



PACTA

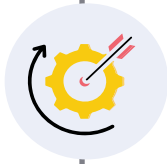
SOCIAL | IMPACT | LEGAL



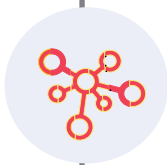
Mitigating Legal & Ethical Risks

for Not-for-Profit Organizations in Use of AI - A Primer

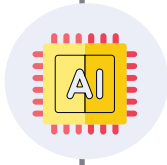
Technical Terms Explained



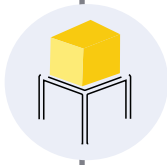
Accuracy, Precision and Recall – Performance metrics for a classification algorithm. Accuracy is the measure of correct predictions; Precision is the measure of correct predictions of the positive class, whereas recall measures the ability of the algorithm to identify true positives from all the positives.



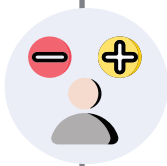
Algorithmic bias – The replication of biases in training data as systematic and repeating errors resulting in outcomes that are different from the intended function of the algorithm, such as privileging one arbitrary group of users over others.



Artificial Intelligence – A broad set of technologies that enable a computing device to replicate human intelligence or to accomplish tasks without being explicitly programmed.



Foundational model – Models trained on large sets of generic data that can be used for various applications with minimal modification. These are used in developing generative AI applications.



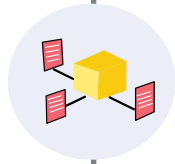
False positives and False negatives – These are two types of errors in the output of classification algorithms where a false positive is an incorrect classification of the presence of a condition, whereas a false negative is an inaccurate classification of the absence of a condition.



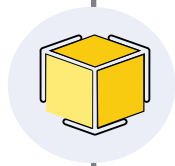
Generative AI – AI techniques for automated creation of synthesised content, including text, code, voice, and images, based on user requirements (otherwise known as prompts).



Generative Pre-trained Transformers (GPT) – is a type of Large Language Model (LLM) developed by OpenAI. At the same time, Chat GPT is OpenAI's generative AI chatbot for generating text-based resources like summaries, research briefs, etc.



Large language model (LLM) – This foundational model was developed to predict word sequences or the subsequent word in a sentence and generate natural language content.



Model – Synonym to a computer program, a model is the output of developing, training, and continually improving the AI or Machine Learning (ML) techniques that accomplish a broad goal with acceptable reliability.



Machine learning – A combination of mathematical and statistical techniques that aid computers in recognising patterns in existing data and making predictions on future data.



Natural language processing (NLP) – A set of algorithms and techniques that use neural networks (a form of ML that mimics the biological structure of brains) to understand and process human language.



Regression – A commonly used form of statistical analysis technique for estimating the relationship between or explaining the variation of a dependent variable based on one or more independent variables.



Software – A set of instructions or algorithms for a computing device to perform specific tasks.



Training and testing data – Training data is the dataset used to teach the machine learning algorithm to recognise patterns. In contrast, testing data is a dataset different from the training data. It is used to test the performance of the algorithm and make adjustment.

Content

Introduction	8
Mapping Popular AI Technologies & Utility for Civil Society Organizations	10
AI Automation	12
Predictive AI	15
Natural Language Processing	21
AI Chatbot	25
Other Generative AI	30
Actionable Recommendations to Mitigate Risks	33
Annexure	37



Introduction

Role of AI in Public Service

Artificial intelligence (AI) is a rapidly growing field of computer science that aims to create intelligent machines that can perform tasks that typically require human intelligence, such as visual perception, speech recognition, decision-making, and language translation.

AI as a concept and academic discipline has been around since the mid-1950s. It caught the public imagination in 1997 when IBM's Deep Blue supercomputer defeated the then-world chess champion. Its rise was primarily driven by advancements and falling costs of computing power and storage capacity, paving the way for the application of machine learning algorithms on big data sets. The launch of ChatGPT into the public domain, followed by the competitive release of other Generative AI models, led to massive adoption and sparked widespread interest in the field, making AI technology mainly accessible to those with an internet connection.

AI technologies use machine learning algorithms and deep neural networks to analyse vast data and make predictions or decisions. AI is transforming various industries, including healthcare, finance, transportation, manufacturing, etc, by enabling automation, optimisation, and personalisation of processes and services. As AI continues gaining prominence in various sectors, countries must develop regulatory responses and policies to ensure responsible development, deployment, and use of AI technologies.

Some AI Basics

Based on its capabilities and applications, AI can be classified into Traditional AI and Generative AI.

Traditional AI

Traditional AI encompasses algorithms and applications that perform a specific task given a set of rules while learning continually and developing new ways of performing those tasks. Predictive analytics, facial recognition, and task-based robots are some forms of Traditional AI.

Generative AI

Generative AI on the other hand, is designed not just to learn from provided data but also to create new data in the form of curated content. ChatGPT, DALL-E, Bing Copilot and Google Gemini are popular generative AI tools. Generative AI mainly has four types of applications: content creation, chatbots, customer services and text summarisation.

Text to image creation: DALL-E, Midjourney

AI pair programmer: Github Copilot

AI writing assistant: Scribe

¹<https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem>

²<https://aiforgood.itu.int/covid-19-how-korea-is-using-innovative-technology-and-ai-to-flatten-the-curve/>

³<https://www.forushealth.com/3nethra-classic.html>

⁴<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9613681/>

⁵<https://economictimes.indiatimes.com/news/economy/policy/rbi-to-extensively-use-ai-mi-driven-tools-for-data-analysis/articleshow/96627679.cms?from=mdr>

⁶<https://indiaai.gov.in/article/do-trains-use-ai-where-is-it-used-in-irctc>

⁷<https://indiaai.gov.in/ministries/ministry-of-defence>

⁸<https://indiaai.gov.in/missions/smart-cities-mission>



What to Expect in this Primer?

AI and Generative AI applications have begun to creep into the public service delivery sector. The Japanese government announced its willingness to adopt ChatGPT¹ if privacy and cybersecurity concerns were solved. With the aid of AI, South Korea was able to develop a COVID-19 testing kit in just three weeks when ordinarily it would have taken two to three months.²

India, with its rapidly growing technology industry and large population, is no exception. The country has recognised the potential benefits of AI and has taken several steps towards AI adoption. AI is used extensively across many sectors in India. In healthcare, AI is being used to diagnose diabetic retinopathy³ and develop an imaging data bank for cancer.⁴ In finance, the RBI is using AI for its data analysis.⁵ AI is used in the Railways to ensure the safety of trains and offer better services.⁶ AI is also augmenting military intelligence⁷ and building Smart Cities.⁸

AI presents immense potential to enhance productivity and improve the efficiency of workers and organisations, thereby reducing costs and freeing up resources for high-value work. In this primer, non-profit organisations can get a sneak peek into how other civil society organisations are using AI. Additionally, being an evolving technology, the use of AI in internal operations or programmatic interventions also carries certain legal and ethical blind spots. This primer is intended to guide NGOs to locate and navigate these moral and legal blind spots.



What Not to Expect from this Primer

This primer is not a technical guide to determine what AI technologies can be applied for specific use cases. Still, we have provided an extensive pool of real-life and illustrative use cases where AI has been implemented by NGOs/CSOs, that we hope you will find easy to help you understand about AI - its merits, utilities and risks.

Mapping Popular AI Technologies & Utility for Civil Society Organizations

In this section we provide a typology of AI technologies, their application in the context of public service delivery by NGOs/ civil society initiatives/ government. We have provided real life examples and suggestions on to how the respective AI technology can be used in an NGO's (external) programs or (internal) operations. Additionally, for each AI technology, we highlight some legal and ethical risks associated with its usage along with corresponding mitigation strategies.

However, it is important to recognize these as potential capabilities rather than definitive solutions that AI can provide. NGOs must remember that AI tools are not the only available solution for a problem. NGOs should consider various systemic hindrances and anticipate the potential consequences if these challenges are not thoroughly addressed and thought through.

There have been several frameworks for responsible AI (RAI frameworks) design, and one is provided below.

Responsible AI Principles Recognized In India⁹



Inclusivity & Non-Discrimination

AI systems must be fair and inclusive, and should not foster prejudices, discrimination or preference for an individual, a community or a group based on their sensitive attributes (e.g., race, gender, ethnicity).



Security

AI system should be robust and secured against adversarial attacks and malicious use. Identifying and mitigating system vulnerabilities is critical.



Accountability

Organisational structures and policies should be created to clarify who is accountable for the outcomes of AI systems. Human supervisory control of AI systems is recommended.



Reliability & Safety

AI systems must produce consistent and reliable outputs in all scenarios. Appropriate grievance redressal mechanisms should be implemented to address adverse impact cases.



Transparency

AI systems should be transparent about their development, processes, capabilities, and limitations to the greatest extent possible.



Transparency

AI systems should be designed and operated such that they align with human values. AI should promote positive human values for the progress of humanity as a whole.



Privacy

AI systems should respect user privacy. Users' right to know what data is collected, why it is collected and who has access to it should be protected. AI systems should not use the data for purposes other than what is stated.



Explainability

AI systems should be explainable to users significantly impacted by their decisions. Explanations must be provided free of cost in non-technical, intuitive language.



Compliance

Throughout their lifecycle, AI systems should comply with all applicable laws, statutory standards, rules, and regulations – Organizations should be watchful of the evolving AI regulatory landscape and ensure

⁹<https://indiaai.gov.in/responsible-ai/homepage>

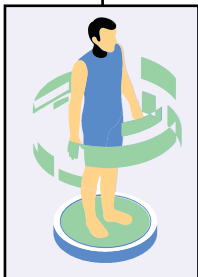
1 AI Automation

Concept

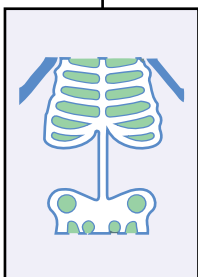
AI assisted automation otherwise known as Intelligent Automation (IA) is a system designed to learn, perform and continually improve the execution of routine, repetitive and predictable tasks.



Example 1: Qure.ai



Qure.ai has employed deep learning techniques to diagnose disease and create automated diagnostic reports from **CT Scans, X-Rays and MRI**.¹⁰



Qure.ai uses artificial intelligence (AI) technology to assist with **X-ray screening** to find pathways to assist clinicians in expediting **tuberculosis diagnosis**, thereby diminishing the duration and complexity of both investigation and treatment of tuberculosis.



Previously the diagnosis required **multiple investigations and examination by radiologists and physicians**. The new technology has the effect of freeing up their capacity to diagnose more patients and prioritise ones displaying symptoms.

¹⁰<https://indiaai.gov.in/startup/qure-ai>

Example 2: DigiYatra

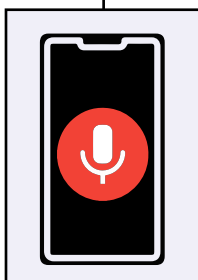


DigiYatra is a biometric facial recognition-based identification system that facilitates paperless and seamless travel for passengers at airports¹¹ by obviating the need to verify their identity at multiple checkpoints.



Passengers are required to create a **digital ID by providing their personal details and valid ID proof**. Once the photo is verified it can be used as a digital identity throughout the journey.

Example 3: Wadhwani AI



Wadhwani AI developed a predictive system to detect presence of COVID-19 infection through an **analysis of cough** sounds that alleviated the problem of limited testing capacity while also enhancing early detection with its ability to identify asymptomatic patients.¹²

¹¹<https://www.india.gov.in/spotlight/digi-yatra-new-digital-experience-air-travellers>

¹²<https://www.wadhwaniai.org/programs/cough-against-covid/>

How NGOs can use AI Automation

1. Operations:

- 01** Automated access/ entry into highly monitored spaces
- 02** Automated filling of statutory forms
- 03** Automated Removal unparliamentary language (based on pre-trained data) from employee chat channels
- 04** Automated onboarding and offboarding of employees (access to email, internal systems and data, creation of ID card etc.)
- 05** Automated reminders and follow up
- 06** Automated creation of standard compliance or audit reports, personalised reports for donors that require data from multiple sources and document formatting.

2. Programs:

- 01** Diagnose diseases/ plant infestations
- 02** Create follow up plans for patients, students, other program participants
- 03** Send reminders to comply with treatment regimen



TIPS

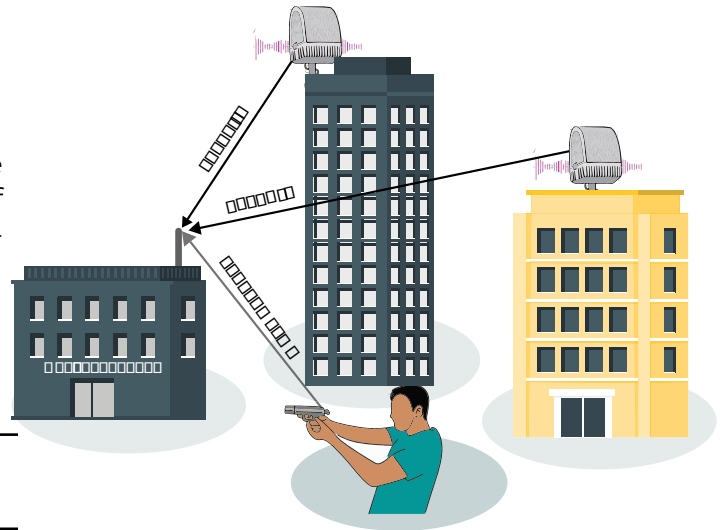
Pacta Pro Tip:

Data related practices – pertaining to consent based collection, use, privacy, security, governance are inextricably linked with AI. Hence while commissioning AI development services or licensing AI tools for use, always be sure to ask and be convinced about the service provider's data practices.

2 Predictive AI

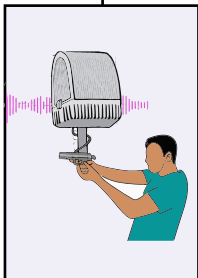
Concept

Predictive AI predicts the outcome of future events based on analysis of past data that aids in the decision-making process.

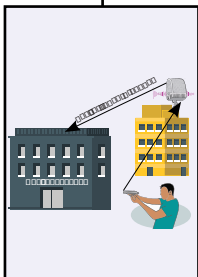


Example 1: Shotspotter¹⁴

Shotspotter is a gunshot detection technology used by law enforcement agencies in the United States and other countries. It aims to:



Detects and locates the place of gunfire incident: Utilises a network of acoustic sensors to pick up loud, impulsive sounds suggestive of gunshots.



Alert law enforcement: Once a potential gunshot is detected, the system analyses the data and sends real-time alerts to authorities with the precise location on a map.

¹⁴<https://www.soundthinking.com>

Example 2: Stanford AI



In collaboration with Stanford University, the **International Rescue Committee** conducted a regression analysis on the last 10 years of **refugee resettlement** across the USA, to create a tool to place refugees in cities that would optimise their overall employment rate. The research team believes that 40% improvement in employment rates is possible if one follows the lessons of the algorithm they have developed.



The Stanford AI employment algorithm is expected to offer massive benefits for refugees searching for a job.¹⁵

Example 3: Akshaya Patra



Akshaya Patra increased efficiency of its **mid-day meal delivery program** by 20% using analytics to optimise logistics for food delivery.¹⁶

Example 4: Save Life Foundation



Save Life Foundation used large scale data from surveys on perception on **road safety to identify and predict accident** risk prone areas.¹⁷

¹⁵ https://verdict-ai.nridigital.com/verdict_ai_summer19/refugees_artificial_intelligence

¹⁶ <https://cio.economictimes.indiatimes.com/news/corporate-news/it-is-the-secret-ingredient-to-akshaya-patras-efficient-operations/58356181>

¹⁷unable to locate

How NGOs can use Predictive AI

1. Operations:

- 01** Select the appropriate candidate/s from a pool of applicants for a job based on pre-selected criteria as best fits for a role
- 02** Use past data to forecast the consumption of raw-material/resources or needs of manpower for similar projects
- 03** Streamlining operations and resource optimisation – Guide deployment of volunteers or staff for field work.

2. Programs:

- 01** Forecast demand for assistance among beneficiaries to improve effectiveness of aid distribution.
- 02** Improve agricultural productivity by providing recommendations on suitable crops, time of sowing, fertiliser requirement etc through soil and crop growth analysis.
- 03** Boost preventive healthcare by predicting likelihood of diseases affecting a person in the future thereby enabling early detection and diagnosis.
- 04** Identify which students may have a higher likelihood of dropping out of school or other programs conducted by the NGO to ensure higher adherence.
- 05** Predict the average time that may be taken by court to pass judgement.
- 06** Predict which areas/ populations are more vulnerable to a particular disease.
- 07** Predict forest fires, flood and natural catastrophes.



Risks and their mitigation:

NGOs are still familiarizing themselves with the risks, potential and use cases of AI. Most AI models are non-transparent, making it difficult for NGOs to assess their reliability. Inadequate testing or insufficient training data can lead to incorrect recommendations, with inbuilt biases affecting beneficiaries.

Some Ways to Mitigate Risks:

1. **Define the AI tool's functionality:** Clear and unambiguously defined algorithms that involve selection, optimization of right metrics and data points will help to reduce unintentional bias.
2. **Human-in-the-Loop to validate AI decisions:** AI systems have the capacity to learn and improve at performing assigned tasks. Using a human to validate the decision of AI will ensure that context appropriate responses are provided to the users.

For example, Google's AI incorrectly flagged a photo as child pornography leading to the wrongful accusation of on a father who was merely sharing a picture of his child's infected groin to a physician for diagnosis.¹⁸ Hence, regular manual reviews (with a human in the loop) are necessary to correct and provide feedback to the system to rectify errors and complete subtasks in ways that are more closely aligned with how a system is expected to work.



3. **Audit:** NGOs must conduct periodic audits of the AI's outputs to objectively check for biases, consistency, and accuracy of responses.
4. **Train AI Tools Based on Indigenous Models:** Over reliance on Western large language models or datasets can lead to outcomes that are skewed toward a

Western context. Indian AI tools and models may help contextualize the responses more effectively.

Specifically, some of the questions you should ask are as follows:

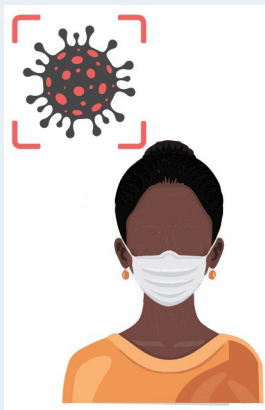
- 01 On what parameters/ data has the AI been trained and whether those parameters are relevant for you and your organisation?
- 02 Whether the problem of false negatives or false positives is likely to be encountered?
- 03 Who would be liable in case the AI makes inaccurate predictions?
- 04 How does the vendor use, process and store the data you are feeding into their system?
- 05 Whether it would be possible to not use user data in ways that you don't want the vendor to. Note that very often models of consent are opt-out models and not opt-in models. One such example is below:

Our Use of content: We may use Content to provide, maintain, develop, and improve our services, comply with applicable law, enforce our teams and policies, and keep our Services safe.

Opt Out: If you do not want us to use your Content to train our models, you can opt out by following the instruction is this Help Center article. Please note that in some cases this may limit the ability of our Services to better address your specific use case.

¹⁸<https://www.nytimes.com/2022/08/21/technology/google-surveillance-toddler-photo.html>

Predictive AI in Action - Prediction of Survival after Covid-19



A system that was trained to predict the chance of survival of adults infected with the Covid-19 virus, if trained predominantly on data with a population from a certain ethnicity, it may predict poorer outcomes for persons of other ethnicities. On these lines, African American people were predicted to have a poorer chance of survival (a higher morbidity), than Caucasian populations and were less preferred for receiving life-saving medical attention in resource constrained settings during the pandemic.²¹ This resulted in an unintentional lack of access to lifesaving treatment to large number of such people.

To address this, we recommend:

- a) Regular review of algorithm decisions to determine its accuracy and educating decision makers about its limitations is necessary to mitigate this problem.
- b) Familiarise with training data, liability and privacy if third party AI tools are used.
- c) In case third party tools or products are used for AI automation, it is imperative to have a lawyer review and understand their terms of service.



Pacta Pro Tip:

Devising a response strategy for Communication, rectification, and prevention when the usage of AI system has caused unintentional harm.

TIPS

Step 1: Issuing an apology and a robust PR management and outreach plan.

Step 2: Short-term fixes like individual corrections or feature discontinuation.¹⁹ For e.g. the Google Photos algorithm incorrectly labelled people of colour as “**Gorillas**”. Hence an immediate fix was prohibiting the algorithm from labelling anything as a gorilla.²⁰

Step 3: Preventing repeat of same or similar incidents - training of staff in usage of AI, improving diversity of training data, tweaking features that are more relevant and better labelling of data.

¹⁹<https://www.trustible.ai/post/what-to-do-when-ai-goes-wrong>

²⁰<https://www.pachyderm.com/blog/when-ai-goes-wrong-and-how-to-fix-it-fast/>

²¹<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7762908/>

Pacta Case Study – AI In Mental Health

As a part of better mental health initiative in their organisation, our client shared a draft contract with the Mental Health Consultancy for our review.



On a perusal of the scope of work the service was being rendered through a digital platform on which our Client's employees could book appointments and avail sessions with a counsellor/ psychiatrist to address any mental health concerns.



The appointment bookings were to be done through a digital platform that enabled clients to choose a counsellor of choice, maintain their case file, progress etc.



AI enabled algorithm would check in with clients asking direct questions to ascertain their mental health status, and flag the same to the Mental Health Consultancy.



Red flags:

- 01** What type of data was the algorithm trained on?
- 02** Was the app going to use the client case data to sharpen and improve its predictions?
- 03** What type of suggestions would the app provide if responses to its check in questions indicated that the user was likely to have a serious or severe mental health condition?
- 04** What type of auto-generated action would the app trigger if the app's algorithm identified an individual of inflicting self harm?

- 05** Who would be liable if a person who was flagged as high risk was not at risk or a person flagged at low risk was actually suffering from a mental health condition that was not detected?
- 06** Whether the app would in-fact be backed by emergency support - should a person feel the need for this? Or whether it would be pre-trained responses that mimicked a human interaction?
- 07** Whether informed consent would be taken from the employees (users) before the AI asked questions, and whether this was an opt-in feature?

Based on clarifications received from the vendor, our client sensitised their user group to know the pros and cons of using the platform. Pacta also recommended that a clause be included in the contract clarifying that their employee data would not be used to train/hone the app.

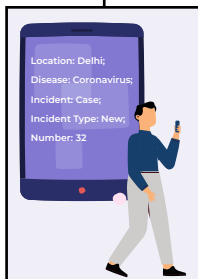
3 Natural Language Processing

Concept

Provides the ability for a computer to extract meaning and insights out of human language content. For example, spell checkers, translators, voice assistants, spam filters, and autocorrect are all NLP applications. The advent of LLM's has improved the ability of the computer to process and understand varied types of inputs as if a human would.

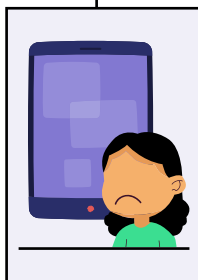


Example 1: Wadhwani AI



AI powered event-based disease outbreak surveillance system developed by Wadhwani AI. This system intelligently screens news articles to extract information on disease outbreaks and their location to notify relevant public health authorities.²² NLP is used to identify relevant articles and extraction of disease names, number of people affected, and location.

Example 2: Save the Children



Save the Children developed an NLP model by analysing various data sources to identify instances and severity of online violence against children.²³

²²<https://www.wadhwaniai.org/2022/12/using-ai-to-automate-the-early-detection-of-disease-outbreaks-in-india/>

²³<https://www.omdena.com/projects/children-violence>

Example 3: Online Mental Health



Crisis Text Line, an NGO providing online mental health services used NLP to analyse over 40 million text messages identifying signs of anxiety, depression and suicide enabling them to prioritise connecting with patients.²⁴

How NGOs can use Natural Language Processing

1. Operations:

01

Text to speech and vice versa conversion to record field observations, to gather data for impact evaluation and translate from one language to another making digital offerings more accessible.

02

Aid qualitative research by identifying common themes from transcripts of interviews or responses of research subjects

03

Identify and categorise named entities in text, such as people, places, and organisations, to sift through a large repository of text data

04

Synthesise an extensive amount of regulations/ executive orders from statutory authorities to identify specific compliances that they must be aware of.

2. Programs:

01

Quickly analyse a large number of feedback responses from users to gain insights into broad sentiment whether positive or negative and key themes originating from them. This can help gauge program effectiveness.

02

Synthesise an extensive amount of quasi-judicial/ judicial orders/ orders to identify common themes/ narratives - this insight can then be used to develop suitable programmatic interventions

03

Conduct subjective evaluation of student's descriptive assessments to identify areas of learning deficiencies in languages like grammar, sentence construction etc.

²⁴ <https://healthitanalytics.com/news/natural-language-processing-helps-shed-light-on-youth-mental-health-crisis>



Risks and their mitigation:

The efficiency of LLMs depends on the training data — for a LLM to understand multiple dialects of Hindi, it should have been trained to do so. Often, diversity in the training dataset is a challenge since this entails higher costs, data contribution, and diversity in the training team.

For other systems that involve the analysis of large amounts of text, the systems struggle to understand context, ambiguity in human language, as well as instances of irony and sarcasm, which can affect the output. Additionally, the same word may have different meanings in different contexts, thus requiring human verification to validate the accuracy of the LLM output.



Pacta Pro Tip:

AI is quite likely to be useful especially in reducing human effort, when highly structured tasks are to be performed repetitively.

Pacta Case Study

Pacta examines access to justice for persons with disabilities. Each year, orders issued by the Chief Commissioner of Persons with Disabilities are reviewed to identify key trends and patterns. This analysis helps us understand commonly litigated issues faced by persons with disabilities and enables us to initiate appropriate policy responses.²⁵



The CCPD makes about 400-500 orders annually. For the orders passed in 2022 we manually parsed through 400 such orders and coded the key themes manually. This involved heavy and repetitive manual effort.



To repeat the exercise in 2023, we used the Open AI's LLM and Named Entity Recognition feature to sift through 500+ orders passed by the Chief Commissioner of Persons with Disabilities.



To identify the following themes: types of complaints filed by persons with disabilities, profile of the complainant (gender, type of disability, percentage of disability) typical geographies pertaining to the complaint, thematic areas of complaint, dispute resolution times etc.



We wrote a Python code that parses the PDF versions of orders uploaded on the official CCPD website, separates each order in the document, and uses the OpenAI API to analyze each order. The output is then generated in an Excel sheet format, containing the results for the themes listed above.



These were then further analysed to create thematic insights such as Percentage of male/ female complainants, average dispute resolution timeline, spread of disabilities among complainants, theme wise complaints - employment, education, access to digital services etc.

How we Mitigated the Risk:

- 01** Identification of cases that could not be analysed and incorporating error handling mechanisms to review them at a later stage.
- 02** Advising manual checks



Red flags:

- 01** Some orders were either missed by the LLM for not possessing definitive start and end markers or could not be processed since they were not in English language.
- 02** OpenAI's free tier API restrictions also meant that orders exceeding defined character limits could not be processed and no mechanism existed to bundle multiple requests for combined processing.
- 03** Accuracy was mostly high but sometimes nuances in the order of the CCPD were missed. This meant that the algorithm did not summarise exactly as a human would.

²⁵https://pacta.in/CCPD_Analysis_2022_Research_Report_2022.pdf

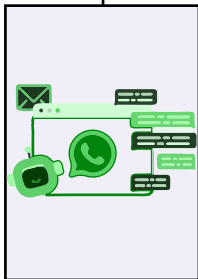
4 AI chatbot

Concept

A chatbot is a generative AI program that is designed to interpret and generate responses to human text or audio input i.e. to simulate a conversation. Generative AI refers to technology that creates new content through learning or analysis of existing data. Typically, chatbots are trained on a knowledge base and would be expected to answer questions based on specifically identified resources.

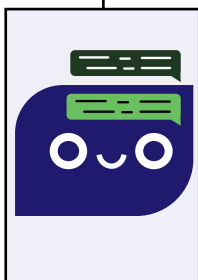


Example 1: Dost Education



Dost Education, an early education-focused non-profit, developed a WhatsApp chatbot to educate parents on activities for young children, enabling them to learn at home. They also developed and shared a home-school curriculum during the COVID-19 pandemic.²⁶ The interactive platform allowed the organization to solicit feedback and incorporate it into their offerings.

Example 2: Roo Chatbot



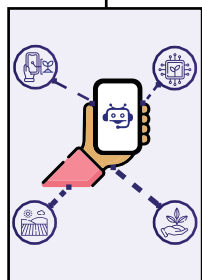
Planned Parenthood developed a chatbot, Roo, which has answered over 8,00,000 sexual and parenting questions in 6 months since its launch enabling informed decision making for young adults.²⁷ rAIInbow, developed a chatbot to create awareness about domestic violence and provides support to those affected.²⁸

²⁶<https://www.wadhwaniai.org/2022/12/using-ai-to-automate-the-early-detection-of-disease-outbreaks-in-india/>

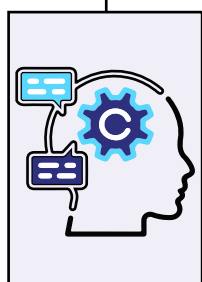
²⁷<https://www.omdena.com/projects/children-violence>

²⁸<https://healthitanalytics.com/news/natural-language-processing-helps-shed-light-on-youth-mental-health-crisis>

Example 3: Kissan GPT



An interactive AI chatbot named 'Kissan GPT' went live on March 15, 2023. It is designed to help India's agriculture sector, allowing farmers to query over an interactive interface from the convenience of their phones.²⁹



It enables a user to ask questions on topics of agriculture such as agricultural cultivation, insect management, and irrigation over voice in nine Indian languages Tamil, Hindi, Bengali, Malayalam, Gujarati, Telegu, Kannada, Marathi, etc. It leverages OpenAI's ChatGPT technology in conjunction with the platform's knowledge store to give the answers to the queries raised.

How NGOs can use AI Chatbots

In their operations, NGOs can use chatbots to assist their own employees or volunteers to get a better understanding of the organisational policies, prior programs.

In their programs, NGOs can use chatbots to assist users by providing personalised information in response to user queries such as procedure to access government services, conditions for entitlements etc. NGOs can also use chatbots as an alternative mode of collecting information as compared to forms – this can be used for registering people for government schemes or to avail NGO's services.

²⁹https://pacta.in/CCPD_Analysis_2022_Research_Report_2022.pdf



Risks and their mitigation:

Inaccuracy

AI enabled chatbots may “hallucinate” (make up factually inaccurate answers), provide inappropriate answers or use inappropriate language owing to misuse or jailbreak techniques employed by some users. The accuracy of information in a chatbot’s responses depends on the quality of information and methods used to train the chatbot. Hence, providing quality training data and maintaining oversight over its sources, regular audits of responses and robust testing before rollout can serve as mitigation measures. In addition, sandboxing or phased release of its capabilities also allows more time to uncover and fix such vulnerabilities.

Copyright breaches

Foundational models are trained on large datasets of publicly available, often copyright-protected information to develop their capabilities. In the USA, The New York Times has filed a lawsuit against OpenAI for violating copyright protections by using its articles published on the web without permission to train OpenAI’s GPT model. This model now competes with NYT as a source of information.³⁰

Cyberattacks

Vulnerability to cyberattacks, such as prompt injection, allows hackers to extract personal information that might have been present in the training data or identify its source. Anonymization or removal of any personally identifiable information from the training data could prevent such data breaches. Another form of attack involves engineering prompts in a manner that generates malicious or unintended responses.

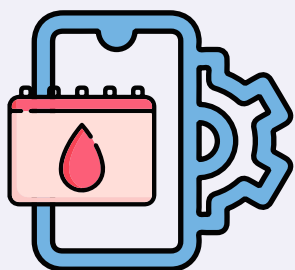
Red teaming is a risk mitigation strategy that involves simulating attacks on the model to identify vulnerabilities and weaknesses before they can be exploited by attackers.³¹ It is the equivalent of ethical hacking in software or IT system for AI models.

³⁰<https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html>

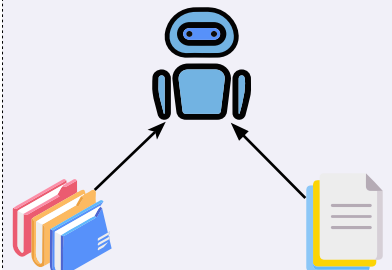
³¹<https://www.nightfall.ai/ai-security-101/ai-model-red-teaming>

Pacta Case Study - A Chatbot For Sexual and Menstrual Health

One of our clients, an NGO working in sexual and reproductive health, reached out to Pacta, requesting our advice on a range of transactions surrounding the launch of a chatbot.



The Chatbot was being developed by a software developing agency, and once ready would answer user questions related to menstrual health. The chatbot was available free of cost to the user.



The Chatbot also provided links to the referred resources to provide the answer. The Chatbot was going to be trained on papers and articles curated by the NGO from a digital platform run by them.



Our role was to identify and mitigate risks across the pipeline of development and deployment of the chatbot.

Risks Identified:

- 01** Does the NGO have the licence/ permissions to use the information resources for the purpose of the chatbot?
- 02** Is the NGO confident about the accuracy of content in these resources?
- 03** If a user took the advice of the chatbot and suffered harm due to that, who would be liable - the NGO or author of the specific resource or developer of the chatbot?
- 04** How could it be ensured that the chatbot did not provide racially inappropriate or other inappropriate answers to the querist?
- 05** Would the chatbot learn from the queries that it received? If yes, how should users be protected?

How we Mitigated the Risk:

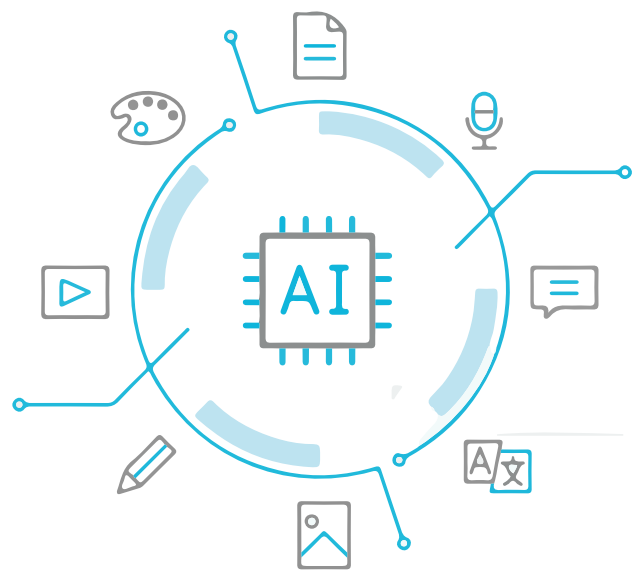
- 01** Since the training and knowledge database had been curated by the NGO for publication on its platform, there was already an understanding that such content could be used for non-commercial purposes, such as awareness building and advocacy, with due credit to the author of the work. Therefore, the NGO could use the resources for the chatbot as long as it explicitly credited the author and ensured the use was non-commercial in nature.
- 02** The NGO had a process of vetting and curating the resources for publication - hence were confident about using the same for knowledge dissemination purposes.
- 03** Additionally, we advised that a suitable disclaimer be placed on the landing page of the chatbot stating that the chatbot was not a diagnostic tool or a medical consultant and if the querist were to have any health concerns, the same should be addressed to the NGO's wellness helpline, where the issue could be resolved.
- 04** We included certain warranties in the development agreement between the NGO and the software development agency, under which the functionalities and features were defined clearly, and damages for breach of warranties were also identified. This included conditions that the Chatbot's knowledge database would be clearly ring-fenced, such that it would not hallucinate.
- 05** We also ensured to define the scope of work very clearly for the developer. A warranty/ feature that was built into the scope of work was the queries of the querist would be recorded and provided to the NGO to help our client understand the nature of queries that users were posing to the chatbot and a weekly analysis of users, demographic, theme and number of queries would be reported to the NGO. However, the technology that was used would be such that the chatbot did not learn from the queries of the querist.

5 Other Generative AI

Concept

Generative AI refers to a subset of artificial intelligence focused on creating or generating new content, data, or artifacts that resemble or are derived from existing data. Unlike traditional AI, which is often tasked with classification, prediction, or optimization, generative AI is primarily concerned with creativity and creation.

Generative AI models are designed to learn the underlying patterns and structures present in the data they are trained on and then generate new instances of data that exhibit similar characteristics. These models can be trained on various types of data, including images, text, audio, video, and more. Generative AI has a wide range of applications across various domains, including:



- 01 Image Generation (Dalle/ Firefly/ Midjourney):** Generating realistic images of objects, scenes, persons, and artwork. This has applications in computer graphics, image editing, and creative design.
- 02 Text Generation (ChatGPT/ Llama/ Anthropic):** Generating human-like text, including stories, poems, articles, and dialogue. This has applications in creative writing, secondary research and content summarisation.
- 03** If a user took the advice of the chatbot and suffered harm due to that, who would be liable - the NGO or author of the specific resource or developer of the chatbot?
- 04** How could it be ensured that the chatbot did not provide racially inappropriate or other inappropriate answers to the querist?

How NGOs can use Generative AI

1. Operations:

01

Employees of NGOs may use Dall-E/ Chat GPT to create illustrations, content, emails, research etc for internal communications, posters, presentations, research etc and visually appealing content with limited resources for programs/ advocacy.

02

It can also enhance or build on developed content such as re-creating a poster in multiple languages.

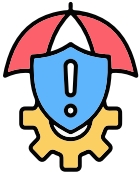
2. Programs:

01

Personalized Education and Skill Development: Tailored learning materials: Generate personalised learning materials and assessments based on their age, individual needs and learning styles, adapting to students' strengths and weaknesses for a more effective learning experience.

02

Precision Healthcare and Personalized Medicine: Support for frontline workers: Generate AI-powered tools supporting community health workers in remote areas, analysing patient data and symptoms to suggest possible treatments and interventions. Tools can also be developed that answer patient's queries and explain the ailments and precaution measures by simplifying medical jargon and in the user's preferred language.



Risks and their mitigation:

The content generated based on provided prompts could contain inaccurate, biased or output that is completely divorced from the intended requirement. Hence, manual reviews of all AI generated content prior to its publication or handover to clients are a necessary step.

Furthermore, development of a risk scale for use of Gen AI-based on its potential of adverse impact would serve as a guide for its usage. The German Data Ethics Commission proposed a 5-tier scale³² (**Figure 1**) to guide AI regulation that can be used as a reference. In case of non-profits, the use of Gen AI for drafting an informal email would be categorised as low risk whereas a report to a donor or regulatory agency would be regarded as high risk on this scale.

- ▶ **Level 1** risk zero or negligible potential harm
- ▶ **Level 2** some potential for harm,
- ▶ **Level 3** regular or significant potential for harm
- ▶ **Level 4** serious potential for harm
- ▶ **Level 5** untenable potential for harm

Figure 1: German Data Ethics committee AI risk scale

By leveraging generative AI technologies, social sector organisations can enhance their capacity to address complex social challenges, deliver more personalised and inclusive services, and empower individuals and communities to thrive. However, it's important to approach the use of AI in the social sector with careful consideration of ethical, privacy, and equity considerations to ensure that AI applications benefit all stakeholders and promote positive social outcomes.



TIPS

Pacta Pro Tip:

Individuals and organisations are advised to use generative AI tools purely as an assistant or productivity enhancement utility and not a complete solution or replacement of human effort in themselves.

³² <https://algorithmwatch.org/en/germanys-data-ethics-commission-releases-75-recommendtions-with-eu-wide-application-in-mind/>

Actionable Recommendations to Mitigate Risks

Identifying Points of Legal & Ethical Exposure Created by AI to Mitigate them

As we saw in the prior sections, there are several legal and ethical blind spots in the path to adoption of AI. This applies to all types of initiatives, be it for-profit or not-for-profit. In this section we identify opportunities at which such legal and ethical risks can be identified, understood and mitigated.

1. Procurement/ Design

The point at which you decide that you need to either commission or design an AI solution to meet your programmatic or operational needs is a moment that offers the opportunity to get it right the first time itself. Do not hesitate to spend time in this phase and consult with several vendors, developers, users etc to not just understand the commercials and functionalities but also to identify legal and ethical determinants of your decisions.

To make it easy for you, here are some questions the Procurement function and the technical team should ask and have answers to at the time of commissioning/ designing AI solutions:

- a.** What type of training data is required for the AI's algorithm?
- b.** What is the purpose/objective of the AI tool and how can we ensure its continued performance in alignment with said purpose/objective without getting side-tracked by short-term incentives and warped data?
- c.** Was such data ethically sourced?
- d.** Can we consider disclosing the architecture of the AI's algorithm to explain the AI's decision making process?
- e.** Would the AI tool collect user data to improve its predictions? If so, is this disclosure made to the user with an option to opt out without any negative repercussions? Additionally, can we discuss the feasibility of using an opt-in approach as the default method for obtaining user consent instead of opt-out?
- f.** What is the likelihood of errors in the outputs? Who would be liable for errors?
- g.** What type of built-in safeguards would the app trigger if the app's algorithm identified a problem?

2. Contracts

While commissioning/licensing AI based solutions, please ensure that your service/ licence agreement provides unambiguous language that minimalises your legal risks. Though contracts may appear unnecessarily long and winding, pay attention to the following clauses and ensure that:

a. Service Providers Liability/Representations: The Service Provider/Licensor's liability is defined and quantified for any wrong, inaccurate, or inappropriate decisions, recommendations, or outputs made by the AI, which result in harm to a user or the commissioning organization.

b. Intellectual Property Rights: Explicit undertaking that the AI tool or solution shall not breach any 3rd party copyright/ intellectual property rights, and an uncapped indemnity in case this clause is breached by the developer/ licensor of the AI.

c. Legal Compliances: Service Provider explicitly undertakes Compliance with all applicable laws including but not limited to laws on personal data privacy, intellectual property rights, other industry specific laws as may be applicable not just on the date of execution of the agreement, but any laws that may be enforceable in the future as well. (See Annexure for generic laws that may apply to AI applications in India)

3. Programs & Practice

When an AI solution is implemented in programs or an NGO's operations, it is important to watch out for any immediate as well as long-term harms that may occur due to the use of the AI tool or solution. Some methods to check for harm include:

a. Sandbox: Repetitive tests and sandboxing of the AI tool/ solution before implementing it to ensure that it performs as it was envisaged to.

b. Capacity Building: Sensitise/train employees to ensure they are familiar with using the tool and inherent risks, by reading the fine print of the license and being aware about its limitations.

c. Third-party audits: Conduct periodic audits on AI decisions/outputs to ensure that the process or decision making:

- i) Is aligned with the program's/ organisation's theory of change;
- ii) Does not unintentionally leave out groups of people who should have ideally benefitted from the program;
- iii) Does not create unintentional but harmful outcomes to persons who were either covered or not covered by the decision output;
- iv) Is factually accurate and is not inflated;
- v) Does not magnify biases in the data-set (Remember the COVID-19 case study).above?)

d. Third-party impact studies: Conduct impact studies to be sure that the AI tool/ solution provides cost-benefit justification and creates the desired impact.

***Third-party** audits and impact studies should be ideally conducted by interdisciplinary experts and affected communities who can give unbiased and objective feedback.

f. Human-in-the-loop: Include a human-in-the-loop wherever possible to verify the accuracy of the output so that AI is a tool that augments but does not replace human decisions.

g. Include disclaimers: Resist the urge to give a human personification to your AI interfaces.

Additionally, you may also consider disclosures and disclaimers in the spirit of full transparency and provenance. Through such disclosures, NGOs can inform users that an AI tool or application was used to make the decision and how their data might be used.

Explainable AI is another aspiration of advocates of Ethical AI. Explainable AI means that users are informed of how the AI tool or application may function, and what factors may influence its decision concerning various stakeholders. Explainability of the tool to the data scientist within the system would look different from that to a regulator or researcher or end-user of the tool. This is in contrast to the typical experience with AI algorithms that leave you wondering why your Instagram page is repetitively showing you how to make bubble tea or drape sarees.

4. Grievance redressal

When AI solutions/tools are used in an NGO's programs or operations, always ensure to also set up a dependable and impartial channel of grievance redressal in case an aggrieved user or other aggrieved persons would like to raise red flags. on AI based decisions. This is essential to help persons build trust in AI and AI based systems, that otherwise have propensity towards unethical and abusive practices.

ANNEXURE

The background of the page is a deep blue color. It features a series of thin, light blue wavy lines that flow from the bottom left towards the top right. Scattered throughout the background are small, faint dots of varying colors, including light blue, purple, and pink, giving it a cosmic or digital feel.

India's Law and Policy Position on AI and Generative AI

Governments across the world have brought in regulations, set up administrative structures or formed committees to study and advise them on the use of AI. While India does not have any specific regulations, generic laws will apply to AI and GenAI matters. These are some major laws that apply to AI and GenAI:

Law	Relevant provision	Application
The Information Technology Act, 2000 ³³	Section 43 A	Compensation for failure to protect data: both developer and the NGO using the AI for its activities can be held liable. For example, An AI solution gets hacked because of inadequate security measures and personal information of users gets compromised. ³⁴
	Section 43 A	Penalty for breach of confidentiality and privacy Obtaining and disclosing information without their consent or knowledge For example, An AI solution listens to user's conversations without their knowledge.

³³https://upload.indiacode.nic.in/showfile?actid=AC_CEN_45_76_00001_200021_1517807324077&type=actfile&file-name=a2000-21.pdf

³⁴[https://upload.indiacode.nic.in/showfile?actid=AC_CEN_45_76_00001_200021_1517807324077&type=rule&file-name=GSR313E_10511\(1\)_0.pdf](https://upload.indiacode.nic.in/showfile?actid=AC_CEN_45_76_00001_200021_1517807324077&type=rule&file-name=GSR313E_10511(1)_0.pdf)

Law	Relevant provision	Application
Information Technology (Reasonable security practices and procedures and sensitive personal data or information) Rules, ³⁵	Rule 3	Privacy The Rules lay down points on disclosure and distinguishes between what is personal data and what is sensitive personal data. For example, an AI solution collecting data on your age might not have as severe repercussions as collecting your biometric information.
	Rule 4 and Rule 5	Drafting privacy and security policy including Obtaining consent from the person whose personal information is collected.
	Rule 8	Security To ensure the security of the AI, compliance with the below standard is recommended: International Standard IS/ISO/IEC 27001 on "Information Technology - Security Techniques - Information Security Management System - Requirements".
Information Technology (Intermediary Guidelines and Digital Media Ethics	Section 2 (w) of the Information Technology Act 2000	Definition of intermediary
	Rule 3 1(a) and (b)	Privacy policy and Terms of Use are essential for the intermediary.
	Rule 3 1(c)	Users should be informed periodically, atleast once a year of complying with the above and not misusing the platform

³⁵<https://www.meity.gov.in/writereaddata/files/Revised-IT-Rules-2021-proposed-amended.pdf>

Law	Relevant provision	Application
	Rule 3 (2)	<p>Grievance redressal mechanism promptly displayed on platform for users to report violations of law and intermediary policy.</p> <p>Intermediary should take swift action within 24 hours of complaint.</p> <p>For example, an organization receives a report that their text to image software is being used to generate child pornography, they need to block the user.</p> <p>Since the case involves They are also under the obligation to mandatorily report this under POCSO Act, 2012 to the local police/ Special Juvenile Police Unit.</p>
	Advisory dated March 15, 2024 issued by MeitY ³⁶	<p>Duties of intermediaries</p> <p>i) Ensure software/AI does not post content that is grossly harmful, defamatory, obscene, pornographic, paedophilic, or other unlawful content.</p> <p>ii) Ensure software/AI does not permit bias, discrimination or threaten the integrity of the electoral process.</p> <p>iii) Ensure that software and AI systems do not allow bias, discrimination, or compromise the integrity of the electoral process</p> <p>iii) Terms of use should also include clauses on any misuse or abuse of AI by users</p> <p>iv) Take measures to prevent misinformation and deepfakes by ensuring the users can be identifiable if the users are allowed to create or modify synthetic content.</p>

³⁶[https://upload.indiacode.nic.in/showfile?actid=AC_CEN_45_76_00001_200021_1517807324077&type=rule&file-name=GSR313E_10511\(1\)_0.pdf](https://upload.indiacode.nic.in/showfile?actid=AC_CEN_45_76_00001_200021_1517807324077&type=rule&file-name=GSR313E_10511(1)_0.pdf)

Law	Relevant provision	Application
Copyright Act, 1957 ³⁷	Sections 51 and 52	A work that qualifies for copyright protection does not require registration to claim that protection. Additionally, under the Berne Convention, this protection applies globally. AI especially used for content generation needs to operate within exceptions of Section 52 to not constitute copyright infringement. For example, you cannot use copyrighted works as training data without the permission of the copyright holders.
Patents Act, 1970 ³⁸	Section 6	AI can be patented by the developer of the AI solution.
	Section 7	A patent cannot be claimed per se. An application needs to be made to the patent office. Some successful patented AI are Niramai Health Analytix and Grahaa Space. ³⁹
	Section 53	Term of the patent is 20 years from the filing date and in case of international patents, the international filing date under the Patent Co-operation Treaty
Consumer Protection Act 2019 ⁴⁰ Guidelines for Prevention and Regulation of Dark Patterns, 2023 ⁴¹	Guideline 2 (e)	Consumer complaints may apply in case of deficiency of service provided using AI, or unfair trade practices such as usage of dark patterns. For example, forcing users to buy a product upgrade without which the AI solution cannot continue to perform.

³⁷<https://www.indiacode.nic.in/bitstream/123456789/1367/1/A1957-14.pdf>

³⁸https://ipindia.gov.in/writereaddata/Portal/IPOAct/1_113_1_The_Patents_Act_1970_-_Updated_till_23_June_2017.pdf

³⁹<https://indiaai.gov.in/research-reports/ai-patents-driving-emergence-of-india-as-an-ai-innovation-hub>

⁴⁰<https://www.indiacode.nic.in/bitstream/123456789/15256/1/a2019-35.pdf>

⁴¹<https://consumeraffairs.nic.in/sites/default/files/The%20Guidelines%20for%20Prevention%20and%20Regulation%20of%20Dark%20Patterns%2C%202023.pdf>

Law	Relevant provision	Application
Indian Penal Code, 1860/ Bhartiya Nyaya Sanhita	Section 499-500 of the old IPC	An AI chatbot might generate content that adversely affects the reputation of a person and amounts to defamation. It may result in imprisonment and/or fine as provided under Bharatiya Nyaya Sanhita. A civil suit for defamation may also be instituted in which case, the person will be sued for compensation. For example, a chatbot generates false legal accusations when a person uses it for background verification ⁴² of a person. AI should be trained with safeguards against bias, hate speech and be based on reliable training datasets.
Article 21 of the Constitution of India ⁴³	Right to privacy	AI cannot infringe the privacy of individuals.
Digital Personal Data Protection Act, 2023 ⁴⁴		In addition to the IT Rules 2011, when the Digital Personal Data Protection Act, 2023 comes into force, organizations need to ensure the AI solutions they develop and/or use comply with the same. For a more comprehensive understanding of this Act, you may refer to Pacta's primer ¹³ on the Act.
Article 14-15 of the Constitution of India ⁴⁵		AI must not exhibit bias as it contravenes the right against discrimination. If AI is designed with exclusionary bias based on religion, caste, sex, gender, disability, etc., it violates fundamental rights.

⁴²<https://www.theverge.com/2023/6/9/23755057/openai-chatgpt-false-information-defamation-lawsuit>

⁴³<https://cdnbbsr.s3waas.gov.in/s380537a945c7aaa788ccfcd1b99b5d8f/uploads/2023/05/2023050195.pdf> not yet effective as on March 1, 2024

⁴⁴https://www.pacta.in/Data_Protection_Bill_For_NGO.pdf

⁴⁵<https://cdnbbsr.s3waas.gov.in/s380537a945c7aaa788ccfcd1b99b5d8f/uploads/2023/05/2023050195.pdf>

Law	Relevant provision	Application
		For example, it would be prudent to follow the Web Content Accessibility Guidelines ⁴⁶ to avoid the AI from excluding persons with disabilities.
The Protection of Children from Sexual Offences Act 2012 ⁴⁷	Section 19 and 20 read with 21	Blatant abuse of AI by end users is a possibility. AI development and deployment should have reasonable safeguards against this to avoid repercussions. For example, Abuse/misuse of AI by end users might involve children which need to be mandatorily reported.
Sector-specific		
Health sector: AI developed and deployed in Healthcare should follow the Guidelines ⁴⁸ framed by the Indian Council for Medical Research. Medical Devices Rules 2017 ⁴⁹ would apply to AI-based solutions used in diagnosis and treatment in healthcare. Developers and re-sellers of such AI solutions need to apply for licences for these devices.		
Finance sector: RBI directions on the Storage of Payment System Data ⁵⁰ in case of AI operating in fintech. The RBI governor also shared a checklist ⁵¹ for financial institutions that are planning to use AI integration.		

⁴⁶<https://www.w3.org/TR/WCAG21/>

⁴⁷https://www.indiacode.nic.in/handle/123456789/2079?sam_handle=123456789/1362


⁴⁸https://main.icmr.nic.in/sites/default/files/upload_documents/Ethical_Guidelines_AI_Healthcare_2023.pdf

⁴⁹https://cdsco.gov.in/opencms/opencms/system/modules/CDSCO.WEB/elements/download_file_division.jsp?num_id=OTg4NQ==

⁵⁰<https://www.rbi.org.in/commonperson/english/scripts/FAQs.aspx?Id=2995>

⁵¹https://rbi.org.in/Scripts/BS_ViewBulletin.aspx?Id=22318



 <https://www.pacta.in/>

 <https://www.linkedin.com/company/pactaindia/>

 @PactaIndia