already begun to <u>advocate</u> for ethical practices before deciding to invest.
	Preparedness for compliance	When the GDPR became law, only <u>31% businesses</u> believed they were ready. RAI helps prepare for such upcoming regulations.

FOR THE USERS

	Increased AI adoption	Security and privacy are the <u>biggest obstacles</u> to AI adoption, especially in heavily regulated sectors. RAI addresses them.
	Increased inclusivity	RAI compels technology companies to <u>perform well</u> across broad user profiles, thereby boosting product value.
	User engagement	<u>Capgemini</u> analysis suggests 70% of consumers expect ethics in AI services, while <u>Salesforce</u> suggests 95% of customers are more loyal to industries they trust.

SECTION 3

Introduction to the Primer



[TABLE OF CONTENTS](#)

Introduction to the Primer

The primer aims to be a collection of myriad practices that one could employ in thinking about and operationalising Responsible AI. The term ‘developer’ used in the primer is an umbrella term meant to indicate a person engaged in any stage of building digital architecture, and includes various AI practitioners like model builders, AI deployers, AI designers, etc. Depending on the context, a developer can pick the most relevant recommendations from this collection to implement.

The reading of this primer can also be accompanied with a deeper exploration of the resources and hyperlinks included.

How to read the primer

There is no one dominant framework when it comes to Responsible AI. However, common themes can be derived from existing literature.

These principles form the basis of the primer:



FAIRNESS



PRIVACY



TRANSPARENCY & ACCOUNTABILITY

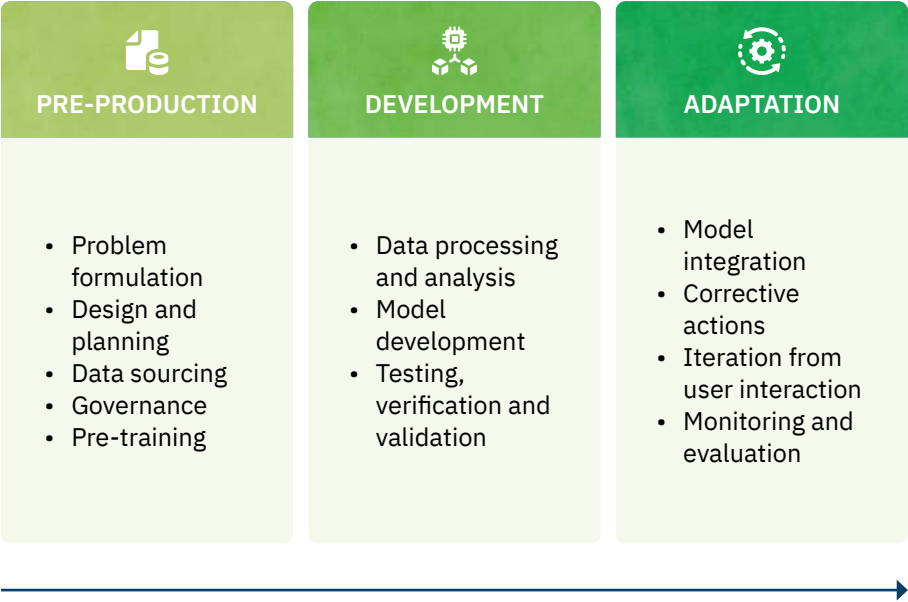


SAFETY & SECURITY

Furthermore, each principle is organised using the value chain approach (right). After discussing the practices, larger recommendations to enable ecosystem efforts for Responsible AI are put forth, including a short note and call-to-action for developers.

The Value Chain Approach

The primer is structured along the lines of the value chain: three stages that determine the AI system lifecycle. The methods of RAI intervention are nestled within these three stages.



Fairness

What is it?

Fairness refers to the principle that allows an AI system to operate equitably — preventing bias, discrimination, inequality, and inequity. Every AI system has its own lens with which it interacts with the world. This lens will always operate with a set of distortions. *Embedded societal assumptions of product developers, called cognitive biases, are one such set of distortions. Errors in analysis or data collection, called algorithmic biases, are another.* Such biases could lead to an AI system causing discriminatory or harmful outputs that are unfair to its stakeholders. It becomes important to intervene actively to mitigate such harms.

Why should you care?

Fairness directly impacts an AI system's effectiveness and social acceptability. A fair AI system attracts a larger, more diverse range of users — making it possible to scale across demographics and geographies. It also improves performance while demonstrating responsibility.

Key Concepts

Bias mitigation

Biases in training data, algorithms and decision-making processes need to be addressed to create a useful and fair AI system.

Equitability

AI systems should ensure that decisions, recommendations, or outputs do not unfairly favour nor disadvantage individuals or groups based on attributes such as race, gender, ethnicity, age, disability, or socio-economic status.

Inclusivity

AI systems should consider the needs and contexts of different populations to ensure their design and deployment are inclusive, driven by open innovation that values diverse perspectives.

Contextual Sensitivity

Specific domains may have varying requirements for fairness. AI systems need to understand the context under which they operate, making sure no group is disproportionately harmed.

Risks it addresses

Discrimination

E.g. A resume screening AI tool trained on urban candidates downgrades graduates from tier-2 and tier-3 city colleges

Disproportional outcomes

E.g. A hiring AI system recommends lower-paying roles to candidates with Muslim or Dalit surnames based on historical hiring patterns

Disproportionate harms

E.g. A loan approval system rejects small business loans from women entrepreneurs in rural areas

Insensitive interactions

E.g. A job assessment AI system penalises mothers for taking time off for caregiving

Lack of consent

E.g. An AI system determines credit worthiness of an individual from their social media activity without disclosure

Unintended consequences

E.g. An AI tool that distributes financial aid prioritises urban applicants as they are overrepresented in training data



FAIRNESS Preproduction

It is important to acknowledge and anticipate potential biases from an early stage of development to ensure accessibility and inclusivity.

Understand existing context

During the stage of problem formulation, spend time understanding the problem ecosystem and how it interacts with your users. Conduct background research and consult domain experts to study the same.

- **Identify the problem ecosystem:**
What is the social, economic, and regulatory landscape?
- **Identify the stakeholders:**
Who are the potential users and community at large being affected?
- **Study biases:**
How do biases impact such stakeholder groups differently in your sector?
- **Find opportunities:**
What are the existing ways of overcoming discriminatory barriers?

Assess datasets and data practices

Investing early in clean and fair datasets and data practices goes a long way in ensuring fairer outcomes.

*Find gaps
in your
dataset(s)*

*Strategies
for mitigating
gaps*

*Tools for
assessing biases
in datasets*

*Ensure
compliance from
the beginning*

*Use fair
collection
and labelling
practices*

Domain experts and community members can help ensure that dataset features and labels are appropriate.

Discern potential biases in your system

- **Evaluate your AI system's value add:** After understanding existing biases, the barriers they produce and existing opportunities to address them, evaluate how your AI system will interact with this ecosystem.
- **Consider broad and long-term impact:** Map out your ideal outcome and functionality in terms of short-term and long-term goals and who these goals impact.
- **Map ways in which biased outcomes may arise:** What are the gaps in your workflow that could lead to biased outcomes, hindering progress to ideal outcome and functionality?

Co-create with potential users

Conducting user research will ensure a human-centric approach that allows for an AI system to be inclusive and user-friendly, while defining barriers and determining different use-cases. Use following design research tools to make sure that user needs, priorities, vulnerabilities and opportunities are mapped.



Affinity
Diagrams



Empathy
Maps



Personas



Stakeholder
Maps



FAIRNESS Development

AI systems need to be evaluated continuously and iteratively to ensure that the outcomes are non-discriminatory.

Train model to understand many variations in input

- Ensure data is diverse and representative. Understand how sensitive attributes (like caste, address, gender) are influencing predictions
- Introduce fairness metrics in the data labelling process or employ double-blind or multi-blind labelling
- Use synthetic data generation techniques like Synthetic Minority Oversampling Technique (SMOTE) to ensure model does not overfit to one group
- Use bias detection tools like IBM AI Fairness 360 or Microsoft FairLearn to automate bias detection
- Incorporate human-in-the-loop (HITL) review, especially in high-stakes decisions

Use red-teaming for stress-testing the model for biases

- Identify and map biases to investigate
- Bring in experts and diverse communities to spot development stage biases
- Use adversarial testing to simulate input scenarios, unintended consequences or harmful outcomes that highlight biases
- Apply bias mitigation strategies like rebalancing training data or using adversarial debiasing techniques
- Continue red-teaming iteratively to catch biases that might emerge with new user behaviours
- Document all findings and communicate risks to stakeholders for informed decision making

Evaluate performance across demographics

Organise simulations of diverse testers to understand performance across demographics using **Disparity Testing**. Look for false positives and false negatives to apprehend disproportionate impact. While evaluating metrics like accuracy, precision and recall, use the following fairness measures:

Demographic Parity

Ensures that the AI model's positive prediction rate is the same across different demographic groups, regardless of their actual outcomes.

Equality of Opportunity

Guarantees that individuals who qualify for a positive outcome (e.g., truly qualified candidates) are equally likely to receive it across different groups.

Counterfactual Fairness

Ensures that an individual's prediction remain the same if their sensitive attributes were changed while keeping all other variables constant.



Often, biases are discovered only after the rollout. Mitigating and remedying such biases are crucial for harm reduction.

Augment Training Data

If you uncover issues during an audit of training data, for instance, missing or skewed data, it becomes important to augment the training data.

- **Input additional data:** Collect additional data that is more diverse and representative
- **Balance the data:** If certain groups are over- or underrepresented, apply data balancing techniques like oversampling the minority class, undersampling the majority class, and using synthetic data generation techniques like SMOTE
- **Normalise sensitive features:** After identifying the correlations between target labels and sensitive attributes like gender and income, reapply preprocessing techniques like normalising sensitive features or debiasing word embeddings for text-based models

Monitor iteratively to detect potential biases

- Continue red teaming iteratively to catch biases that might emerge with new user behaviours
- Regularly update model with fresh and diverse data
- Incorporate ethical reviews and audits to assess the model's alignment with fairness metrics
- Engage with ethicists, subject experts and community members continually to identify and address biases
- Create clear escalation paths when bias or harm is detected
- Plan ways to sunset or deprecate models when they cannot meet fairness standards or are beyond repair

Adjust model's optimisation function

In cases where collecting additional training data is not viable, another approach for mitigating bias is to adjust how loss is calculated during model training.

MINDIFF

MinDiff aims to balance the errors for two different slices of data (male/female students versus nonbinary students) by adding a penalty for differences in the prediction distributions for the two groups. It is a regularisation technique that reduces differences in model performance or predictions between predefined groups by introducing an additional term to the model's loss function.

COUNTERFACTUAL LOGIT PAIRING

CLP aims to ensure that changing a sensitive attribute of a given example doesn't alter the model's prediction for that example. It ensures predictions remain consistent for an individual, helping the model focus on relevant features rather than biases linked to sensitive attributes. E.g. if two examples whose feature values are same except for gender, CLP will add a penalty if the predictions for these two are different.

Fairness for India

On embracing India's diversity

India is home to many diverse communities. Therefore, there is a deeper need to create a fairer model that understands diverse inputs.

- Having a diverse and inclusive research team is a good way to make sure your initial forays into developing AI systems include multiple perspectives that can help spot gaps early.
- You can begin with basic developing and testing the model across 2-3 major regional languages.
- AI4Bharat's [IndicNLP](#) library can be useful for data augmentation.
- In case of low-resource languages, Bhashanet.in's [transliteration tools](#) can be used.
- When working with code-mixed data, specialized tools from IIIT-H's code-mixing toolkit are useful.
- AI4Bharat's IndicBERT can be helpful in checking for biases in Indian language models.

Fairness evaluation is systemic

When evaluating for fairness in India, systemic issues need to be kept in mind.

- Test across different Indian language scripts and regional dialects.
- Check for biases in code-mixed language.
- Test with different literacy levels.
- Assess performance with varying access to internet connectivity.

When red-teaming is inaccessible

Red-teaming can be resource-intensive for early startups. Some low-cost high-impact strategies for stress-testing the model:

- Use simple statistical testing using Python libraries.
- Partner with open-source groups, online penetration-testing communities, and local tech communities like CITRIS AI for peer reviews and feedback.
- Engage with CSOs and NGOs for domain expertise and reaching out to marginalized communities.
- Connecting with AI ethics groups and universities that can help with cost-effective student-led testing.

Bring community to the forefront

Community engagement needs to be sensitive and responsible, especially if the marginalized communities need trauma-informed consideration:

- Run focus group discussions with communities that would engage with and/or could be affected.
- Provide fair compensation to community participants.
- Offer support when discussions might involve sensitive experiences.
- Plan for building respectful long-term relationships rather than extractive interactions.
- If possible, ensure transparent reporting of fairness metrics back to the community.
- Post deployment, have online/ offline engagement strategies for community to report fairness issues.

Privacy

What is it?

Privacy refers to the safeguarding of personal data and respecting an individual's right to control their personal information when AI systems are designed, developed and deployed. AI systems are built on records of human details and behaviour. It is not impossible to infer preferences and sensitive information and to link it to individuals in a dataset. Thus, it is imperative that systems guarantee that such data is protected when it is collected, processed, stored, transmitted, analysed, and retrieved. Privacy involves processes to handle sensitive data responsibly, comply with regulations, and maintain user trust.

Why should you care?

Privacy is the most stringent aspect of AI regulation; thus, meeting privacy requirements is non-negotiable for an AI system. More than just fulfilling legal or compliance obligations, addressing privacy is about generating user trust, fostering societal values, ensuring sustainable and safe AI.

Key Concepts

Data minimisation

To avoid misuse of personal data, it is necessary to limit the collection of data to only relevant data that is specific to the purpose disclosed to the users.

Informed consent

Individuals must be fully informed about how their data will be collected, used, stored, and shared and must provide explicit permission for its use.

Data protection

Data should be protected in whichever way possible. Sensitive data, in particular, should be encrypted or anonymised to protect against unauthorised access.

Right to Privacy

Privacy safeguards autonomy by ensuring individuals control how their personal information is used. AI systems must adhere to privacy laws to protect the individual right to privacy.

Risks it addresses

Data breaches

E.g. A cloud storage service used by an AI system is breached, leaking private medical records

Data Misuse

E.g. A UPI-linked fintech app assesses user health using payment history without consent, affecting their insurance eligibility

Re-identification

E.g. A study identifies people from insufficiently anonymised health data by linking it to public voter records

Stigma and harm

E.g. An AI system analysing social media activity inadvertently exposes users' mental health conditions

Autonomy loss

E.g. Employees using an AI productivity tool feel constantly surveilled, leading to decreased morale

Non-compliance

E.g. An AI marketing system fails to provide users with a way to opt out of data collection, leading to fines



Privacy should be the default in design of the AI system, especially in the beginning - balancing functionality, user centricity, and security



Understand compliance requirements

If your geography is India, then the Digital Personal Data Protection Act applies. Find comparative laws if you seek international markets (e.g. EU AI Act for Europe). The [DPDP Act](#) emphasises the following AI entities:

Data Principals

Person (e.g. user) to whom personal data relates. Personal data is information that can directly or indirectly identify a data principal.

Data Processor

Entity that processes data on behalf of the data fiduciary
Liabilities: Security Safeguards

Data Fiduciary

Entity that determines the purpose and/or means of processing personal data
Liabilities: Notice and Consent, Security Safeguards, Data Quality, Reporting Data Breaches, Maintenance of Records, Audit, Impact Assessment

Limit dataset to purpose

The DPDP Act mandates data minimisation and purpose limitation. Ensure the following principles during the processing of data collection

Data should be adequate, relevant, and limited to what is necessary

Personal data collected for a purpose should not be used for new, incompatible purpose

Don't retain data longer than necessary

Be accountable for purpose of personal data, ensuring compliance

Methods to reduce data: [Generative adversarial networks \(GANs\)](#), [synthetic data](#), [federated learning](#), [matrix capsules](#), [transfer learning](#)



Empower the data principal

The DPDP Act describes the rights of the data principals that developers need to be aware of.

Rights

Data principals have the right to:

Confirmation and access

Data portability

Correction and erasure

Be forgotten

Notice and Consent

It is **mandatory** to notify principals of the following before collecting their personal data:

- Purpose of data use
- Nature of personal data
- Contact of fiduciary
- Process of (and right to) withdrawing consent
- Source, if it is not principal
- Entities with whom data is shared
- If applicable, details of cross-border transfer
- Period of data retention

Define protocols for data security

Set mechanisms for ensuring “**Privacy By Design**” to build preventative measures over prescriptive

Role-based training

Train employees on data privacy, security protocols, and breach response

Access control

Restrict access to authorised personnel throughout the lifecycle

Vendor and third-party management

Ensure that third-party vendors or data processors follow data privacy protocols



This involves building capacity for monitoring and protecting rights of the principals while instilling techniques to make datasets secure



De-identify (pseudonymise/anonymise) personal data

Ensure that no datapoint can be traced back to a person even when the privacy of a dataset could be breached

TECHNIQUES

Suppression

Remove variables or values you don't need

Permutation

Swap values between data subjects

Replacement

Replace sensitive data with non-sensitive data (pseudonyms)

STATISTICAL APPROACH

K-anonymity

Each record is indistinguishable from at least $k - 1$ others, minimising re-identification risks

T-closeness

Distribution of sensitive attributes in each anonymised group to match the dataset's distribution

Generalisation

Reduce granularity of data

Perturbation

Add noise to obfuscate data

Top & Bottom Coding

Recode unique extreme values to set min/max values

L-diversity

Each group of indistinguishable records contains at least l diverse and distinct sensitive values

Differential Privacy

Calibrate noise to data queries, making reidentification mathematically impossible

Embed security in computational processes

If dataset includes personal data, consider computational needs

Computing Facility:

- Eg, Local (laptop) or external (cloud)?
- Which country hosts the data?

Software:

- Is root user access needed?
- Paid licenses needed?

Who else has access to the data?

CRYPTOGRAPHIC TECHNIQUES:

Encryption

Instill code/cypher that needs a key to decipher, e.g. symmetric, asymmetric, hybrid cyphers

Homomorphic Encryption

Encrypts not just data but also computation, either partially or fully

Confidential Computing

Use Trusted Execution Environment (TEE) to maintain data confidentiality

Secure Multiparty Computation

Techniques that let parties jointly analyse without sharing data, like secure set-intersection

Symmetric encryption algorithms: DES, AES, 3DES, Twofish, Blowfish, IDEA

Asymmetric algorithms: RSA, elliptic curve cryptography



Develop capacity for continuous validation

- Periodic Assessment to delete irrelevant data
- Annual Data Protection Impact Assessments (DPIA)
- Annual audits of data policies by independent auditors
- Consistent records of privacy-related documents in all components
- Mechanisms through which written records from principals can be processed (e.g. consent manager)
- Data Protection Officer to deal with all liability concerns related to privacy



Addressing a privacy breach in India involves prompt breach containment, regulatory notifications, and comprehensive remediation efforts



Detect and contain breaches using failsafe protocols

Rapidly identify and contain the breach by:

1. Isolating the affected systems
2. Halting suspicious activities
3. Shutting down compromised endpoints
4. If feasible, deleting compromised data from external sources

Activate fail-safe protocols to stop any ongoing data access or processing that may be affected by the breach by:

1. Revoking third-party API access
2. Restricting user accounts
3. Disabling specific features

Conduct thorough analysis of the root cause

Document the root cause analysis, as it will be critical in reporting to the [Data Protection Board \(DPB\)](#) and may help demonstrate compliance and mitigation efforts if regulatory inquiries or penalties arise.



Examine Logs



Review updates



Assess access



Notify the relevant stakeholders

Data Protection Board



If the breach involves personal data, the DPDP Act mandates notifying the DPB. A well-prepared notification can minimise penalties:

- Nature of data breach
- Cause of the breach
- Types of data involved
- Number of data principals affected
- Possible consequences
- Action taken to mitigate breach

Data Principals



If it is impacting principals, the DPDP Act emphasises notifying affected data principals without undue delay. Notify them through messages, emails, etc.

- What data was potentially compromised
- Any steps they can take, e.g. changing passwords
- Whom to contact for assistance

Implement corrective actions

Based on the root cause, remedy the concern by remediation (direct corrective plans to fix the issue), strengthen security measures, and make relevant changes to data protection and breach response policies to improve resilience

Patching

Protocol updates

Additional employee training

Additional authentication

Additional encryption

Risk assessment

Post incident monitoring

Privacy for India

Notify but sensitively

When it comes to notifying data principals, sometimes concerns can arise. Breaches can affect community dynamics and immediate notifications could put vulnerable people at risk. For example, in domestic violence cases, if abuser is notified about the breach before the survivor, it could put the survivor at risk. Thus, in sensitive cases, consider:

- Tiered notification systems starting with most vulnerable users first.
- To establish preliminary safety measures (like storing sensitive data separately) in place before broader notifications go out.
- In addition to safety measures, to also create alternative support systems ready before the notifications go out.
- Create careful guidelines that consider who gets notified and how.

Data is messier than you think

Current privacy techniques assume clean, structured data. However, in real life scenarios, collected data can appear in different forms, such as screenshots, voice notes, informal complaints, text messages, etc. In this case, a tailored approach for specific challenges of each data type needs to be accounted for:

- Such data needs rigorous classification. Identifiable elements in its content (e.g. faces, voices, etc.) as well as its metadata (E.g. timestamps, geolocation) need to be analysed.
- Different techniques to be used for different types of data. For example, using Natural Language Processing to redact sensitive terms in text data, blurring faces and other identifiable elements in video data, speech-to-text transcription in audio data, etc
- Techniques for mixed data formats: tokenisation or pseudonymisation of identifiers in all data, separately storing personally identifiable information from the core dataset, adding noise to prevent re-identification.

Informed Consent works differently in the Indian context

The global standard for gaining informed consent does not always translate in the Indian social structures:

- Often, in India, consent cannot be taken on an individual level. Sometimes, it needs to be community-level as they are interacting with the AI system on shared devices. This will include discussions around group privacy rights, especially for vulnerable communities.
- Power hierarchies between an organisation and the last mile should always be considered when it comes to informed consent. Sometimes, vulnerable people might consent only due to the power dynamic, or if they are unaware about the full risks. Additional processes to balance this dynamic could help:
 - a. Going back to the community to show how their data was used.
 - b. Getting their approval after processing as well, not just before.
 - c. Finding ways of reducing the power gap, like involving community members as co-researchers or having community representatives who understand both tech and the local context.

Transparency & Accountability

What is it?

While transparency ensures stakeholders have the ability to interpret AI systems, accountability allows stakeholders to intervene in its design and correct automated decisions. AI systems, with their usual 'black box' nature, are often at the risk of blurring the decision-making process. This could lead to unreasonable outcomes that could victimise its stakeholders. Creating frameworks and structures that allow a peek into the black box is the crucial first step towards building long-term user trust.

Why should you care?

As the key element to building trustworthy AI, transparency and accountability makes an AI system easy to understand, navigate and interact with long-term – allowing diverse participants to iterate, helping engineers identify and solve safety concerns, and highlighting privacy risks.

Key Concepts

Explainability By making AI systems interpretable, both users and developers are empowered to co-create iteratively — making the AI system work as intended and celebrating trustworthiness.

Human-centricity From ethical sourcing of data to designing a user-friendly interface, positioning human-centricity at the core of an AI system ensures an ethical lifecycle. It also allows an AI system to be proactive in terms of accountability and compliance.

Openness Revealing the nature of the 'ingredients' (data, algorithms, models, design, etc.) can help provide visibility into decision-making, including biases, limitations, and trade-offs — thereby setting realistic expectations across the AI ecosystem. Disclosures should be role-based.

Feedback Feedback and grievance redressal mechanisms should not just be relegated to high-stakes use-cases. Keeping communication open, clear, and simple allows for better user experience.

Risks it addresses

Black box decisions

E.g. An AI program in the justice system recommends detention for an individual but judge cannot explain why

Hidden biases

E.g. An AI system denies loans to minorities due to historical redlining data, offering no paths to contestation

Data misuse

E.g. A voice recording AI uses user conversations without disclosure for targeted ads but users don't know how to challenge

Erosion of public trust

E.g. Scandals tarnish reputation of tech companies, slowing adoption of beneficial AI tech

Lack of liabilities

E.g. A navigation app optimised for cars leads cyclists into dangerous routes, but there is no one held accountable

Non-compliance

E.g. An AI tool promises 100% fraud detection but misses subtle frauds, granting users false sense of security











Ensure ethical and open data collection and documentation

- **Source your data ethically**
If using an existing dataset, ensure,
 - a. it is collected responsibly and with consent, and
 - b. it is representative of user context, demographics and needs.If you are building your own, consider fairness and privacy risks
- **Aim for high-quality data**
Audit your dataset for gaps and imbalances

Clean your data, e.g. transform missing values, handle incorrect data, remove duplicates

Try to match input data to real-world data
- **Establish proper documentation**
Document using tools like [datasheets for datasets](#):
 - a. What does it represent?
 - b. What is the source?
 - c. What preprocessing was done?
 - d. What use cases can it be used for, with responsibility?
 - e. What are its limitations/ biases?

Encourage practices for open communication

-  **[Design for labelers](#)**
-  **[Map and share workflows](#)**
-  **[Iterative feedback loops](#)**
-  **[Get users to participate in usability testing](#)**
-  **[Facilitate cross-disciplinary collaborations](#)**
-  **[Early engagement with HR and legal experts](#)**
-  **[Share knowledge across teams](#)**
-  **[Value inclusivity](#)**
-  **[Establish regular communication channels](#)**
-  **[Maintain regular progress reports](#)**

Explain decision-making and its justifiability

STAKEHOLDER	WHAT THEY NEED TO KNOW	PURPOSE
End-users	How decisions affect them and how to appeal	Build trust, ensure fairness, and empower users
Engineers	Algorithms, parameters, and processes	Debugging, improving performance and robustness
Regulators and Policymakers	Ethical considerations, compliance, and audit trails	Assess legal compliance, mitigate risks, and ensure adherence to ethical standards
Domain Experts	Decision rationale	Appropriateness and improve alignment with real world
Ethics Committees	Bias checks, fairness metrics, and decision impacts	Ensure the AI aligns with organisational and societal ethical standards
Business Leaders	High-level summaries of decision logic and trade-offs	Align AI decisions with business goals, manage risks, and make strategic decisions
Communities	Potential impacts, biases, and recourses	Ensure social responsibility and mitigate harm

Define accountability and liability structures





Being transparent and accountable to users should be at the forefront of the modelling stage to ensure the outcomes match the user expectations.

Ensure data transformations are interpretable

Each dataset should have documentation for the baseline quality of ingested data and a record of all operations and transformed conducted on them.

- Use Explainable AI solutions like Local Interpretable Model Agnostic Explanations (LIME), decision trees, and Shapley additive explanations (SHAP)
- Transparency techniques like K-Nearest Neighbours, Generalised Additive Models, Bayesian Models, and Rule Based Learning and post-hoc explainability models supplement such solutions
- Maintain simplicity, consistency, and standardisation in transformations as much as possible
- Provide interpretability through data visualisations

Understand users' mental model

To build a usable AI system, it is necessary to understand the context of how users might interact with your AI system. After studying such processes or mental models users employ to accomplish tasks, it becomes important to map multiple flexible paths to reach outcomes.

CONSIDER:

What are the existing mental models in your ecosystem?

What are the gaps and how can you design better UX?

What aspects need to be explained to users?

How to integrate communication like marketing and feedback?

Intuitive

Existing mental models & cross disciplinary collaboration can help create appropriate interactions that feel intuitive

Flexible

Design so that users can experiment. Keep room for new features and new forms of interactions

Intentional

Ensure handholding during onboarding or when new features are added but refrain from overexplaining

Trustworthy

Set the right expectations with each feature, and communicate how the user can trust its service

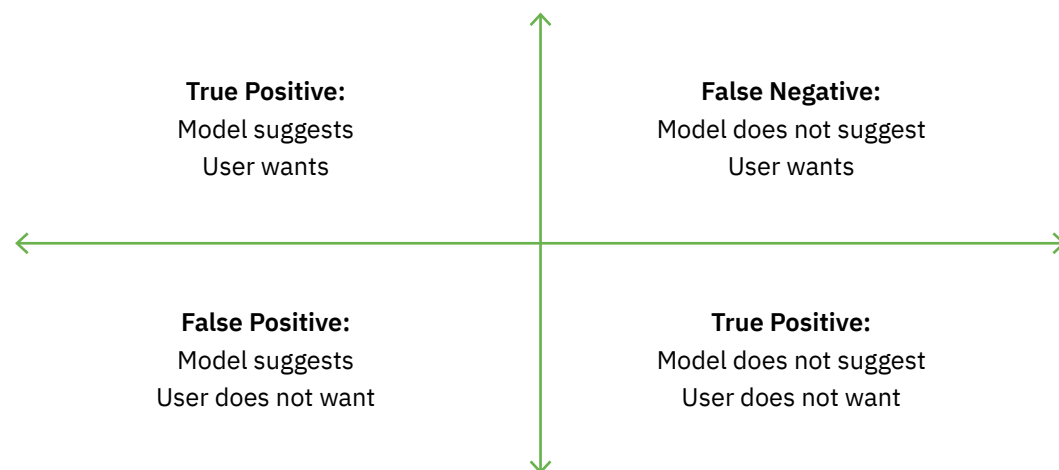
Human-like

Not all interactions need to sound human, but some features might benefit from doing so



Tune the model's reward function appropriately

Depending on the use case, the reward or loss function of the model needs to be weighed appropriately. Ensure cross-disciplinary collaborations, especially with domain experts, to achieve the correct outcomes



This suggests two types of errors:

1. If the model is dedicated to **precision**, it will predict accurately but some outcomes that the user might also want are filtered out (false negatives/ missed detections)
2. If the model is dedicated to **recall**, it will produce broader outcomes, but some could be irrelevant results (false positives/ false alarms)

Precision and Recall trade-offs need to be designed on a use-case basis.

Such trade-offs and their confidence should be communicated if it improves decision-making

Visualisations to show model confidence:



Categorical



N-best alternatives



Numeric



Graphic



It is necessary to install gracious iterative failure mechanisms to ensure that the user trust is not broken

Test and iterate for optimising user trust

Users take time to build trust. Transparent communication with defined accountability structures are the first step in this process. The following should be tested iteratively to improve user experience of your AI system.

Competence

If the system meets the set expectations communicated to the users

Consistency

If the system delivers reliably across features within their set error margins

Control

If the system allows the user to feel agential and in control of their decisions

User Centricity

If the system's interface is accessible, inclusive, and intuitive

Maintain robust accountability systems

IDENTIFY ERRORS

System Limitations

Inherent limitations prevent correct or any outcomes to be generated. Output errors like having low confidence or generating irrelevant outputs can occur.

User Errors

Users can give incomplete or incorrect input, while anticipating the system to understand regardless. The user could also be a novice or require assistance.

User-perceived Errors

Users can feel the output to be incorrect due to poor assumptions of user needs. E.g. Context errors like receiving unrelated updates, or if the model cannot provide an output for the given input

INVESTIGATE

Find the source of the error, e.g.,

- Reward function might need further optimisation
- Errors in training data like missing or mislabelled data
- Deficiencies in data modelling

Categorise and prioritise errors based on frequency, severity, and whether it is a high-stakes failure.

PATHS FORWARD

Opportunities for feedback and redressal, e.g.,

- For user-perceived errors, ask user what output they expected instead
- For output error, apologise and explain why a result was not produced and suggest alternatives

Allow users more control to define outcomes, especially in high-stakes failures.



Embed mechanisms for feedback and grievance redressal

Feedback



- If the feedback is **implicit**, i.e. usage data being collected over time, then make sure it is mentioned clearly in the terms of use (what data is being collected and how it is used)
- If the feedback is **explicit**, i.e. prompted commentary on experience and outputs, then they should be simple, structured, and unintrusive. Try to communicate how the feedback will help the user as it might encourage them to do so. Thank the user after receiving feedback as it builds trust

Grievance Redressal



- Maintain **clear reporting channels**:
 - **Help centers**: with FAQs, tutorials, and guides for common concerns
 - **Support tickets**: to report unresolved issues
 - **Contact options**: multiple ways to escalate concerns, including chatbots, email support, or human agents
- Incorporate **appeal processes** to challenge high stakes decisions (e.g. loan denials, bans). Trigger human-in-the-loop systems for them
- Acknowledge **receipts** of complaints with **resolution timelines** or next steps



Transparency & Accountability for India



Being transparent in the Indian context

When being transparent with Indian audiences it is necessary to understand which issues they might face in accessing this information.

- If possible, try to maintain documentation in the language potential users find accessible.
- It is important to understand how to generate tech explanations for low-literacy users.
- In general, it is good practice to invest in initiatives that attempt at increasing digital literacy.
- Create accessible offline feedback channels for community members.
- Maintain methods of building trust with people who have had negative experiences with technology.
- Data labelling work matters. It is important to communicate clearly the purpose and expectations of the AI system to data labellers. Create guidelines for labelling with inputs from domain experts and community.

Transparency needs to be balanced with safety and privacy

Openness in data sharing can be harmful, particularly if the data concerns vulnerable communities.

- For instance, when working with sensitive caste/ gender discrimination data, being completely open might put communities at risk. Thus, there should be clearer guidelines on balancing openness with protection of vulnerable communities.
- Perfect security can also be a hindrance to openness and accessibility. For example, Discord might have better security than WhatsApp, but it will require more compute that vulnerable communities may not possess. If you are building an online grassroots community in India, you will have to resort to using WhatsApp instead.
- There should also be guidelines for transparent communication about such trade-offs and model limitation so that stakeholders are aware of the benefits and risks of using the AI system.

Explainability is the crux of user feedback

In India, due to the lack of standardization and the lack of resources, developers rely on untraditional processes to get the same results. It becomes critical to spearhead explainability to ensure such processes get their due feedback.

- For instance, in Indian languages, libraries for simple transformations like ‘stopwords’ removal and ‘stemming’ are not available. Thus, sometimes, Indian developers create their own algorithms for normalization. If this process is done incorrectly, it can lead to losing context or meaning. Thus, leveraging the principle of explainability to implement ethical practices in designing and documenting normalization algorithms, you can build long-term trust with domain experts, tech communities, and users.

Open Source AI

India needs to foster Open Source AI opportunities and enable its growth.

- Open Source democratizes AI, especially in a resource-constrained environment like India.
- There is a need for open flows and collaborations between the large development community in India to iterate accessible tools to address challenges at a local level.
- Community plays a key role in propagating Open Source. Organisations like FOSS United help incentivise the growth of Open Source by building robust communities.
- While amplifying Open Source tools, biases, exclusion, and misuse need to be kept in check. There is also a need to employ better data protection methods.



Safety & Security

What is it?

Safety and security ensure AI systems operate reliably, predictably, and securely, minimising risks of harm to individuals, organisations, or society. While safety means calibrating AI systems towards mitigating broad physical, digital, ethical, and psychological risks to humans, security entails protecting them from internal or external threats. With increasingly advanced cybersecurity threats in today's digital space, it has become important to foresee risks and make every effort to mitigate them. Combatting ecosystem concerns like deepfakes, misinformation, and spam will also allow the global AI ecosystem to take a revolutionary step towards building long-term and large-scale public trust.

Why should you care?

Creating safe, secure, and technically robust AI should be prioritised from the very beginning to ensure the reproducibility, trustworthiness, and scalability of the AI system. Meeting regulatory safety standards can allow an AI system to become a reliable industry product.

Key Concepts

Reliability

AI systems should function as intended, producing consistent and predictable results across diverse conditions.

Harm prevention

AI systems should include mechanisms to detect, prevent, and correct errors that could lead to harm. Such mechanisms can include robust governance structures, data traceability, adversarial testing, continuous monitoring, and feedback systems.

Human oversight

The most reliable way to create a safe AI system is to ensure human-in-the-loop mechanisms at crucial stages of the AI lifecycle.

Cyber-security

AI systems should be safeguarded against hacking, data breaches, and other malicious attacks. By restricting access, data misuse should be prevented.

Risks it addresses

Cyber threats

E.g. Malicious actors seek to damage an AI model for their own advantage"

Misaligned values

E.g. An AI system keeps generating pornographic images, considering that as the most wanted output

Unauthorised use

E.g. TaskRabbit, IKEA's online marketplace, was hacked using AI to gain unauthorised access to large-scale business data.

Unpredictable behaviours

E.g. An agricultural AI tool failing during monsoon periods as it was not trained for Indian weather, practices, and geography

Emergent risks

E.g. An AI tool bypasses scheduled maintenance of robots to increase short-term productivity but puts humans interacting with them at risk

Manipulation

E.g. A facial recognition AI system cannot resist being fooled by altered images



SAFETY & SECURITY

Preproduction

Efforts put into understanding safety and security risks can go a long way into building long-term stakeholder trust

Identify possibilities where safety can be compromised

Before modelling, analyse how the AI system can be used in your ecosystem and what safety and security concerns can arise from such interactions. Conduct background research and consult domain experts.

- Analyse patterns of use among stakeholders of the problem ecosystem
- Recognise how your AI system can be misused
- Identify existing cases of security breaches in your ecosystem
- Comprehend the regulatory landscape to meet safety and security standards

Understand risks to your lifecycle

Ensure safety and security audits throughout the lifecycle of your AI system and categorise them into types

Anticipate risks:

Ethical
Legal
Social

Assess the severity of the threats:

Categorise risks
Use risk matrices to map severity, likelihood, impact
Prioritise risks and document them

Plan mitigation:

Study countermeasures and failsafes

Set up monitoring systems

Tools: Microsoft Threat Modelling Tool, MITRE ATLAS

Bring in domain experts and security professionals to guide this process



Protect the data

- a. [Multifactor authentication](#)
- b. [Role-based access control \(RBAC\)](#)
- c. [Network segmentation and encryption using SSL/TLS, VPNs](#)
- d. [Data Loss Prevention \(DLP\) tools](#)

Define protocols for a safe lifecycle

Create a robust structure to safeguard the AI system throughout its lifecycle, ensuring that it operates securely, ethically, and effectively while being resilient to evolving risks

Roles and responsibilities

Define who will be responsible (eg. AI Safety Officer) and ensure their training accordingly

Governance frameworks

Use established frameworks [IEEE's Ethically Aligned Design](#), [NIST AI Risk Management Framework](#), and industry standards like [IEC 27001](#)

Approval processes

Set up risk assessment and approval processes and checkpoints at each lifecycle point

Response playbooks

Prepare responses like rollback mechanisms and isolation procedures for identified breaches

Prepare a guiding document that includes:

Company values | Process | Identified risks | Risk mitigation strategies | Protocols with internal and external stakeholders



Analysing such risks ensures reliable operation, safeguards data, maintains compliance, fosters trust, and mitigates vulnerabilities

Instill traceability in datasets and processes

Data Version Control

Optimised for tracking large datasets and ML-specific workflows, DVC enables reproducibility, collaboration, and data change management

Data Provenance

It provides detailed history of origins, lifecycle, and transformations of data, ensuring transparency and trust

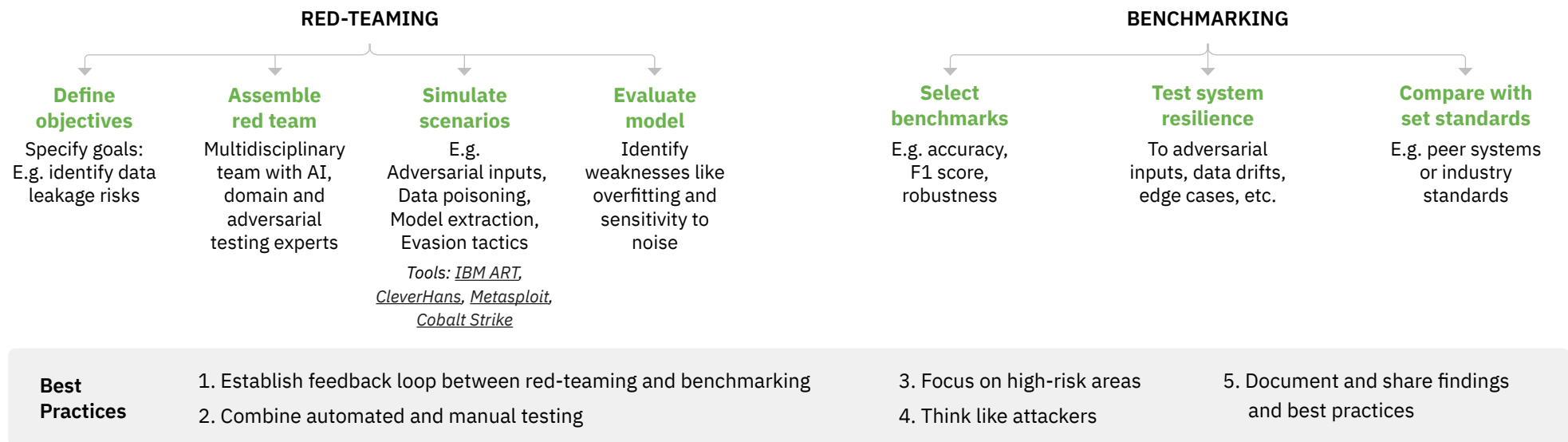
Data Lineage

It tracks and visualises the data lifecycle, focusing on flow and dependencies, giving a broader overview than data provenance

Audits and Logs

Detailed records are essential in tracking access and activities, allowing for easier root cause analysis

Use red-teaming and benchmarking for evaluations





Create human oversight and control measures

Human-in-the-loop systems (HITL) to ensure safety:

- Involve humans at critical points like data labeling and model validation, while ensuring proper compensation and support systems for them
- Involve domain experts to review outputs
- Define confidence threshold that require human review, e.g. diagnosing output with <90% confidence
- Design protocols to flag and report anomalies
- Use tools like [SHAP](#) to explain decisions
- Use dashboards to track and evaluate metrics
- Develop escalation protocols/ manual overrides

Reinforcement Learning with Human Feedback (RLHF)



Human feedback to train foundational models to value safety in decision-making:

- Collect human feedback on outputs through surveys, labelling tasks, direct evaluations
- Adjust the reward model to reflect human values
- Evaluate performance using human-defined goals
- Update as requirements allow

Tools: [RL4LMs](#), [Spinning Up](#), [RLlib](#), [Appen RLHF](#), [Scale](#)

Develop capacity for continuous validation

Real-time Monitoring

- KPIs
- Data drift
- Concept drift
- Model outputs
- Data pipelines

Iterative Monitoring

- Feedback loops
- Retraining pipelines
- Compliance audits
- Robustness audits

Anomaly Detection

- Statistical methods
- ML- based methods
- HITL methods
- Automated alerts

Risk Management

- Personnel training
- Explainability tools
- Stress testing
- Regular patches and updates

Contain the incident

Understand the incident:

Use monitoring systems to detect anomalies and classify severity by determining impact (e.g. data leakage, model corruption or service disruption)

Tools: [Prometheus](#), [Tenable Nessus](#), [Datadog](#), [Evidently AI](#), [Snort](#), [Suricata](#), [MLFlow](#)

Contain the threat

Block unauthorised access, temporarily shut down vulnerable endpoints. Terminate ongoing adversarial attacks, such as denial-of-service (DoS) or adversarial input submissions

Tools: [Azure AD](#), [Okta](#), [Qradar](#), [Elastic Security](#), [Kong Gateway](#), [AWS IAM](#)

Isolate affected components:

Quarantine compromised components (e.g. APIs, models, or datasets) and implement network segmentation

Tools: [Cisco ASA](#), [Vmware NSX](#), [Docker](#), [Kubernetes](#), [MLFlow](#)

Implement system isolation

Switch to backup systems, implement air-gapped isolation, move affected components to isolated environments for further analysis

Tools: [AWS Backup](#), [Azure Site Recovery](#), [Cuckoo Sandbox](#)

Address the root cause

Gather evidence

Collect logs & metrics, identify suspicious patterns, and identify stakeholders to define the issue's scope

Identify causes

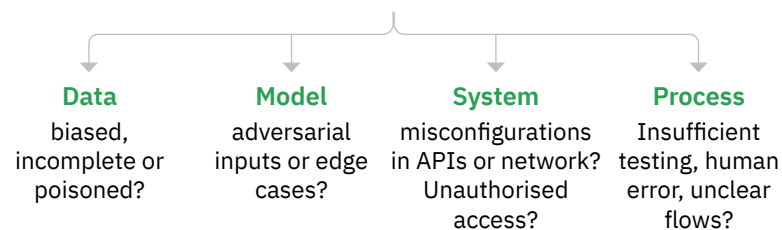
Use fault trees or fishbone diagrams to map contributing factors. Use RCA platforms like [TapRoot](#)

Investigate causes

Systematically evaluate each and test hypothesis in controlled environment

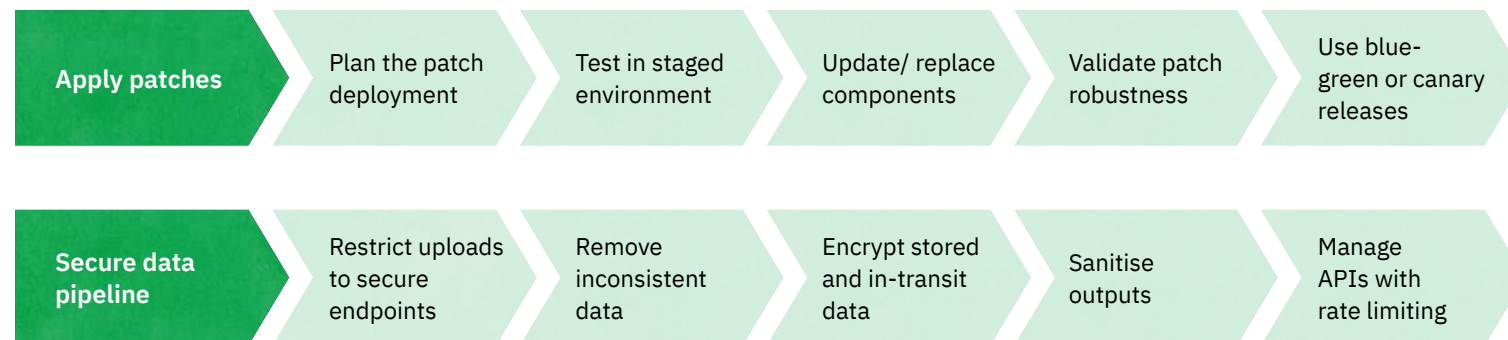
Validate findings

Validate root cause hypotheses by implementing fixes in a test environment, using [Docker](#) or [Kubernetes](#)





Patch and secure the data lifecycle



TOOLS

[Snyk](#), [Tenable Nessus](#), [ART](#), [Atera](#), [Docker](#), [Kubernetes](#), [NinjaOne](#)

TOOLS

[TFDV](#), [OpenSSL](#), [AWS KMS](#), [VeraCrypt](#), [Pachyderm](#), [Kubernetes](#)

Strengthen system safeguards post-breach



Stress test the patch

Tools: [Chaos monkey](#), [IBM ART](#), [Prometheus](#)



Train personnel and revise playbooks

Tools: [KnowBe4](#), [AttackIQ](#), [Cymulate](#)



Analyse incident collectively

Tools: [Confluence](#), [Notion](#), [Miro](#), [Slack](#)



Implement full-system audit

Tools: [Splunk](#), [ELK Stack](#), [Azure Security Center](#)



Deploy stronger security measures and policies

Tools: [Azure AD](#), [Okta](#), [Terraform](#)



Build long-term resilience

Tools: [Safe Security](#)



Inform affected stakeholders

Tools: [StatusPage](#)



Safety & Security for India

Look beyond just technical safety

India is a complex country where the risk assessment of an AI system can go beyond technical and individual safety concerns. It is necessary to consider broader social implications in terms of AI safety.

- Consider how your AI system will interact with existing sociocultural dynamics and power structures.
- Conduct regular risk assessments to determine potential harms like job displacement, misinformation, or erosion of trust.
- Monitor the system's impact on society, both positive and negative.
- Develop crisis response plans to address unintended consequences or harmful outcomes.
- Establish capacity building and training programmes for developers and other stakeholders to promote ethical AI practices.
- Inform the public at large about benefits and risks of AI.
- By establishing whistleblower protections, encourage stakeholders to report concerns without fear of reprisal.
- Maintain ESG as one of the long-term goals for the AI system and share best practices with the tech community

Simplifying monitoring and evaluation (M&E)

Considering smaller startups might not have the capacity to set up proper oversight infrastructure, M&E can be time-consuming and complex. Some best practices that can simplify the M&E process are as follows:

- Focus evaluation efforts on critical parts of the AI system and key metrics first. Expand gradually to other parts and metrics as the AI system scales.
- Instead of creating new infrastructure for monitoring and evaluation (e.g. dashboards) from scratch, explore existing open-source tools in the market. Many cloud providers also offer free tiers that can provide useful solutions.
- If tools prove too expensive, develop lightweight custom scripts to monitor your AI systems. For example, use Python libraries like Pandas and Scikit-learn to calculate metrics. Python can also be useful in implementing statistical methods for detecting anomalies and drift.

Documentation is the first step towards safety

Indian startups might feel constrained to implement security measures, both in terms of the learning curve and resources required. Some quick simple first steps that can be prioritised:

- Start by investing in proper documentation using data provenance techniques. Documenting data sourcing and data transformations allow for easier interventions later.
- Track performance variations across different model generations, and across different demographics. Log modifications made for different use-cases.
- If unable to carry out resource-intensive red-teaming processes, try to establish robust feedback loops by involving subject and/or technical experts.
- Similarly, HITL structures can be resource-intensive. Getting participation from wider pool of stakeholders like AI ethicists, consultants, peers, community members and students can allow for receiving and maintaining feedback.

SECTION 4

Macro Recommendations



Macro Recommendations



Shared repositories of resources

Sharing resources like tools, knowledge, frameworks, and practices will help smaller players overcome resource gaps. Open-source libraries, checklists, templates and domain-specific and localised case studies will tailor practices to industry standards as well as to legal, cultural and contextual nuances.



Community of Practitioners

Collective and bottom-up problem-solving, cross-disciplinary inputs, and peer accountability will democratically diffuse RAI. Sharing successes and failures will help developers avoid common pitfalls and replicate effective strategies. This approach will empower developers to negotiate better with regulation and to address emergent risks.



Collaboration between relevant actors

This will ensure that regulatory oversight and community insights shepherds AI towards responsibility without stifling innovation. Techniques like sandboxes allow developers to help policymakers establish realistic and informed regulation and policymakers to standardise best practices. Network building between different actors will ensure multi-stakeholder knowledge exchange and feedback loops.



Better handholding and training

Industry leaders, pioneers and early adopters can diffuse RAI awareness among developers. Academics, policymakers, researchers, and subject experts can also contribute to better handholding. Organisations can train personnel while maintaining iterative engagement with community and experts.



Certifications and external standards

Increasing public trust in AI systems will require advocating and advertising of RAI. Guidelines and certifications verified independently will incentivise developers to adhere to RAI. Exhibiting commitment to RAI will promise competitive advantage, expanded market reach, and a healthier AI ecosystem.

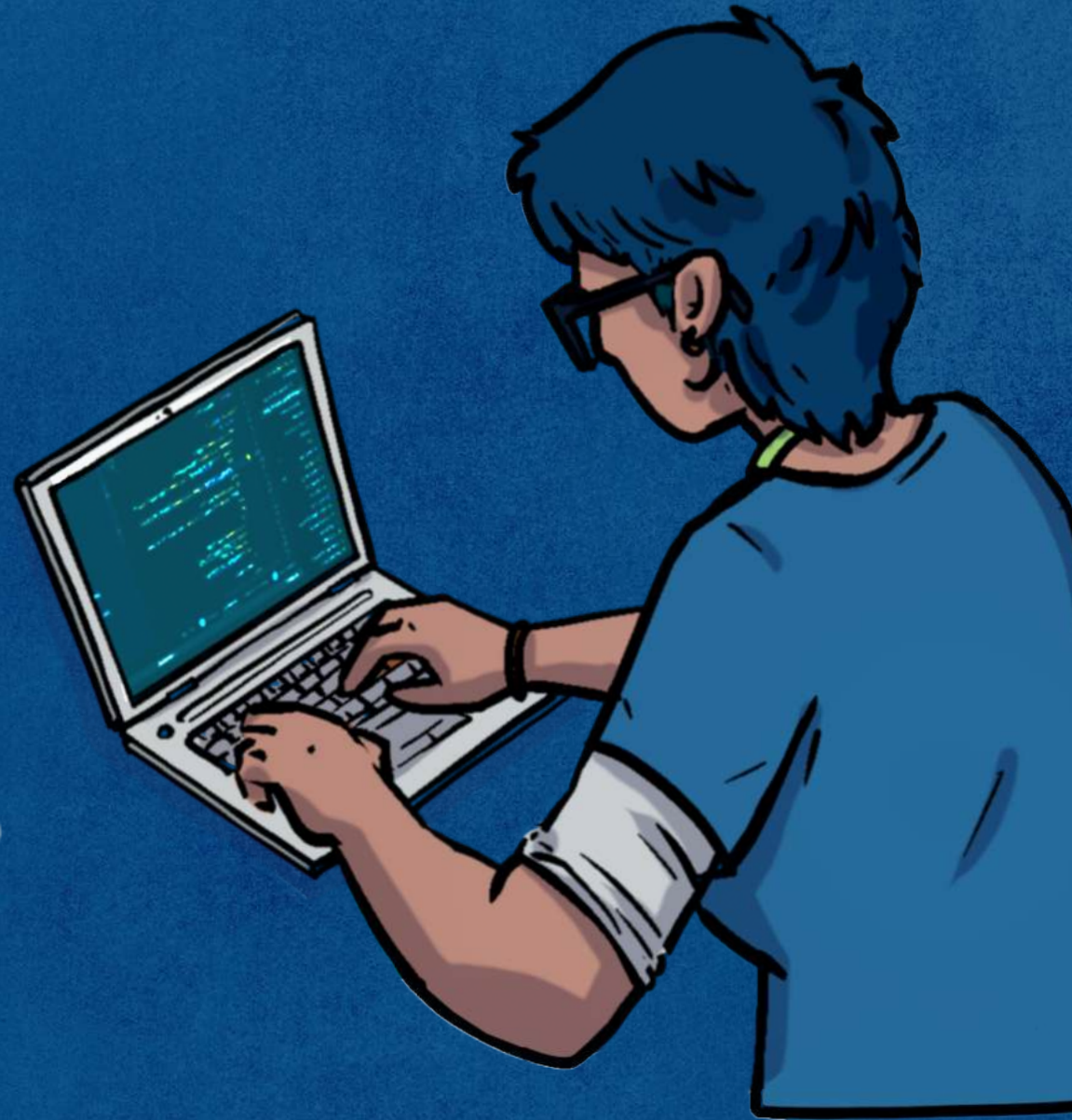


Greater public awareness

Public awareness can create a cultural shift that demands ethical AI systems. Broader discussions and transparent communication through methods like reporting, social media, and interactive tools and platforms will inform the public of the potential and limitations of AI, combatting unfounded fear mongering.

SECTION 5

A Short Note to Developers



[TABLE OF CONTENTS](#)

A Short Note to Developers

Dear Developers,

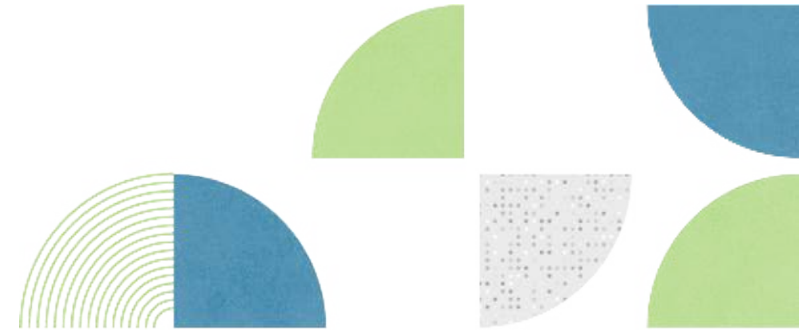
As creators of transformative technology, you carry the profound responsibility of ensuring that innovations serve humanity ethically and inclusively. Responsible AI is not merely a technical goal—it is a moral imperative that demands collaboration, foresight, and transparency.

AI’s potential is vast, but so are its risks. Bias in data, lack of accountability, or insufficient awareness can have real-world implications. Addressing these challenges requires us to work together across disciplines, incorporating diverse perspectives from ethicists, designers, researchers, and users.

Effective communication is central to this endeavour. Engaging stakeholders early and often can help anticipate concerns and foster trust. Documenting decisions, openly sharing methodologies, and welcoming scrutiny from peers and communities strengthen the integrity of your work.

Responsible AI is a shared journey, not a solitary task. By embracing collaboration and maintaining open channels of communication, you can create AI systems that not only push technological boundaries but also enhance societal well-being. Let’s continue to build tools that inspire trust, promote fairness, and make a lasting positive impact.

Together, you can lead the way in defining what ethical technology looks like for future generations.



Aapti is a public research institute that works at the intersection of technology and society.
Aapti examines the ways in which people interact and negotiate with technology both offline and online.

contact@aapti.in | www.aapti.in

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 2.5 India License.

View a copy of this license at creativecommons.org/licenses/by-nc-sa/2.5/in/