THE DEVELOPER'S PLAYBOOK

for **Responsible AI** in India

November 2024

ACKNOWLEDGEMENTS & CREDITS

This playbook was developed under the direction and oversight of **Sangeeta Gupta** (Senior Vice President and Chief Strategy Officer at nasscom), **Ankit Bose** (Head of AI at nasscom), **Pravin Anand** (Managing Partner at Anand and Anand), and **Dr. Ajai Garg** (Director of Digital Tech and Law Group at Anand and Anand).

Raj Shekhar (Responsible AI Lead at nasscom), and **Vibhav Mithal** (Associate Partner at Anand and Anand) led the research and drafting, with support from **Simrandeep Singh** (Associate, Responsible AI at nasscom), and **Alvin Antony** (Associate, Tech and Law at Anand and Anand).

The playbook greatly benefited from the insights and feedback of the following stakeholders and subject matter experts: **Abhik Chaudhuri**, Digital Transformation Governance expert at Tata Consultancy Services • **Dr. Abhijnan Chakraborty**, Assistant Professor, Department of Computer Science & Engineering at Indian Institute of Technology Kharagpur • **Abilash Soundararajan**, Founder and CEO at PrivaSapien • **Abhishek Upperwal**, Founder and CEO at Socket Labs • **Ajit Ashok**, Data & AI Excellence Lead at Philips Innovation Campus • **Akbar Mohammed**, Head of Fractal Dimension at Fractal • **Amrita Sengupta**, Research and Program Lead at Centre for Internet and Society • **Dr. Amritendu Mukherjee**, Co-founder and CTO at NeuroPixel.AI Labs • **Anand Venkatanaryanan**, CTO at DeepStrat • **Andeed Ma**, President and Chairperson of Governing Council at Risk and Insurance Management Association of Singapore • **Anwesha Sen**, Assistant Programme Manager at The Takshashila Institution • **Arti Khanijo**, Product Head at United We Care • **Ashutosh Kumar**, Director and Head of Engineering at Corrosion Intel • **Atul Gandre**, Technology Head at Tata Consultancy Services AI.Cloud Microsoft Practice • **Dr. Avik Sarkar**, Senior Research Fellow and Visiting Faculty at Indian School of Business Mohali • **Avinash Babu M.**, CEO at Samasti Health Technologies • **Axel Beelen**, Legal Consultant in Data Protection and AI • **Ayushi Agarwal**, Head of Data Science & Analytics at United We Care • **Bhavana Mittal**, Co-founder, Executive Director and Chief Growth Officer at Bert Labs • **Bishakha Bhattacharya**, Head of Public Policy at AWS India and South Asia • **Devesh Raj**, Director, AI and Data Science at Fidelity Investments • **Errol Finkelstein**, Co-founder and COO at ChipX Limited • **Jeyandran Venugopal**, Chief Product and Technology Officer at Flipkart • **Jyothi VK**, General Counsel and Senior Vice President at Aditya Birla Fashion and Retail • **Karen Silverman**, Founder and CEO at The Cantellus Group • **Karthik Rao Bappanad**, Cybersecurity and Public Policy Advisor at DeepStrat • **Krity Kansara**, Product Manager, Responsible AI at Bosch Global Software Technologies • **M. Chockalingam**, Technical Director at nasscom ai • **Markus Krebsz**, Project Lead, AI & Other Digital Embedded Technologies, United Nations Economic Commission for Europe • **Merve Hickok**, Founder at AIethicist.org • **Michael Borelli**, Director at AI & Partners • **Nehaa Chaudhari**, Partner at Ikigai Law • **Nita Khare**, Technology Champion at Tata Consultancy Services AI.Cloud Microsoft Practice • **Pavan M. Laxmeshwar**, AI Program Manager at Bosch Global Software Technologies • **Preethika Pilinja**, Senior Manager (Legal) at Zensar Technologies • **Prinkan Pal**, Co-founder and CEO at LEGOAI Technologies • **Ravi Vijaya Raghavan**, Chief Data & Analytics Officer and Senior Vice President at Flipkart • **Rohit Kochar**, Founder, Executive Chairman and CEO at Bert Labs • **Rubal Chib**, Co-founder and CEO at qZense Labs • **Ryan Carrier**, Executive Director at ForHumanity • **Dr. Sachin Baliga**, Consultant Psychiatrist at Fortis Healthcare • **Dr. Saikat Saha**, Technical Director at nasscom ai • **Dr. Sameep Mehta**, Distinguished Engineer at IBM Research • **Satish Grampurohit**, Co-founder and CEO at Cogniquest Technologies • **Saurabh Singh**, Leader, Technology Policy at AWS India & South Asia • **Sendil Kumar**, Senior Director of Advanced Analytics and Data Science at Eli Lilly and Company • **Shweta Gupta**, Senior Director, Engineering at Microsoft • **Dr. Sivaramakrishnan R. Guruvayur**, Chief AI Scientist at aaquarian.ai,

Member, Multistakeholder Experts Group at Global Partnership on AI and Research Advisor at Centre for Responsible AI, Indian Institute of Technology Madras • **Srinivas Aluri**, Founder at Smart Machines and Structures • **Srishti Batra**, Co-founder and CTO at Qzense Labs • **Sourav Banerjee**, Founder and CTO at United We Care • **Subhabrata Debnath**, Co-founder and CTO at Neural Garage • **Sudeshna Mukhopadhyay**, Co-founder and CEO at Intelekt AI • **Sundar Narayanan**, Working Group Member, Standard on AI Robustness Assessment, Department of Telecommunications, Government of India • **Syed Ahmed**, Head of Responsible AI Office at Infosys • **Tanusree De**, Executive Director, AI & Data Practice at Ernst & Young Global Delivery Services • **Tirthankar Choudhuri**, Vice President, Digital Data Science at American Express • **Dr. Tutan Ahmed**, Assistant Professor of Economics and Public Policy at Indian Institute of Technology Kharagpur • **Vishwam Jindal**, Co-founder and CEO at Webnyay

CONTENTS

FOREWORD

In an era where artificial intelligence (AI) is reshaping industries and redefining how we interact with technology, embedding responsible AI practices has never been more crucial. The global focus has moved beyond what AI can achieve to how it should be ethically developed and implemented. Responsible AI is not just about meeting regulatory requirements or securing a competitive edge; it represents a commitment to values prioritising ethics, transparency, accountability, and inclusivity. This commitment ensures that technological advancements do not outpace the principles that safeguard public trust, fostering innovation that is both transformative and equitable.

For India, a country poised to harness the immense potential of AI across diverse sectors such as healthcare, education, finance, agriculture, and public services, the stakes are particularly high. The need for voluntary industry adherence to responsible AI principles is critical. By embedding these principles into core practices, developers and organisations can build an ecosystem that promotes public trust and aligns innovation with societal well-being. This proactive commitment not only prepares the industry for evolving regulations but also positions India as a leader on the global stage for ethical and responsible AI development.

The Developer's Playbook for Responsible AI in India is a step in this direction. Created by nasscom with support from Anand and Anand, this playbook highlights an approach to building a robust, ethical, and inclusive AI landscape. It serves as a guide for developers and businesses to navigate the complexities of responsible AI development. The playbook equips innovators with the guidance they need to identify potential risks, align with best practices, and adopt responsible AI methodologies while allowing for evolving approaches towards AI development and deployment.

Aligned with the IndiaAI Mission's Safe and Trusted AI pillar, the playbook seeks to harmonise public and private sector risk management frameworks, offering a path that integrates ethics into the fabric of AI development. IndiaAI Mission seeks to ensure that the future of AI in India is not only groundbreaking but also secure, inclusive, and socially responsible. By championing responsible AI practices, we are laying the foundation for a future where technological progress supports human dignity and sustainable growth. As technology evolves, feedback from industry, academia, startups, and researchers will further help develop this playbook into an implementable and actionable guide for all.

Sh. Abhishek Singh

Additional Secretary, Ministry of Electronics and Information Technology,
Government of India

PREFACE

This playbook—developed by the nasscom Responsible AI Hub with support from Anand and Anand—underscores the shared commitment of the Government of India and nasscom to establishing a unified industry framework for artificial intelligence (AI) risk identification and mitigation in India. Through this collaboration, we aim to foster safe and trustworthy AI-enabled futures for all.

The playbook is designed to provide a voluntary framework for developers to systematically identify and mitigate the potential risks associated with the commercial development, deployment, and use of AI in India. The risk mitigation guides provided in the playbook contain consolidated AI risk libraries with corresponding sets of risk mitigation prompts, based on our evaluation of existing public and private AI risk management frameworks and hard and soft regulation—most relevant and applicable to the commercial implementation of AI in India. The references section appended to this playbook provides a list of the regulatory and non-regulatory resources that informed the formulation of the risk mitigation guides.

The playbook adopts an inclusive definition of "developers," covering all primary functions integral to AI development and deployment. By organising these functions within a unified AI lifecycle, it seeks to provide a comprehensive perspective on the risks inherent across the development and deployment processes, along with the corresponding risk mitigation prompts to ensure safe and responsible AI practices. However, it is important to recognise that safe and responsible AI practices rely on the coordinated involvement of multiple actors throughout the AI lifecycle and may require different levels of control from various actors over specific AI safety functions. Therefore, the playbook should not be construed as an accountability framework for assigning responsibility among different actors for specific safety functions.

The risk mitigation guides provided in the playbook are designed to be sector- and use case-agnostic and do not prescribe exhaustive guidance on risk mitigation during the development and deployment of any specific type of AI model or application—while acknowledging that precise requirements of AI trust and safety may vary in response to the demands and constraints of a given commercial context. Therefore, none of the risk mitigation prompts in the playbook should be interpreted by developers in an absolute or rigid manner. Developers are advised to refer to the relevant risk mitigation guide and use the risk mitigation prompts within as guidance to implement responsible AI practices—depending on the type of AI they are developing, their role and level of control in the AI lifecycle and the trust and safety requirements relevant to their context.

For instance, certain risk mitigation prompts such as those related to post-deployment monitoring, may not apply to developers of open-source AI models, as they might lack the ability to monitor or control how their models are adapted and used downstream once open-sourced. Similarly, developers of AI models with wide applicability across different contexts might not be able to anticipate all potential uses and users of their models, and therefore, certain risk mitigation prompts, such as those related to ensuring model responsiveness to geographical considerations of language, culture, laws and regulations, etc. may not apply to them.

The playbook uses technical or legal jargon at a minimum, only when necessary to maintain precision or accuracy. The *glossary* section contains the definitions of the jargon and special terms used in the playbook.

Users can expect future versions of this playbook, given the rapidly evolving landscape of AI risks and safety practices.

GLOSSARY

NOTE: The definitions of technical terms used in the playbook do not necessarily correspond to any one or more existing frameworks or rules for responsible AI implementation (e.g., OECD's AI principles and Biden Administration's Executive Order on Safe, Secure, and Trustworthy AI). Legal definitions are used only when there is sufficient clarity in the Indian context.

AI application	The term refers to a software solution that leverages a discriminative AI or a generative AI model, or a combination of both, to perform specific or general tasks.
Artificial intelligence (or AI)	The term refers to any machine learning model or an AI application designed using such a model to perform specific or general tasks.
Data principal	The term refers to a natural person to whom the personal data relates—as per the Digital Personal Data Protection Act, 2023.
Developer	The term refers to any natural person who designs, builds and implements AI models and/or applications and is engaged in one or more of the following activities: algorithm development, data handling, model training or fine-tuning, software engineering, model deployment and monitoring.
Discriminative AI model	The term refers to a machine learning model designed to classify data points or predict outcomes by learning the decision boundary that separates different classes.
Generative AI model	The term refers to a deep learning model designed to generate artefacts such as image, text, audio, video, and various forms of multi-modal content. Such a model can be designed either for a specific task or for deployment and use across multiple contexts (e.g., multimodal models like OpenAI's GPT 4).

GLOSSARY

Grievance

The term refers to any complaint alleging a violation of applicable laws, terms of use and organisational policies for safe and responsible AI practices, contractual terms for AI procurement and use, etc.

Non-personal data

The term refers to any data which is not personal data.

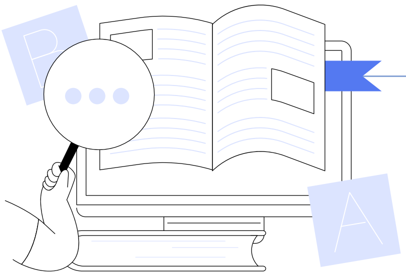
Personal data

The term refers to any data about an individual who is identifiable by or in relation to such data as per the Digital Personal Data Protection Act, 2023.

Publicly available personal data

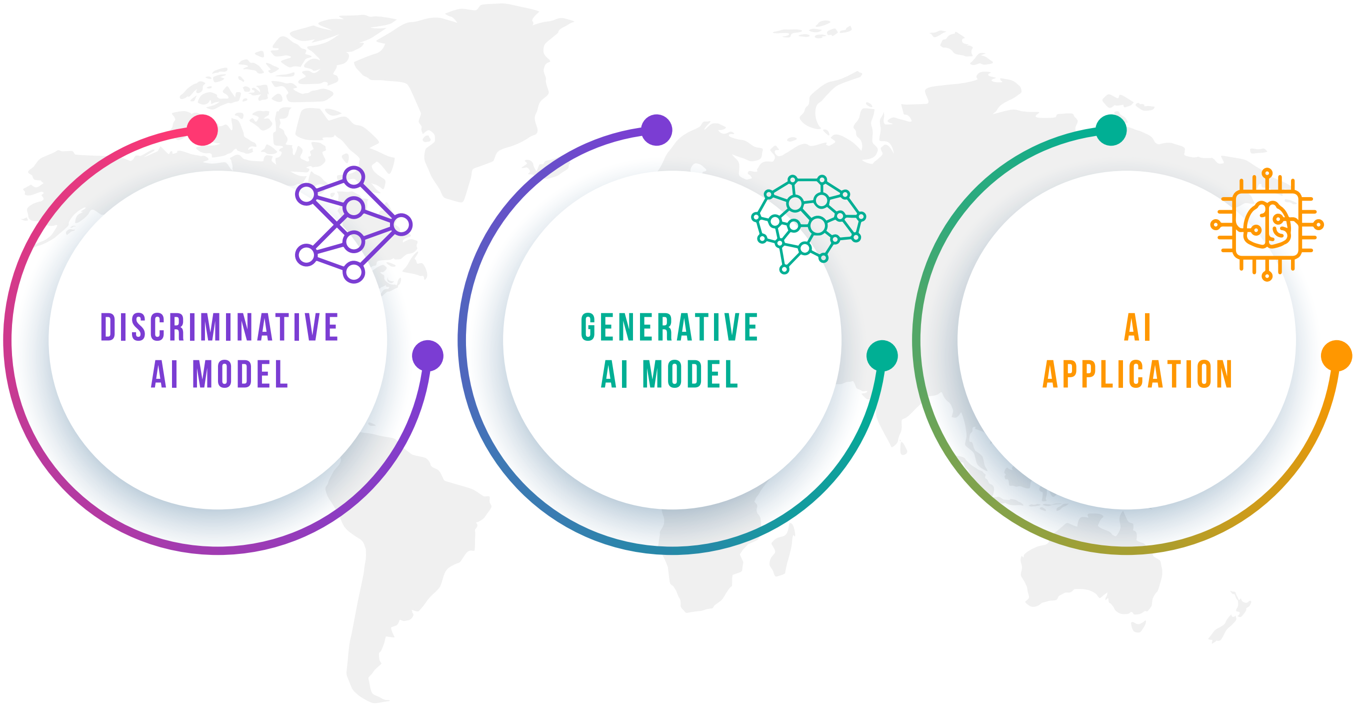
The term refers to personal data that is made or caused to be made publicly available by (a) the data principal to whom such personal data relates; or (b) any other person who is under an obligation under any law for the time being in force in India to make such personal data publicly available—as per the Digital Personal Data Protection Act, 2023.

Note that "publicly available" is not defined under Digital Personal Data Protection Act, 2023; if certain data qualifies as "publicly available personal data",the Digital Personal Data Protection Act, 2023 shall not apply to such data.

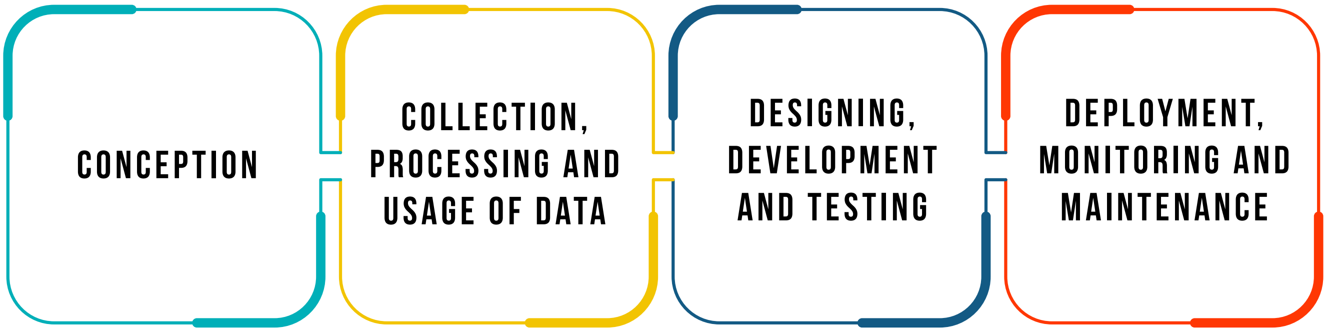


HOW TO NAVIGATE THE PLAYBOOK

The playbook is divided into three risk mitigation guides, each focusing on a particular AI type (see *glossary* for a definition of each of the AI types listed below):

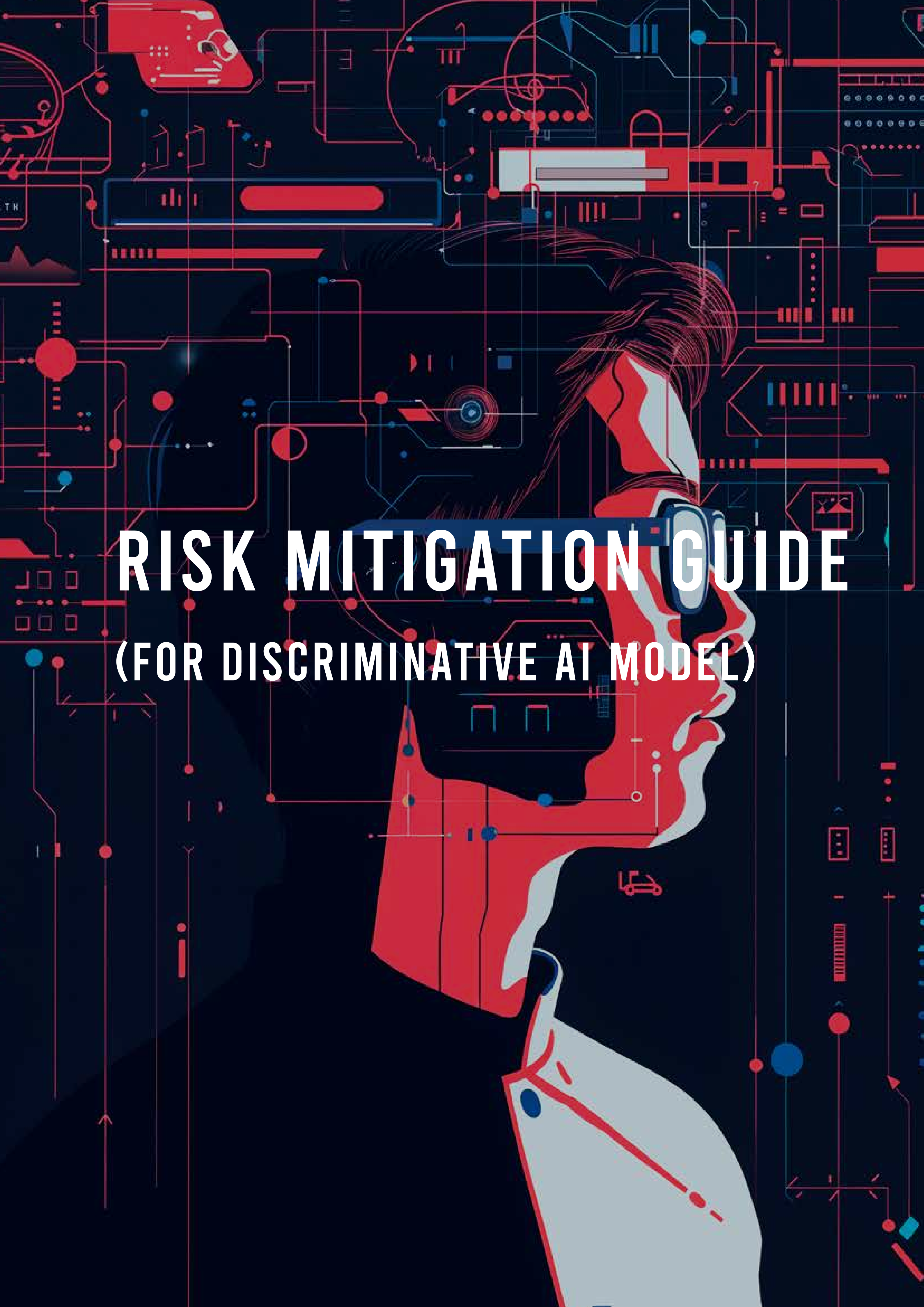


Each risk mitigation guide provides a comprehensive library of potential risks associated with developing a specific type of AI, along with tailored risk mitigation prompts organised across the following commonly recognised stages of the AI lifecycle:


CONCEPTION
COLLECTION, PROCESSING AND USAGE OF DATA
DESIGNING, DEVELOPMENT AND TESTING
DEPLOYMENT, MONITORING AND MAINTENANCE

This structure seeks to ensure that risks are effectively addressed at every critical phase of AI development through robust safety practices.

References to relevant guidance for developers on how to effectively respond to risk mitigation prompts are also included, where appropriate.



RISK MITIGATION GUIDE

(FOR DISCRIMINATIVE AI MODEL)

RISK MITIGATION GUIDE
(FOR DISCRIMINATIVE AI MODEL)

STAGE 1 : Conception

Risk A: Model not being fit for purpose
(i.e., not responsive to the context in which it is sought to be deployed)

□ Define and document the intended purpose or use of the model you are conceiving for development and ensure that the model is fit for the intended purpose or use; note that this is notwithstanding that the model may have to undergo multiple iterations before it is deemed fit for purpose.

- □ Identify and describe the stakeholders who are intended to deploy the model and those who are likely to be directly impacted by its functioning; identify the potential benefits and harms for each stakeholder.

Further Guidance: To learn more about how to conduct AI impact assessment, refer to Microsoft's Responsible AI Impact Assessment Template.

- □ Describe the geographic area(s) where the model is intended to be deployed and take measures to ensure it is responsive to the geographical considerations of language, culture, laws and regulations, etc.

Further Guidance: To learn more about how developers can ensure participation of relevant stakeholders and integrate their feedback throughout the AI lifecycle, refer to Centre for Responsible AI, IIT Madras and Vidhi Centre for Legal Policy's paper on Participatory AI Approaches in AI Development and Governance Case Studies and Partnership on AI's Draft Guidelines for Participatory and Inclusive AI.

- □ Describe the data requirements for the intended use of the model and ensure that the required data is available in desired quality, quantity, and format for model development; identify any risks associated with sourcing any specific dataset(s) for model development.

- □ Define the model performance metrics and error types; and prepare an evaluation plan for each of the performance metrics and error types. Note that this is notwithstanding that the performance metrics and error types may have to be revised at later stages of the model lifecycle to meet contextual requirements.

Further Guidance: To learn more about assessing if AI systems are fit for purpose, refer to Microsoft's Responsible AI Standard v2 (see *Goal A3: Fit for Purpose*).

Risk B: Unaccountable development and use of the model

- □ Ensure clarity on the roles and responsibilities of
  - □ external entities that provide services or resources to support the development of the model, (e.g., data providers, cloud service providers, third-party software libraries, etc.), and
  - □ users or stakeholders who interact with the model

  through clearly defined SOPs, valid contracts and licensing agreements.
- □ Ensure clarity on human oversight and control responsibilities from conception to deployment of the model.

STAGE 2 : Collection, processing and usage of data

Risk A: Violation of applicable data privacy regulations

- □ Consult appropriate experts within or outside your company to determine whether any personal data that you intend on using to develop the model qualifies as "publicly available" under the Digital Personal Data Protection Act, 2023. If yes, note that the Digital Personal Data Protection Act, 2023 would not apply to such data. However, verify and document the source of such data.
- □ Consult appropriate experts within or outside your company to determine whether your intended use of any personal data to develop the model would qualify as "legitimate use" under the Digital Personal Data Protection Act, 2023. If yes, note that the consent requirements under the Digital Personal Data Protection Act, 2023 would not apply to such data. However, document this data for review by internal risk assessors and/or external auditors.
- □ If any personal data you intend to use for model development does not qualify as "publicly available" under the Digital Personal Data Protection Act, 2023, and/or if your intended use of such data to develop the model does not meet the criteria for "legitimate use" under the Digital Personal Data Protection Act, 2023, ensure you have the consent from the concerned data principal for using such data for the intended purpose or use of the model.
  - □ The notice for obtaining consent from the data principal should include: particulars of the kind of data being collected, purpose of data collection, manner in which data is being collected, manner of consent withdrawal, manner of usage of collected data, procedure for grievance redressal and manner in which complaint can be made to the data protection board.
  - □ Check the accuracy and completeness of such data, especially if the model under development will be used to make a decision that would affect the data principal.

  - □ If you later decide to use such data for another purpose (such as developing another model), ensure you have the consent to do so from the concerned data principal.

  - □ Determine when it is reasonable to conclude that the specified purpose for which such data was processed is no longer being served. If the specified purpose is no longer being served and if data retention is not necessary to ensure compliance with any law, ensure that the data is erased. If the data erasure is technically infeasible, register your constraints for internal risk assessment and/or external audits. Consult appropriate experts within or outside your company.

  - □ Comply with any requests to correct, complete or update such data. If compliance is technically infeasible, register your constraints for internal risk assessment and/or external audits.

  - □ If the data principal concerned withdraws her consent for the processing of her data:

    - □ determine if processing of such data is required to ensure compliance with any law. If not, cease processing of the data. If it is technically infeasible to do so, register your constraints for internal risk assessment and/or external audits.

    - □ determine if retention of such data is necessary to ensure compliance with any law. If not, ensure erasure of the data. If it is technically infeasible to do so, register your constraints for internal risk assessment and/or external audits.

  Consult appropriate experts within or outside your company.

  - □ Check if your company has put a grievance redressal mechanism in place to address grievances filed by the data principal.

- □ If your company has engaged a data processor to process personal data for model development, confirm the existence of a valid contract defining the terms and conditions of such data processing. Ensure that the data contracts are updated to reflect evolving data practices. Consult appropriate experts within or outside your company.

- □ Deploy safeguards to prevent the leakage of personal data and employ privacy-preserving techniques in data collection and use for model development (e.g., data minimisation, anonymisation, pseudonymisation, abstraction, segregation etc.)

- □ Ensure adherence to data protection regulations in addition to the Digital Personal Data Protection Act, 2023 that may apply to your model development and use (e.g., handling of sensitive personal data in model development for a high-risk context). Consult appropriate experts within and outside your company.

- □ Ensure compliance with applicable data protection regulations during model development is clearly disclosed in model documentation, user-facing policies, procurement documents, etc.

- □ Ensure the model has a privacy policy which complies with applicable laws, such as the Digital Personal Data Protection Act, 2023. Consult appropriate experts within or outside your company.

Further Guidance: To learn how to assess and mitigate data protection-related risks throughout the AI lifecycle, refer to United Kingdom Information Commissioner's Office's AI and Data Protection Risk Toolkit.

Risk B: Generation of biased, dangerous, or illegal outputs

- □ Ensure adherence to a documented procedure for robust data governance to ensure data quality for model development.
  - □ Ensure the data you are using to train the model is of acceptable quality and representative of the stakeholders and geographical considerations identified in stage 1.
  - □ Ensure the data you are using to develop the model is devoid of CSAM (Child Sexual Abuse Material), NCII (non-consensual intimate imagery), or information related to development of chemical, biological, radiological, or nuclear (CBRN) weapons or other dangerous materials or agents.

Further Guidance: To learn about how to identify and mitigate bias in data, refer to nasscom's Responsible AI Architect's Guide (see section on *Data Collection and Processing*).

Risk C: Unauthorised use of non-personal and/or proprietary data

- □ Maintain data catalogs.
- □ Before using any datasets in the possession of your company for model development, ensure you have obtained the necessary internal permissions.
- □ If you are using open-source datasets to develop the model, ensure you comply with the terms and conditions that apply.
- □ If you are sourcing data from a third party to develop the model, ensure that the third party is authorised to license the data and accordingly you have legitimate authorisation to source it.
- □ Consult appropriate experts within or outside your company to ensure compliance with intellectual property regulations applicable to the types of data you intend to use to develop the model.
- □ Ensure appropriate training of the persons involved in model development on responsible data collection, processing, and use, in adherence with applicable regulations, including intellectual property rights laws, etc.

Risk D: Data security breach

□ Check internal security measures to guard against potential breaches of data collected and stored by your company for model development.

□ Check with appropriate experts within or outside your company to ensure that the internal security measures qualify as reasonable security safeguards.

□ Check with appropriate experts within or outside your company if applicable regulations mandate notifying relevant authorities and affected stakeholders of data breaches through specific procedures.

STAGE 3 : Designing, development, and testing

Risk A: Unauthorised, unlawful, or irresponsible use of the model procured from third party

□ Check if your purported use of the model is

- □ authorised by such third party,
- □ compliant with the model usage policies and applicable laws and regulations, and
- □ internal model access guidelines at your company.

□ Review the checks and safeguards that have been put in place by the model developer to pre-empt unauthorised or unlawful output generation by model users.

Risk B: Unauthorised, unlawful or irresponsible use of software code sourced from third party or generated through generative AI solutions

□ If a software code has been sourced from a third party, verify the third party's rights to the code and their authority to grant such rights.

□ If a software code has been procured from a third party, ensure and check if your purported use of the code is:

- □ authorised by such third party, and
- □ compliant with applicable laws and regulations.

□ Ensure you have been able to deploy necessary controls against unauthorised or unlawful use of such software code.

□ If you are using generative AI solutions to generate software code, ensure that you have obtained the necessary internal permissions. Also, apply checks to ensure functionality and hygiene of such code.

Risk C: Compromised security and robustness

□ Adopt reasonable security safeguards against data breaches and unauthorised access to the model.

□ Implement measures to validate the output of the model and tackle inaccurate or unreliable outputs due to glitches or approximations, data bias, model bias etc.

□ Implement measures to handle sudden peaks in workload that may lead to system failures and substantial loss and damage.

□ Implement technical measures to ensure the model is robust and can withstand known adversarial attacks.

   □ Deploy procedures to stress-test the model under different scenarios for unintended harms (e.g., unfair treatment of certain stakeholder groups) that might arise from model deployment.

   □ Employ red teaming to identify model vulnerabilities from a security and ethical standpoint.

   □ Ensure you have documented results from the stress-testing and/or red teaming for review by internal risk assessors and/or external auditors.

   □ Implement measures to maintain the ability to exercise ultimate human control such as circuit breakers, kill switches, or equivalent mechanisms.

Further Guidance: To see how to defend AI systems against adversarial attacks, refer to the Adversarial Robustness Toolbox (ART). To validate the performance of AI systems against international best practices, refer to Singapore Infocomm Media Development Authority's AI Verify framework. To learn about adversarial techniques and mitigation measures, and perform threat assessment and internal red teaming, refer to MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems).

□ Consult with appropriate experts within or outside your company and ensure that the model development and deployment remain in compliance with standards mandated by the sectoral regulator(s).

□ Before model deployment, verify that:

   □ all relevant documents in relation to model development are in order, and

   □ the model produces output in line with its intended use and purpose.

Risk D: Lack of transparency about the model development process, capabilities, limitations, and eventual use among stakeholders

□ Develop documentation to enable public transparency about the model's architecture, capabilities and limitations of the model, computational resources, mechanisms and datasets utilised for training the model, safety and security evaluations conducted on the model, etc. Existing industry best practices for documentation include:

- □ Datasheets that document the data collection and use procedures and related safety and security measures in a model pipeline. Also document dataset characteristics and limitations of the model training and validation data.
- □ Model cards that record the model's purpose, capabilities, limitations, training datasets and evaluation results, and are usually intended for a technical audience.
- □ Factsheets that contain information about the model captured throughout its entire lifecycle, incorporating input from multiple actors in the lifecycle, and is typically tailored to meet the specific needs of the target audience.

Further Guidance: To learn about the guiding principles for designing AI documentation, refer to the CLeAR Documentation Framework for AI Transparency. Refer to nasscom's Responsible AI Architect's Guide (see section on *Prototyping: process and tools to design a responsible system prototype*). To learn more about model cards, refer to Hugging Face's primer on Model Cards. To learn more about factsheets, refer to IBM's AI Factsheets 360.

Risk E: Unexplainable outputs from the model in context(s) that demand explainability

□ Check if the outputs generated by the model that you are developing would need to be explainable, either under applicable laws and/or to adequately meet user expectations. Note that the requirement for model explainability is likely to be most critical in situations where fundamental rights or matters of life and safety are involved.

- □ If the outputs generated by the model that you are developing *must be* explainable, deploy measures to ensure model explainability while ensuring that the model explanations are tailored to the end user's technical proficiency, background and other relevant characteristics.
- □ Ensure human fallback, especially for models intended for deployment in high-risk contexts.

Further Guidance: To learn about how to build explainable AI models, refer to IBM's AI Explainability 360, and Google's Vertex Explainable AI.

Risk F: Model being uninterpretable in context(s) that demand interpretability

- □ Check if the model you are developing would need to be interpretable, either under applicable laws or prevailing ethical mandates. Note that the requirement for model interpretability is likely to be most critical in situations where fundamental rights or matters of life and safety are involved.

- □ If the model that you are developing *must* be interpretable, deploy measures to ensure model interpretability while considering the requirements of the internal risk assessors and/or external evaluators and auditors.

STAGE 4 : Deployment, monitoring and maintenance

Risk A: Uncontrolled deployment of the model

- □ Consider deployment of the model in a phased manner by initially allowing access to a small group of users to enable early detection and mitigation of potential issues before broader deployment to reduce the likelihood of large-scale harms.

- □ If you decide to make the model open source, document the method and terms of its release, including the extent of access to model weights, source code, underlying algorithms, data used, etc.

- □ Implement procedures for continuous monitoring of the model to check for data and model drifts, ensure compliance with applicable ethical and legal requirements or standards for model performance and safety.

  - □ Ensure you have documented results from such monitoring for review by internal risk assessors and external auditors.

- □ Adopt mechanisms for incident reporting and collection of user feedback on model performance and safety.

- □ Adopt measures to reduce the impact on individual privacy in the event of model malfunction.

- □ Ensure there is an appropriate grievance redressal mechanism to resolve issues with the model after its deployment and that there is an internal team to monitor and respond to such grievances.

- □ Assess and document the feasibility of rolling back the model; roll it back, as necessary, as per established procedures and protocols at your company.

Risk B: Unintended or malicious use of the model by third parties leading to adverse impacts

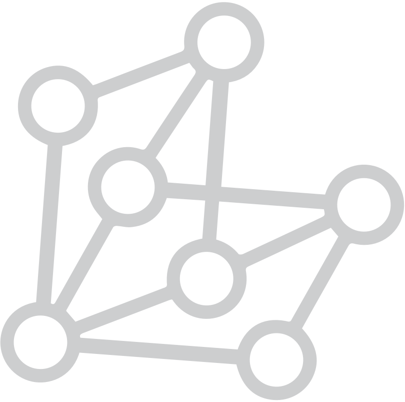
- □ Ensure you have been able to deploy necessary controls against unauthorised, not unauthorized or unlawful use of the model that you have developed.
  - □ Document and publish known limitations of the model, with guidance on its safe and responsible use.
  - □ Check if the terms of service for the model incorporate your concerns about its unauthorised or unlawful use.

  Consult appropriate experts within or outside your company.

- □ Discuss adoption of mechanisms that enable notification of unintended or malicious use or results within the company.

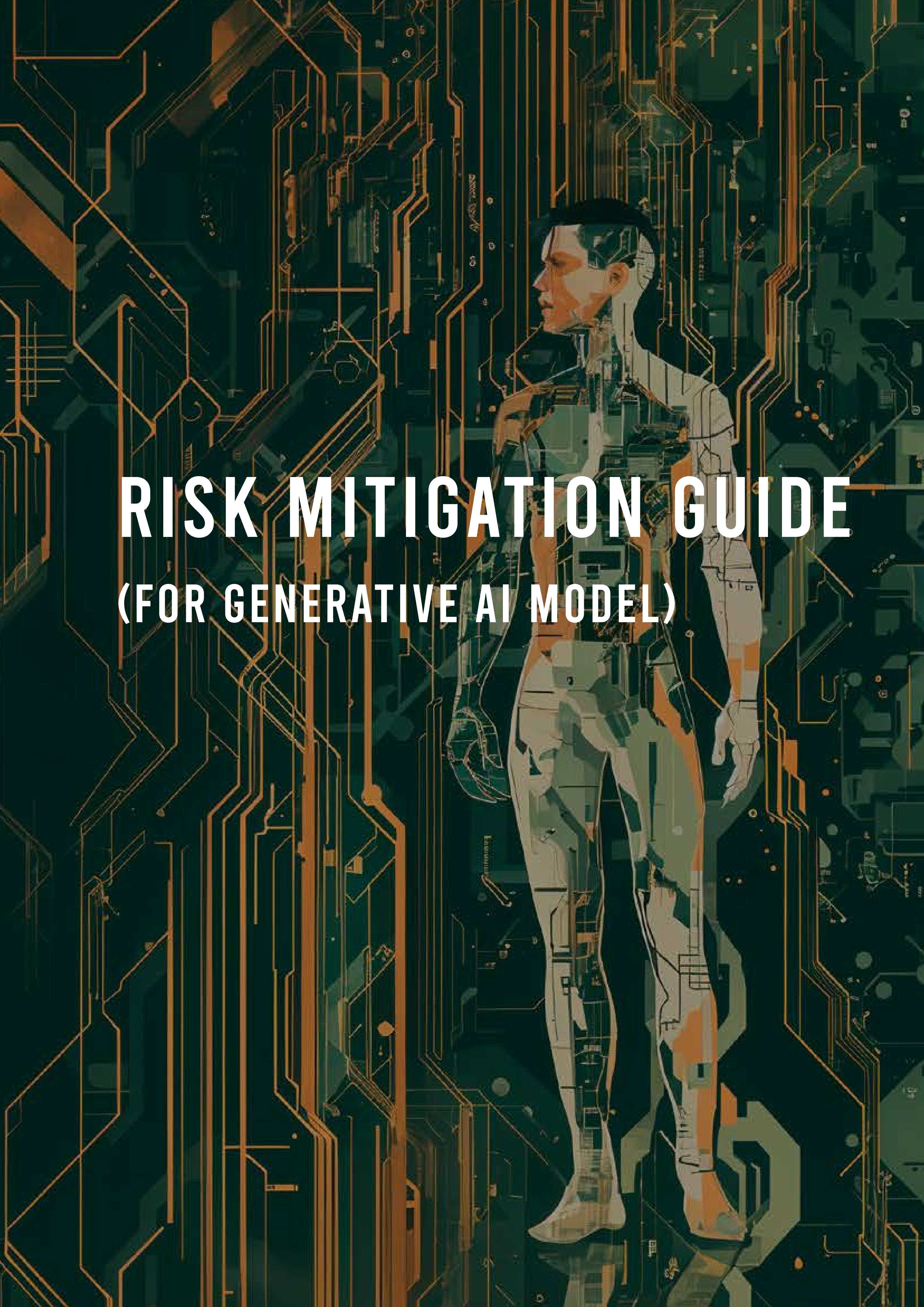
- □ Ensure you have a disaster recovery plan in place in case of an unexpected event with the option of rolling back the model, if necessary. The plan should provide:
  - □ mechanisms to minimise harm to stakeholders and your company, and
  - □ communication channels to facilitate effective interaction amongst stakeholders.

Further Guidance: To learn about best practices for developing and using machine learning systems responsibly refer to nasscom's Responsible AI Architect's Guide and AWS' Responsible Use of Machine Learning (v. 1.2).



RISK MITIGATION GUIDE

(FOR GENERATIVE AI MODEL)



RISK MITIGATION GUIDE
(FOR GENERATIVE AI MODEL)

STAGE 1 : Conception

Risk A: Model not being fit for purpose
(i.e., not responsive to the context in which it is sought to be deployed)

□ Define and document the intended purpose or use of the model you are conceiving for development and ensure that the model is fit for the intended purpose or use; note that this is notwithstanding that the model may have to undergo multiple iterations before it is deemed fit for purpose.

- □ Identify and describe the stakeholders who are intended to deploy the model and those who are likely to be directly impacted by its functioning; identify the potential benefits and harms for each stakeholder.

- □ Describe the geographic area(s) where the model is intended to be deployed and take measures to ensure it is responsive to the geographical considerations of language, culture, laws and regulations, etc.

Further Guidance: To learn more about how developers can ensure participation of relevant stakeholders and integrate their feedback throughout the AI lifecycle, refer to Centre for Responsible AI, IIT Madras and Vidhi Centre for Legal Policy's paper on Participatory AI Approaches in AI Development and Governance Case Studies and Partnership on AI's Draft Guidelines for Participatory and Inclusive AI.

- □ Describe the data requirements for model development and ensure that the required data is available in desired quality, quantity, and format; identify any risks associated with sourcing any specific dataset(s) for model development.

- □ Define the model performance metrics and error types; and prepare an evaluation plan for each of the performance metrics and error types. Note that this is notwithstanding that the performance metrics and error types may have to be revised at later stages of the model lifecycle to meet contextual requirements.

Risk B: Unaccountable development and use of the model

□ Ensure clarity on the roles and responsibilities of

  - □ external entities that provide services or resources to support the development of the model, (e.g., data providers, cloud service providers, third-party software libraries, etc.), and

  - □ stakeholders who are likely to interact with the model, AI education or training provider, marketing partners etc.

  through clearly defined SOPs, valid contracts and licensing agreements.

- □ Ensure clarity on human oversight and control responsibilities from conception to deployment of the model.

STAGE 2 :

Risk A: Violation of applicable data privacy regulations

- □ Consult appropriate experts within or outside your company to determine whether any personal data that you intend on using to develop the model qualifies as "publicly available" under the Digital Personal Data Protection Act, 2023. If yes, note that the Digital Personal Data Protection Act, 2023 would not apply to such data. However, verify and document the source of such data.

- □ Consult appropriate experts within or outside your company to determine whether your intended use of any personal data to develop the model would qualify as "legitimate use" under the Digital Personal Data Protection Act, 2023. If yes, note that the consent requirements under the Digital Personal Data Protection Act, 2023 would not apply to such data. However, document this data for review by internal risk assessors and/or external auditors.

- □ If any personal data you intend to use for model development does not qualify as "publicly available" under the Digital Personal Data Protection Act, 2023, and/or if your intended use of such data to develop the model does not meet the criteria for "legitimate use" under the Digital Personal Data Protection Act, 2023, ensure you have the consent from the concerned data principal for using such data for intended purpose or use of the model.

  - □ The notice for obtaining consent from the data principal should include: particulars of the kind of data being collected, purpose of data collection, manner in which data is being collected, manner of consent withdrawal, manner for usage of collected data, procedure for grievance redressal and manner in which complaint can be made to the data protection board.

  - □ Check the accuracy and completeness of such data, especially if the model under development will be used to make a decision that would affect the data principal.

  - □ If you later decide to use such data for another purpose (such as developing another model), ensure you have the consent to do so from the concerned data principal.

  - □ Determine when it is reasonable to conclude that the specified purpose for which such data was processed is no longer being served. If the specified purpose is no longer being served and if data retention is not necessary to ensure compliance with any law, ensure that the data is erased. If data erasure is technically infeasible, register your constraints for internal risk assessment and/or external audits. Consult appropriate experts within or outside your company.

  - □ Comply with any requests to correct, complete or update such data. If compliance is technically infeasible, register your constraints for internal risk assessment and/or external audits.

  - □ If the data principal concerned withdraws her consent for the processing of her data:

    - □ determine if processing of such data is required to ensure compliance with any law. If not, cease processing of the data. If it is technically infeasible to do so, register your constraints for internal risk assessment and/or external audits.

    - □ determine if retention of such data is necessary to ensure compliance with any law. If not, ensure erasure of the data. If it is technically infeasible to do so, register your constraints for internal risk assessment and/or external audits.

  Consult appropriate experts within or outside your company.

  - □ Check if your company has put a grievance redressal mechanism in place to address grievances filed by data principal.

- □ If your company has engaged a data processor to process personal data for model development, confirm the existence of a valid contract defining the terms and conditions of such data processing. Ensure that the data contracts are updated to reflect evolving data practices. Consult appropriate experts within or outside your company.

- □ Deploy safeguards to prevent the leakage of personal data and employ privacy-preserving techniques in data collection and use for model development (e.g., data minimisation, anonymisation, pseudonymisation, abstraction, segregation etc.)

- □ Ensure adherence to data protection regulations in addition to the Digital Personal Data Protection Act, 2023 that may apply to your model development and use (e.g., handling of sensitive personal data in model development for a high-risk context). Consult appropriate experts within and outside your company.

- □ Ensure compliance with applicable data protection regulations during model development is clearly disclosed in model documentation, user-facing policies, procurement documents, etc.

- □ Ensure the model has a privacy policy which complies with applicable laws, such as the Digital Personal Data Protection Act, 2023. Consult with appropriate experts within or outside your company.

Further Guidance: To learn how to assess and mitigate data protection-related risks throughout the AI lifecycle, refer to United Kingdom Information Commissioner's Office's AI and Data Protection Risk Toolkit.

Risk B: Generation of biased, dangerous, or illegal outputs

□ Ensure adherence to a documented procedure for robust data governance to ensure data quality for model development.

  □ Ensure the data you are using to develop the model is of acceptable quality and representative of the stakeholders and geographical considerations identified in stage 1.

  □ Ensure the data you are using to develop the model is devoid of CSAM (Child Sexual Abuse Material), NCII (non-consensual intimate imagery), or information related to development of chemical, biological, radiological, or nuclear (CBRN) weapons or other dangerous materials or agents.

Risk C: Unauthorised use of non-personal and/or proprietary data

□ Maintain data catalogs.

□ Before using any datasets in the possession of your company for model development, ensure you have obtained the necessary internal permissions.

□ If you are using open-source datasets to develop the model, ensure you comply with the terms and conditions that apply.

□ If you are sourcing data from a third party to develop the model, ensure that you have the legitimate authorisation to source it.

□ Consult appropriate experts within or outside your company to ensure compliance with intellectual property regulations applicable to the types of data you intend to use to develop the model.

□ Ensure appropriate training of the persons involved in application development on responsible data collection, processing, and use, in adherence with applicable regulations, including intellectual property rights laws, etc.

Risk D: Data security breach

□ Check internal security measures to guard against potential breaches of data collected and stored by your company for model development.

□ Check with appropriate experts within or outside your company to ensure that the internal security measures qualify as reasonable security safeguards.

□ Check with appropriate experts within or outside your company if applicable regulations mandate notifying relevant authorities and affected stakeholders of data breaches through specific procedures.

STAGE 3 : Designing, development, and testing

Risk A: Unauthorised, unlawful, or irresponsible use of the model procured from third party

- □ Check if your purported use of the model is
  - □ authorised by such third party,
  - □ compliant with the model usage policies (e.g., OpenAI's Usage Policies, Anthropic's Usage Policies, Google's Generative AI Prohibited Use Policy, Meta's Llama 3 Acceptable Use Policy), and applicable laws and regulations, and
  - □ internal model access guidelines at your company.
- □ Review the checks and safeguards that have been put in place by the model developer to pre-empt unauthorised or unlawful output generation by model users.

Risk B: Unauthorised, unlawful or irresponsible use of software code sourced from third party or generated through generative AI solutions

- □ If a software code has been sourced from a third party, verify the third party's rights to the code and their authority to grant such rights.
- □ If a software code has been procured from a third party, ensure and check if your purported use of the code is
  - □ authorised by such third party, and
  - □ compliant with applicable laws and regulations.
- □ Ensure you have been able to deploy necessary controls against unauthorised or unlawful use of such software code.
- □ If you are using generative AI solutions to generate software code, ensure that you have obtained the necessary internal permissions. Also, apply checks to ensure functionality and hygiene of such code.

Risk C: Compromised security and robustness of the model

- □ Adopt reasonable security safeguards against data breaches and unauthorised access to the model.
- □ Implement measures to validate the model and tackle generation of inaccurate or unreliable outputs by the model due to glitches or approximations, data bias, model bias etc.

Further Guidance: To test if your LLMs generate harmful or offensive outputs, refer to Meta's Llama Guard and Prompt Guard.

- □ Implement measures to handle sudden peaks in workload that may lead to system failures and cause substantial loss and damage.
- □ Implement technical measures to ensure the model is robust and can withstand known adversarial attacks.
  - □ Deploy procedures to stress-test the model under different scenarios for unintended harms (e.g., unfair treatment of certain stakeholder groups) that might arise from the use of the model.
  - □ Employ red teaming to identify model vulnerabilities from a security and ethical standpoint.
  - □ Ensure you have documented results from the stress-testing and/or red teaming for review by internal risk assessors and/or external auditors.
  - □ Implement measures to maintain the ability to exercise ultimate human control such as circuit breakers, kill switches, or equivalent mechanisms.

Further Guidance: To see how to defend AI systems against adversarial attacks, refer to the Adversarial Robustness Toolbox (ART). To measure if your LLMs are resilient to adversarial prompts, refer to PromptBench, and Meta's CyberSecEval. To learn about adversarial techniques and mitigation measures, and perform threat assessment and internal red teaming, refer to MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems).

- □ Before release of the model verify that:
  - □ all relevant documents in relation to model development are in order, and
  - □ unapproved changes are not made in the model.

Risk D: Lack of transparency about the model's development process, capabilities, limitations, and downstream use

- □ Develop documentation to enable public transparency about the model's architecture, capabilities and limitations, computational resources, mechanisms and datasets utilised for model development, safety and security evaluations conducted on the model, etc. Existing industry best practices for documentation include:

- □ Datasheets that document the data collection and use procedures and related safety and security measures in a model pipeline. Also document dataset characteristics and limitations of the model training and validation data.

- □ Model cards that record the model's purpose, capabilities, limitations, training datasets and evaluation results, and are usually intended for a technical audience.

- □ Factsheets that contain information about the model captured throughout its entire lifecycle, incorporating input from multiple actors in the lifecycle, and is typically tailored to meet the specific needs of the target audience.

Further Guidance: To learn about the guiding principles for designing AI documentation, refer to The CLeAR Documentation Framework for AI Transparency. To learn about model cards, refer to Hugging Face's primer on Model Cards. To learn more about factsheets, refer to IBM's AI Factsheets 360.

- □ Make reasonable efforts to implement technical mechanisms for ensuring that the outputs generated using the model are labelled as AI-generated.

Further Guidance: To learn about tools and techniques for AI watermarking, refer to Hugging Face's AI Watermarking 101: Tools and Techniques

STAGE 4 : Deployment, monitoring and maintenance

Risk A: Uncontrolled deployment of the model

- □ Consider deployment of the model in a phased manner by initially allowing access to a small group of users to enable early detection and mitigation of potential issues before broader deployment to reduce the likelihood of large-scale harms.

- □ If you decide to make the model open source, document the method and terms of its release, including the extent of access to model weights, source code, underlying algorithms, data used, etc.

Further Guidance: To learn about the key considerations for guiding decision-making around the extent of model access, refer to Irene Solaiman's paper on The Gradient of Generative AI Release: Methods and Considerations.

- □ Implement procedures for continuous monitoring of the model to check for data and model drifts, ensure compliance with applicable ethical and legal requirements or standards for model performance and safety.

    - □ Ensure you have documented results from such monitoring for review by internal risk assessors and external auditors.

- □ Adopt mechanisms for incident reporting and collection of user feedback on model performance and safety.

- □ Ensure there is an appropriate grievance redressal mechanism to resolve issues with the model after its deployment and that there is an internal team to monitor and respond to such grievances.

- □ Assess and document the feasibility of rolling back the model, if necessary, as per established procedures and protocols at your company.

Risk B: Unintended or malicious use of the model by third parties leading adverse impacts

- □ Internally discuss and identify prohibited uses (such as generating artefacts prohibited under prevailing laws).

- □ Ensure you have been able to deploy necessary controls against unauthorised or unlawful use of the model.

    - □ Document and publish known limitations of the model, with guidance on its safe and responsible use.

    - □ Check if the terms of service, and usage policies for the model incorporate your concerns about its unauthorised or unlawful use.

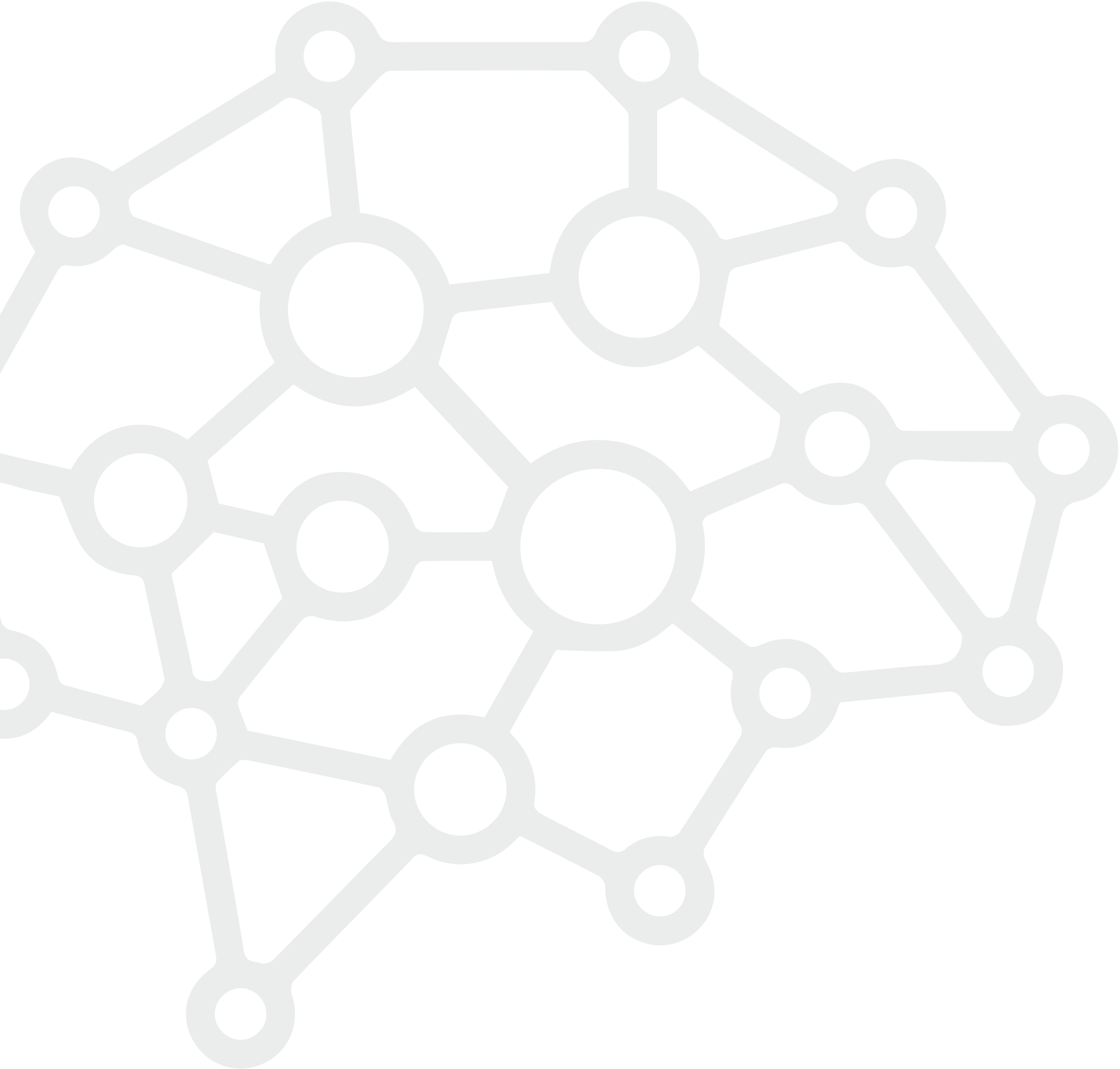
  Consult appropriate experts within or outside your company.

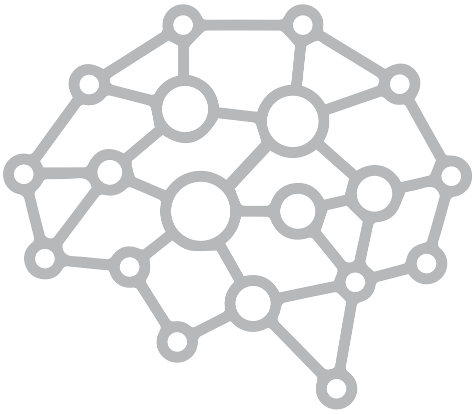
Further Guidance: To learn more about how to develop guidance on safe and responsible use of the model, refer to Meta Llama 3's Acceptable Use Policy, OpenAI's Usage Policies, Anthropic's Usage Policies, Google's Generative AI Prohibited Use Policy, Sarvam AI's Terms of Use, and Krutrim's Acceptable Use Policy.

- □ Discuss and define mechanisms that enable notification of unintended or malicious use or results within the company.

- □ Ensure you have a disaster recovery plan in place in case of an unexpected event with the option of rolling back the model, if necessary. The plan should provide:

- □ mechanisms to minimise harm to stakeholders and your company, and
- □ communication channels to facilitate effective interaction amongst stakeholders.





RISK MITIGATION GUIDE

(FOR AI APPLICATION)

RISK MITIGATION GUIDE (FOR AI APPLICATION)

STAGE 1 : Conception

Risk A: Application not being fit for purpose
(i.e., not responsive to the context in which it is sought to be used)

□ Define and document the intended purpose or use of the application you are conceiving for development and ensure that the application is fit for the intended purpose or use; note that this is notwithstanding that the underlying model(s) may have to undergo multiple iterations before the application is deemed fit for purpose.

- □ Identify and describe the stakeholders who are intended to use the application and those who are likely to be directly impacted by its functioning; identify the potential benefits and harms for each stakeholder.

Further Guidance: To learn more about how to conduct AI impact assessment, refer to Microsoft's Responsible AI Impact Assessment Template.

- □ Describe the geographic area(a) where the application is intended to be deployed and take measures to ensure it is responsive to the geographical considerations of language, culture, laws and regulations, etc.

Further Guidance: To learn more about how developers can ensure participation of relevant stakeholders and integrate their feedback throughout the AI lifecycle, refer to Centre for Responsible AI, IIT Madras and Vidhi Centre for Legal Policy's paper on Participatory AI Approaches in AI Development and Governance Case Studies and Partnership on AI's Draft Guidelines for Participatory and Inclusive AI.

- □ Describe the data requirements for application development and ensure that the required data is available in desired quality, quantity, and format; identify any risks associated with sourcing any specific dataset(s) for application development.

- □ Identify suitable model(s) for your application based on your examination of available information on the model(s)' intended purpose(s) or use(s), training datasets and parameters, limitations, and evaluation results.

- □ Ensure you have legitimate access to the required model(s) or necessary tools and resources to customise or fine-tune it for developing the application.

    - □ Define the performance metrics and error types; and prepare an evaluation plan for each of the performance metrics and error types. Note that this is notwithstanding that the performance metrics and error types may have to be revised at later stages of the application lifecyle to meet contextual requirements.

Further Guidance: To learn more about assessing if AI systems are fit for purpose, refer to Microsoft's Responsible AI Standard v2 (see *Goal A3: Fit for Purpose*).

Risk B: Unaccountable development and use of the application

- □ Ensure clarity on the roles and responsibilities of

    - □ external entities that provide services or resources to support the development of the application, (e.g., data providers, cloud service providers, third-party software libraries, etc.), and

    - □ end users or stakeholders who interact with the application, AI education or training provider, marketing partners etc.

  through clearly defined SOPs, valid contracts and licensing agreements.

- □ Ensure clarity on human oversight and control responsibilities from conception to deployment of the application.

STAGE 2 : Collection, processing and usage of data

Risk A: Violation of applicable data privacy regulations

- □ Consult appropriate experts within or outside your company to determine whether any personal data that you intend on using to develop the application qualifies as "publicly available" under the Digital Personal Data Protection Act, 2023. If yes, note that the Digital Personal Data Protection Act, 2023 would not apply to such data. However, verify and document the source of such data.

- □ Consult appropriate experts within or outside your company to determine whether your intended use of any personal data to develop the application would qualify as "legitimate use" under the Digital Personal Data Protection Act, 2023. If yes, note that the consent requirements under the Digital Personal Data Protection Act, 2023 would not apply to such data. However, document this data for review by internal risk assessors and/or external auditors.

- □ If any personal data you intend to use for application development does not qualify as "publicly available" under the Digital Personal Data Protection Act, 2023, and/or if your intended use of such data to develop the application does not meet the criteria for "legitimate use" under the Digital Personal Data Protection Act, 2023, ensure you have the consent from the concerned data principal for using such data for the intended purpose or use of the application.

- □ The notice for obtaining consent from the data principal should include: particulars of the kind of data being collected, purpose of data collection, manner in which data is being collected, manner of consent withdrawal, manner for usage of collected data, procedure for grievance redressal and manner in which complaint can be made to the data protection board.

- □ Check the accuracy and completeness of such data, especially if the application under development will be used to make a decision that would affect the data principal.

- □ If you later decide to use such data for another purpose (such as developing another application), ensure you have the consent to do so from the concerned data principal.

- □ Determine when it is reasonable to conclude that the specified purpose for which such data was processed is no longer being served. If the specified purpose is no longer being served and if data retention is not necessary to ensure compliance with any law, ensure that the data is erased. If data erasure is technically infeasible, register your constraints for internal risk assessment and/or external audits. Consult appropriate experts within or outside your company.

- □ Comply with any requests to correct, complete or update such data. If compliance is technically infeasible, register your constraints for internal risk assessment and/or external audits.

- □ If the data principal concerned withdraws her consent for the processing of her data:

  - □ determine if processing of such data is required to ensure compliance with any law. If not, cease processing of the data. If it is technically infeasible to do so, register your constraints for internal risk assessment and/or external audits.

  - □ determine if retention of such data is necessary to ensure compliance with any law. If not, ensure erasure of the data. If it is technically infeasible to do so, register your constraints for internal risk assessment and/or external audits.

Consult appropriate experts within or outside your company.

- □ Check if your company has put a grievance redressal mechanism in place to address grievances filed by data principal.

- □ If your company has engaged a data processor to process personal data for application development, confirm the existence of a valid contract defining the terms and conditions of such data processing. Ensure that the data contracts are updated to reflect evolving data practices. Consult appropriate experts within or outside your company.

- □ Deploy safeguards to prevent the leakage of personal data and employ privacy-preserving techniques in data collection and use for application development (e.g., data minimisation, anonymisation, pseudonymisation, abstraction, segregation etc.)

- □ Ensure adherence to data protection regulations in addition to the Digital Personal Data Protection Act, 2023 that may apply to your application development and use (e.g., handling of sensitive personal data in application development for a high-risk context). Consult appropriate experts within and outside your company.

- □ Ensure compliance with applicable data protection regulations during application development is clearly disclosed in application documentation, user-facing policies, procurement documents, etc.

- □ Ensure the application has a privacy policy which complies with applicable laws, such as the Digital Personal Data Protection Act, 2023. Consult appropriate experts within or outside your company.

Further Guidance: To learn how to assess and mitigate data protection-related risks throughout the AI lifecycle, refer to United Kingdom Information Commissioner's Office's AI and Data Protection Risk Toolkit.

Risk B: Generation of biased, dangerous, or illegal outputs

- □ Ensure adherence to a documented procedure for robust data governance to ensure data quality for application development.
    - □ Ensure the data you are using to develop the application is of acceptable quality and representative of the stakeholders and geographical considerations identified in stage 1.
    - □ Ensure the data you are using to develop the application is devoid of CSAM (Child Sexual Abuse Material), NCII (non-consensual intimate imagery), or information related to development of chemical, biological, radiological, or nuclear (CBRN) weapons or other dangerous materials or agents.

Risk C: Unauthorised use of non-personal and/or proprietary data

- □ Maintain data catalogs.

- □ Before using any datasets in the possession of your company for application development, ensure you have obtained the necessary internal permissions.

- □ If you are using open-source datasets to develop the application, ensure you comply with the terms and conditions that apply.

- □ If you are sourcing data from a third party to develop the application, ensure you have the legitimate authorisation to source it.

- □ Consult appropriate experts within or outside your company to ensure compliance with intellectual property regulations applicable to the types of data you intend to use for model finetuning.

- □ Ensure appropriate training of the persons involved in application development on responsible data collection, processing, and use, in adherence with applicable regulations, including intellectual property rights laws, etc.

Risk D: Data security breach

□ Check internal security measures to guard against potential breaches of data collected and stored by your company for application development.

□ Check with appropriate experts within or outside your company to see that internal security measures qualify as reasonable security safeguards.

□ Check with appropriate experts within or outside your company If applicable regulations mandate notifying relevant authorities and affected stakeholders of data breaches through specific procedures.

STAGE 3 : Designing, development, and testing

Risk A: Unauthorised, unlawful, or irresponsible use of the model procured from third party

□ Check if your purported use of the model is

  □ authorised by such third party,

  □ compliant with the model usage policies (e.g., OpenAI's Usage Policies, Anthropic's Usage Policies, Google's Generative AI Prohibited Use Policy, Sarvam AI's Terms of Use, and Krutrim's Acceptable Use Policy), and applicable laws and regulations, and

  □ internal model access guidelines at your company.

□ Review the checks and safeguards that have been put in place by the model developer to pre-empt unauthorised or unlawful output generation by model users.

Risk B: Unauthorised, or unlawful, or irresponsible use of software code sourced from third party or generated through generative AI solutions

□ If a software code has been sourced from a third party, verify the third party's rights to the code and their authority to grant such rights.

□ If a software code has been procured from a third party, ensure and check if your purported use of the code is

  □ authorised by such third party, and

  □ compliant with applicable laws and regulations.

□ Ensure you have been able to deploy necessary controls against unauthorised or unlawful use of such software code.

- □ If you are using generative AI solutions to generate software code, ensure that you have obtained the necessary internal permissions. Also, apply checks to ensure functionality and hygiene of such code.

Risk C: Compromised security and robustness of the application

- □ Adopt reasonable security safeguards against data breaches and unauthorised access to the application.
- □ Implement measures to validate the application and tackle generation of inaccurate or unreliable outputs by the application due to glitches or approximations, data bias, model bias, etc.
- □ Implement measures to handle sudden peaks in workload that may lead to system failures and substantial loss and damage.
- □ Confirm that the developers of the model you are using have implemented technical measures to ensure that the model is robust and can withstand known adversarial attacks.
- □ Implement technical measures to ensure the application is robust and can withstand known adversarial attacks.
    - □ Deploy procedures to stress-test the application under different scenarios for unintended harms (e.g., unfair treatment of certain stakeholder groups) that might arise from the deployment and use of the application.
    - □ Employ red teaming to identify vulnerabilities in the application from a security and ethical standpoint.
    - □ Ensure you have documented results from the stress-testing and/or red teaming for review by internal risk assessors and/or external auditors.
    - □ Implement measures to maintain the ability to exercise ultimate human control such as circuit breakers, kill switches, or equivalent mechanisms.

Further Guidance: To see how to defend AI systems against adversarial attacks, refer to the Adversarial Robustness Toolbox (ART). To validate the performance of AI systems against international best practices, refer to Singapore Infocomm Media Development Authority's AI Verify framework. To learn about adversarial techniques and mitigation measures, and perform threat assessment and internal red teaming, refer to MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems).

- □ Consult with appropriate experts within or outside your company and ensure that the application follows standards mandated by the sectoral regulator(s).
- □ Before deployment of the application, verify that

- □ all relevant documents in relation to application development are in order, and
- □ unapproved changes are not made in the application.

Risk D: Lack of transparency about the AI application development process, capabilities, limitations, and eventual use among stakeholders

□ Develop documentation to enable public transparency about the application development process, capabilities, limitations, safety and security evaluations, etc. Existing industry best practices for AI documentation include:

- □ Datasheets that document the data collection and use procedures and related safety and security measures. Also document dataset characteristics and limitations of the model training and validation data.
- □ Factsheets that contain information about the application captured throughout its entire lifecycle, incorporating input from multiple actors in the lifecycle, and is typically tailored to meet the specific needs of the target audience.

Further Guidance: To learn more about factsheets, refer to IBM's AI Factsheets 360. To learn about the guiding principles for designing AI documentation, refer to The CLeAR Documentation Framework for AI Transparency.

□ If the application you are developing is intended for content generation based on user prompts, ensure that the model you are using contains appropriate technical mechanisms to label outputs as AI-generated.

Further Guidance: To learn about tools and techniques for AI watermarking, refer to Hugging Face's AI Watermarking 101: Tools and Techniques

Risk E: Unexplainable outputs from the application in context(s) that demand explainability

□ Check with internal and external experts if the outputs generated by the application that you are developing would need to be explainable, either under applicable laws or prevailing ethical mandates. Note that the requirement for model explainability is likely to be most critical in situations where fundamental rights or matters of life and safety are involved.

- □ If the outputs generated by the AI application that you are developing *must* be explainable, deploy measures to ensure model explainability while ensuring that the model explanations are tailored to the end user's technical proficiency, background and other relevant characteristics.
- □ Ensure human fallback, especially for applications intended for deployment in high-risk contexts.

Risk F: Model being uninterpretable in context(s) that demand interpretability

- □ Check with internal and external experts if the model powering the application that you are developing would need to be interpretable, either under applicable laws or prevailing ethical mandates. Note that the requirement for model interpretability is likely to be most critical in situations where fundamental rights or matters of life and safety are involved.

- □ Check with internal and external experts if the model powering the application that you are developing would need to be interpretable, either under applicable laws or prevailing ethical mandates. Note that the requirement for model interpretability is likely to be most critical in situations where fundamental rights or matters of life and safety are involved.

STAGE 4 : Deployment, monitoring and maintenance

Risk A: Uncontrolled deployment of the application

- □ Consider deployment of the application in a phased manner by initially allowing access to a small group of users to enable early detection and mitigation of potential issues before broader deployment to reduce the likelihood of large-scale harms.

- □ If you decide to make the application open source, document the method and terms of its release, including the extent of access to the application's source code, underlying model(s) and algorithms, data used, etc.

- □ Implement procedures for continuous monitoring of the application to check for data and model drifts, ensure compliance with applicable ethical and legal requirements or standards for model performance and safety.

  - □ Ensure you have documented results from such monitoring for review by internal risk assessors and external auditors.

- □ Adopt mechanisms for incident reporting and collection of user feedback on the safety and performance of the application.

- □ Adopt measures to reduce the impact on individual privacy in the event of application malfunction.

- □ Ensure there is an appropriate grievance redressal mechanism to resolve issues with the application after its deployment and that there is an internal team to monitor and respond to such grievances.

- □ Ensure an appropriate audit mechanism is created for monitoring and maintenance of the application.

- □ Ensure that the end users of the application understand that they are interacting with an AI-based machine.

□ Assess and document the feasibility of rolling back the application, if necessary, as per established procedures and protocols at your company.

Further Guidance: To learn about best practices for deploying secure AI systems in high-risk environments, refer to United States National Security Agency's Artificial Intelligence Security Center (NSA AISC) guidance on Deploying AI Systems Securely.

Risk B: Unintended or malicious use of the application by third parties leading to adverse impacts

□ Ensure you have been able to deploy necessary controls against unauthorised or unlawful use of the application that you have developed.

- □ Document and publish information about unsupported uses and known limitations of the application, with guidance on its safe and responsible use.
- □ Check if the terms of service, and usage policies for the application incorporate your concerns about its unauthorised or unlawful use.

Consult appropriate experts within or outside your company.

□ Discuss adoption of mechanism that enables notification of unintended or malicious use or results within the company.

□ Ensure you have a disaster recovery plan in place in case of an unexpected event with the option of rolling back the application, if necessary. The plan should provide:

- □ mechanisms to minimise harm to stakeholders and your company, and
- □ communication channels to facilitate effective interaction amongst stakeholders.

Further Guidance: To learn about best practices for building LLM-based products, refer to Meta's Responsible Use Guide and for developing and using machine learning systems responsibly refer to nasscom's Responsible AI Architect's Guide and AWS' Responsible Use of Machine Learning (v. 1.2).

REFERENCES

- AI Verify Foundation. (2024, May 30). Model AI Governance Framework for Generative AI: Fostering a Trusted Ecosystem. Infocomm Media Development Authority. https://aiverifyfoundation.sg/wp-content/uploads/2024/05/Model-AI-Governance-Framework-for-Generative-AI-May-2024-1-1.pdf

- Amazon Web Services. (2023, June 21). Responsible use of machine learning version 1.2. https://d1.awsstatic.com/responsible-machine-learning/AWS_Responsible_Use_of_ML_Whitepaper_1.2.pdf

- Anthropic. (2024, June 6). Usage Policies. https://www.anthropic.com/legal/aup

- Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), 6:1-6:13. https://doi.org/10.1147/JRD.2019.2942288

- Artificial Intelligence Security Center. (2024, April 15). Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems (U/OO/143395-24). National Security Agency, Cybersecurity and Infrastructure Security Agency, the Federal Bureau of Investigation, the Australian Signals Directorate's Australian Cyber Security Centre, the Canadian Centre for Cyber Security, the New Zealand National Cyber Security Centre, and United Kingdom National Cyber Security Centre. https://www.ic3.gov/CSA/2024/240415.pdf

- Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., & Zhang, Y. (2019). One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques (arXiv:1909.03012). arXiv. http://arxiv.org/abs/1909.03012

- Bhatt, M., Chennabasappa, S., Li, Y., Nikolaidis, C., Song, D., Wan, S., Ahmad, F., Aschermann, C., Chen, Y., Kapil, D., Molnar, D., Whitman, S., & Saxe, J. (2024). CyberSecEval 2: A Wide-Ranging Cybersecurity Evaluation Suite for Large Language Models (arXiv:2404.13161). Meta. http://arxiv.org/abs/2404.13161

- Chmielinski et al. (2024, May 21). The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners and Context for Policymakers [Discussion Paper]. Harvard Kennedy School Shorenstien Center on Media, Politics and Public Policy. https://shorensteincenter.org/clear-documentation-framework-ai-transparency-recommendations-practitioners-context-policymakers/

- DHR-ICMR Artificial Intelligence Cell. (2023). Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare. Indian Council of Medical Research. https://www.icmr.gov.in/ethical-guidelines-for-application-of-artificial-intelligence-in-biomedical-research-and-healthcare

- Digital Personal Data Protection Act, No.22 of 2023. https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf

REFERENCES

- Github.(n.d). Adversarial Robustness Toolbox.
  https://github.com/Trusted-AI/adversarial-robustness-toolbox

- Google. (2024, March 14). Generative AI Prohibited Use Policy.
  https://policies.google.com/terms/generative-ai/use-policy

- Google Cloud. (n.d.). Introduction to Vertex Explainable AI.
  https://cloud.google.com/vertex-ai/docs/explainable-ai/overview

- Hugging Face. (n.d.). Model cards.
  https://huggingface.co/docs/hub/en/model-cards

- IBM Research. (2019, August 8). Introducing AI Explainability 360.
  https://research.ibm.com/blog/ai-explainability-360

- Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., & Khabsa, M. (2023). Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. Meta. arXiv.
  http://arxiv.org/abs/2312.06674

- Information Commissioner's Office. (2022, May 24). AI and Data Protection Risk Toolkit.
  https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/

- Krutrim. (n.d.). Terms and Conditions.
  https://www.olakrutrim.com/terms-and-conditions#:~:text=You%20must%20only%20use%20our,%E2%80%9D%20includes%2C%20without%20limitation%2C%20the

- Luccioni, S., Jernite, Y., Thomas, D., Witko, Ozoani, E., Fukano, J., Srivastav, V., Tousignant, B., Moctchell, M. (2024, February 26). Hugging Face, AI Watermarking 101: Tools and Techniques. Hugging Face.
  https://huggingface.co/blog/watermarking

- Meta Llama. (2024, April). Responsible Use Guide: Resources and best practices for responsible development of products built with large language models.
  https://ai.meta.com/static-resource/responsible-use-guide/

- Meta Llama. (n.d.). Meta Llama 3 Acceptable Use Policy.
  https://www.llama.com/llama3/use-policy/

- Microsoft. (2022, June). Responsible AI Impact Assessment Template.
  https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf

- Microsoft. (2022, June). Responsible AI Standard v2.
  https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

- MITRE ATLAS. (n.d.). ATLAS Matrix.
  https://atlas.mitre.org/matrices/ATLAS

REFERENCES

- nasscom Responsible AI Resource Kit. (2023, October 11). Risk identification and assessment tool. https://indiaai.gov.in/responsible-ai/pdf/risk-identification-and-assessment-tool.pdf.

- NITI Aayog. (2018, June). National Strategy for Artificial Intelligence. https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf.

- NITI Aayog. (2021, February). Responsible AI, Approach Document for India, Part 1 – Principles for Responsible AI. https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf

- NITI Aayog. (2021, August). Responsible AI: Approach Document for India: Part 2 - Operationalizing Principles for Responsible AI. https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf

- NITI Aayog. (2022, November). Adopting the Framework: A Use Case Approach on Facial Recognition Technology. https://www.niti.gov.in/sites/default/files/2022-11/Ai_for_All_2022_02112022_0.pdf

- nasscom. (2023, June). Responsible AI Guidelines for Generative AI https://www.nasscom.in/ai/img/GenAI-Guidelines-June2023.pdf

- nasscom. (2022, October 11). Responsible AI Architect's Guide. https://indiaai.gov.in/responsible-ai/pdf/architect-guide.pdf

- OpenAI. (2024, January 10). Usage Policies. https://openai.com/en-GB/policies/usage-policies/

- Partnership on Artificial Intelligence. (n.d). PAI's Guidance for Safe Foundation Model Deployment: A Framework for Collective Action. https://partnershiponai.org/modeldeployment/#:~:text=A%20-Framework%20for%20Collective%20Action,to%20evolving%20capabilities%20and%20uses

- Partnership on Artificial Intelligence (2024, September 17). [Draft] Guidelines for Participatory and Inclusive AI. https://partnershiponai.org/stakeholder-engagement-for-responsible-ai-introducing-pais-guidelines-for-participatory-and-inclusive-ai/

- Sanyal, S., Sharma, P, Dudani, C. (2024, January 1). A Complex Adaptive System Framework to Regulate Artificial Intelligence (Report No. EAC PM/WP/26/2024). Economic Advisory Council to the Prime Minister. https://eacpm.gov.in/wp-content/uploads/2024/01/EACPM_AI_WP-1.pdf

- Parthasarathy, A., Ravindran, B., Krishnan, G., Jauhar, A., Phalnikar, A. (2024, April 17). Participatory AI Approaches in AI Development and Governance Case Studies. Centre for Responsible AI, Robert Bosch Centre for Data Science and AI, IIT Madras & Vidhi Centre for Legal Policy. https://vidhilegalpolicy.in/research/participatory-ai-approaches-in-ai-development-and-governance/

REFERENCES

- Sarvam AI. (2024, August 21). Terms of Use.
  https://www.sarvam.ai/terms-of-use

- Securities and Exchange Board of India. (2019, January 4). Reporting for Artificial Intelligence (AI) and Machine Learning (ML) applications and systems offered and used by market intermediaries, Circular No.: SEBI/HO/MIRSD/DOS2/CIR/P/2019/10.
  https://www.sebi.gov.in/legal/circulars/jan-2019/reporting-for-artificial-intelligence-ai-and-machine-learning-ml-applications-and-systems-offered-and-used-by-market-intermediaries_41546.html

- Securities and Exchange Board of India. (2019, May 09). Reporting for Artificial Intelligence (AI) and Machine Learning (ML) applications and systems offered and used by Mutual Funds, Circular No.: SEBI/HO/IMD/DF5/CIR/P/2019/63
  https://www.sebi.gov.in/legal/circulars/may-2019/reporting-for-artificial-intelligence-ai-and-machine-learning-ml-applications-and-systems-offered-and-used-by-mutual-funds_42932.html

- Solaiman, I. (2023, February 5). The Gradient of Generative AI Release: Methods and Considerations (arXiv:2302.04844). arXiv. http://arxiv.org/abs/2302.04844

- Telecommunication Engineering Centre. (2023, July 7). Fairness Assessment and Rating of Artificial Intelligence Systems. Department of Communications, Ministry of Communications.
  https://tec.gov.in/pdf/SDs/TEC%20Standard%20for%20fairness%20assessment%20and%20rating%20of%20AI%20systems%20Final%20v5%202023_07_04.pdf

- Telecom Regulatory Authority of India. (2023, July 20). Recommendations on Leveraging Artificial Intelligence and Big Data in Telecommunication Sector.
  https://www.trai.gov.in/sites/default/files/Recommendation_20072023_0.pdf

- The Information Technology Act, 2000 (No. 21 of 2000).
  https://www.meity.gov.in/writereaddata/files/act2000_0.pdf

- The Information Technology (Security Procedure) Rules, 2004.
  https://i4c.mha.gov.in/theme/resources/actRule/Information%20Technology%20(security%20Procedure)%20Amendment%20Rules,%202004.pdf

- The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.
  https://www.meity.gov.in/writereaddata/files/Information%20Technology%20%28Intermediary%20Guidelines%20and%20Digital%20Media%20Ethics%20Code%29%20Rules%2C%202021%20%28updated%2006.04.2023%29-.pdf

- U.S. Cybersecurity and Infrastructure Security Agency & and the UK National Cyber Security Centre. (2023, November 26). Guidelines for secure AI system development.
  https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development/guidelines

- Zhu, K., Zhao, Q., Chen, H., Wang, J., & Xie, X. (2024). PromptBench: A Unified Library for Evaluation of Large Language Models (arXiv:2312.07910). arXiv.
  http://arxiv.org/abs/2312.07910

nasscom

ANAND & ANAND

इलेक्ट्रॉनिकी एवं
सूचना प्रौद्योगिकी मंत्रालय
MINISTRY OF
ELECTRONICS AND
INFORMATION TECHNOLOGY

सत्यमेव जयते

INDIAai
A MEITY INITIATIVE

Digital India
Power To Empower