Response dated 15 January 2021 to the Working Document: Enforcement Mechanisms for Responsible #AIforAll released by the NITI Aayog in November 2020

Dvara Research is an independent Indian not-for-profit research institution guided by our mission of ensuring that every individual and every enterprise has complete access to financial services. Our work seeks to address challenges for policy and regulation in India given the waves of digital innovation sweeping financial services, focussing on the impact on lower income individuals in the country. The regulation and protection of consumer data has been a core area of our recent research.

In this document we present our response to the **Working Document: Enforcement Mechanisms for Responsible #AIforAll** (**Working Document**) released by the NITI Aayog in November 2020.

This response is divided into in **two sections**. The **first section** (**Section I: A Framework to Identify High Risk Applications of AI**) responds to the specific research request for suggestions for a framework to identify high risk applications of AI, made at page 31 of the Working Document. This section presents early thinking on the indicators and variables that can be used to design a **risk matrix for AI** and rank risks from the use of AI across use cases, regardless of the sector they belong to.

The **second section** (**Section II: Feedback on the roles of the Oversight Body**) provides feedback on the seven roles that have been set out for the proposed Oversight Body in the Working Document. The discussion in this section focusses on identifying **seventeen specific** functions that the Oversight Body will need to discharge to perform the roles set out for it in the Working Document. Performing these seventeen functions will also help the Oversight Body to operationalise the principles of responsible AI as set out in *Working Document: Towards Responsible #AIforAll* and thus, constitute a complete implementation strategy. A finer understanding of the roles that need to be performed by the Oversight Body also helps size up its resource requirements and the enforcement powers required by the Oversight Body to perform the roles earmarked for it.

This note presents our early thinking on the governance of artificial intelligence. We are keen to develop this thinking over the course of the coming months. We will be happy to assist the department by offering research assistance on the topic and developing these ideas further.

Table of Contents

Executive Summary 3

Section I: A Framework to Identify High Risk Applications of AI 6

1. The rationale for risk-based regulation 6
2. Limitations of a risk-based approach to regulation and regulation of AI 6
3. A proposed AI risk matrix 6

Section II: Feedback on the roles of the Oversight Body 15

Appendix: Describing and operationalising the Principles for Responsible AI 22

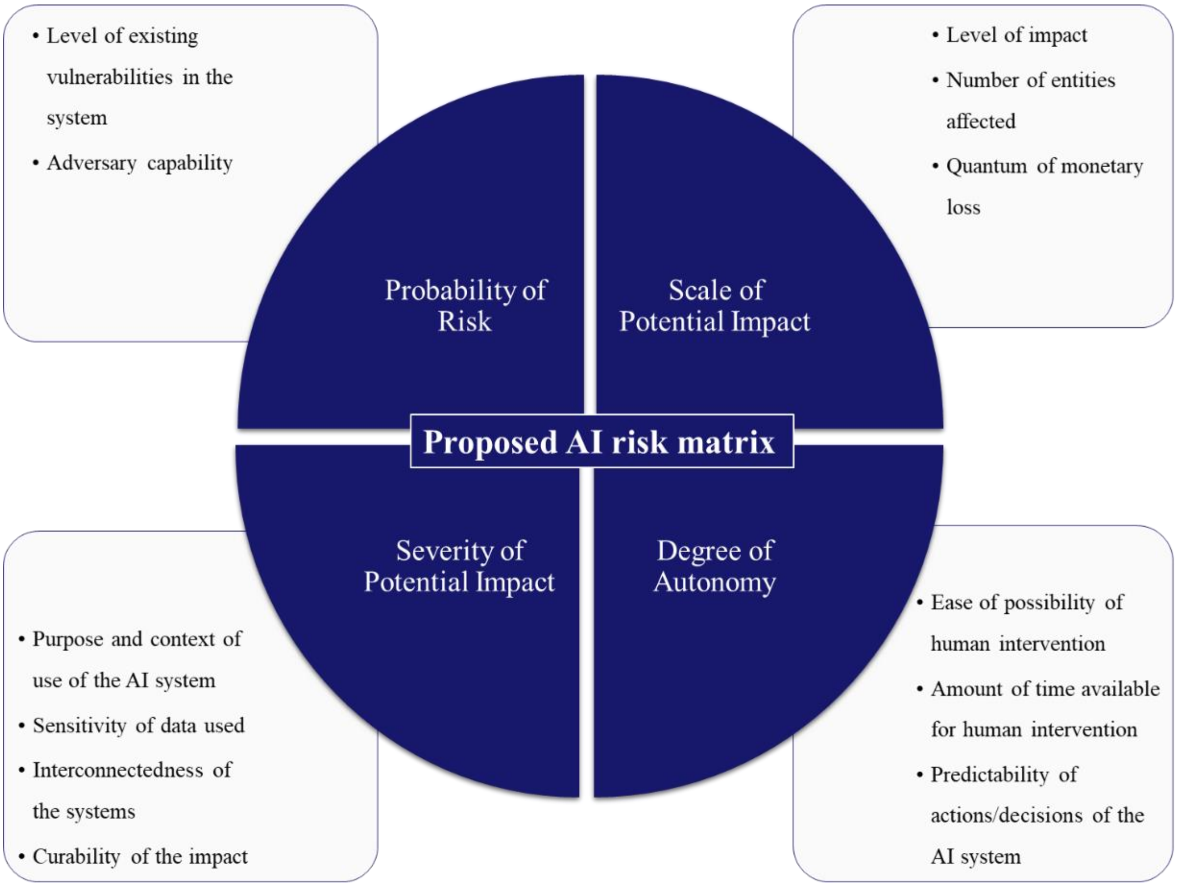
References 37

Executive Summary

This research note intends to offer a solution-oriented feedback to the Working Document: Enforcement Mechanisms for Responsible #AIforAll (**Working Document**) released by the NITI Aayog in November 2020. This research note is divided into two sections:

- **Section I: A Framework to Identify High Risk Applications of AI.** This section presents early thinking on a matrix that can *ex-ante* measure the risk from an AI system. **Risk** is identified as a perverse behaviour of an AI system (Bradley, 2019). The harms arising from such behavior can include actual or potential injury or loss to a consumer. Such injury or loss may be economically quantifiable or non-quantifiable (e.g., discouragement) or purely social in nature (Federal Deposit Insurance Corporation, 2019; Dvara Research, 2018). The **proposed AI risk-matrix (Figure 1)** helps regulators gauge the riskiness of AI use-cases and identify high-risk applications of AI from among the universe of AI applications. High-risk use cases indicate that a potential malfunction in the application's AI system adversely affect the social and economic lives of people on a large scale. By ranking the riskiness of each use-case of AI, this matrix can help regulators regulate AI proportionately.

Figure 1: An AI risk matrix for identifying high-risk use cases.

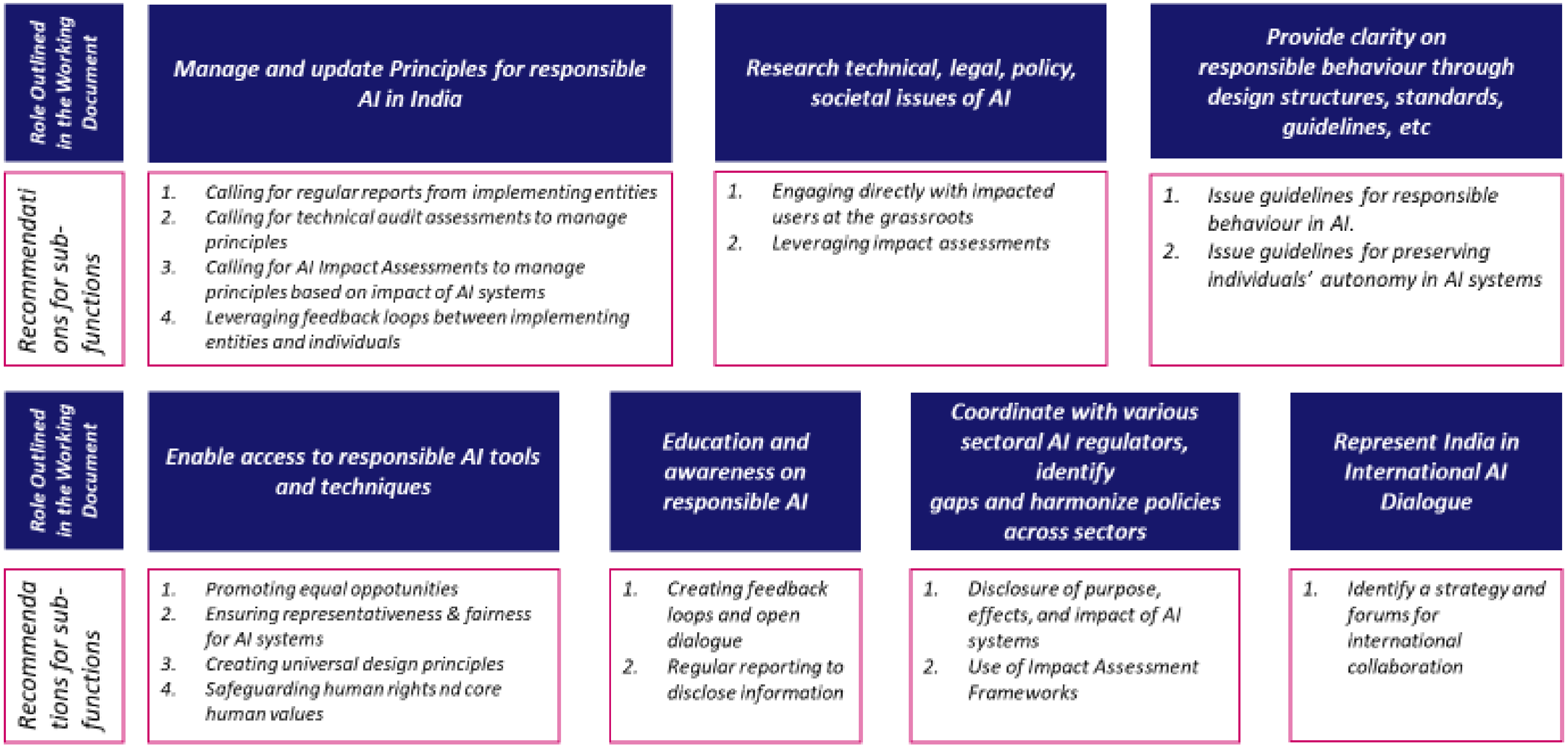

Proposed AI risk matrix
Probability of Risk
Scale of Potential Impact
Severity of Potential Impact
Degree of Autonomy
• Level of existing vulnerabilities in the system
• Adversary capability
• Level of impact
• Number of entities affected
• Quantum of monetary loss
• Purpose and context of use of the AI system
• Sensitivity of data used
• Interconnectedness of the systems
• Curability of the impact
• Ease of possibility of human intervention
• Amount of time available for human intervention
• Predictability of actions/decisions of the AI system

- **Section II: Feedback on the roles of the Oversight Body.** This section proposes seventeen functions that the proposed Oversight Body must perform under the seven broad roles earmarked for it. This list of seventeen functions is firmly grounded in the Principles of Responsible AI mentioned in *Working Document: Towards Responsible #AIforAll* (NITI Aayog, 2020a). These functions are also entrenched in the operations of similar organisations i.e. government entities created to guide and supervise the use of AI in other jurisdictions such as the United States of America (USA), the United Kingdom (UK), European Union (EU), Australia and Singapore. Incorporating these functions under the relevant roles set out in the Working Paper is useful because they:

  i. Provide substance to the broader roles of the Oversight Body set out in the Working Document and gives an indication of the full range of actions that the Oversight Body must perform.
  ii. Ensure that the roles of the Oversight Body are well-grounded in the Principles of Responsible AI identified in *Working Document: Towards Responsible #AIforAll* (NITI Aayog, 2020a) and well recognised globally. As such it creates a well-rounded implementation strategy by closely aligning the activities performed by the Oversight Body with the wider Principles of Responsible AI.
  iii. Help in assessing the resource requirement, institutional designs and enforcement powers required by the Oversight Body to perform the roles earmarked for it.

  The infographic below sets out the functions that we propose the Oversight Body must perform under each role.

Figure 2: Recommendations for the functions of the Oversight Body.



Role Outlined in the Working Document	Manage and update Principles for responsible AI in India	Research technical, legal, policy, societal issues of AI	Provide clarity on responsible behaviour through design structures, standards, guidelines, etc
Recommendations for sub-functions	1. Calling for regular reports from implementing entities 2. Calling for technical audit assessments to manage principles 3. Calling for AI Impact Assessments to manage principles based on impact of AI systems 4. Leveraging feedback loops between implementing entities and individuals	1. Engaging directly with impacted users at the grassroots 2. Leveraging impact assessments	1. Issue guidelines for responsible behaviour in AI. 2. Issue guidelines for preserving individuals' autonomy in AI systems

Role Outlined in the Working Document	Enable access to responsible AI tools and techniques	Education and awareness on responsible AI	Coordinate with various sectoral AI regulators, identify gaps and harmonize policies across sectors	Represent India in International AI Dialogue
Recommendations for sub-functions	1. Promoting equal oppotunities 2. Ensuring representativeness & fairness for AI systems 3. Creating universal design principles 4. Safeguarding human rights nd core human values	1. Creating feedback loops and open dialogue 2. Regular reporting to disclose information	1. Disclosure of purpose, effects, and impact of AI systems 2. Use of Impact Assessment Frameworks	1. Identify a strategy and forums for international collaboration

Section I: A Framework to Identify High Risk Applications of AI

This section presents some early thinking on the design of a matrix that can *ex-ante* measure the risk from an AI system. Risk is identified as a perverse behaviour of an AI system (Bradley, 2019). The harms arising from such behavior can include actual or potential injury or loss to a consumer. Such injury or loss may be economically quantifiable or non-quantifiable (e.g., discouragement), or purely social in nature (Federal Deposit Insurance Corporation, 2019; Dvara Research, 2018). The proposed AI risk-matrix helps regulators gauge the riskiness of AI use-cases and identify high-risk applications of AI from among the universe of AI applications. High-risk use cases indicate that a potential malfunction in the application's AI system could have an adverse bearing on the social and economic lives of people on a large scale. This matrix will help regulators rank the riskiness of each use-case of AI help them regulates for AI, proportionately.

1. The rationale for risk-based regulation

Risk-based regulation allows regulators and governments to (a) anticipate risk and (b) design policies that helps reduce the occurrence of risk.

Risk-based regulation acknowledges that risk cannot be reduced to zero. It therefore works to "*instil processes and practices – training programmes, regular simulations, audits, crisis management units – that help prepare public and private organisations to recognise and manage these potentially catastrophic events*" (Heijden J. v., 2019). In essence, the approach uses the riskiness of an activity as a criterion to allocate regulatory capacity, and guide regulation.

2. Limitations of a risk-based approach to regulation and regulation of AI

The most pronounced limitation of a risk-based approach is its heavy reliance on probabilistic modelling. Although the modelling seems to work on paper, it may not stand the test of practice (Heijden, 2019). A risk-based approach may in fact give a false sense of security and encourage people to take on bigger risks than they would have otherwise taken. Further, the approach tends to underplay low risk cases by focussing on high-risk cases to the exclusion of low-risk cases (Black & Robert, 2012). Often, low-rated risks can be unstable and potentially accumulate and aggravate into higher risks.

This limitation of risk-based regulation is extremely relevant for regulators seeking to develop a risk-based approach for regulating AI wherein the risks are dynamic and ever-evolving. Knowledge about the way in which AI functions and the impact it can have is still being developed. In this context, the significance of a risk may not be fully understood. Cases that appear low-risk may avalanche into high-risk cases. An appreciation of harms from AI may require analysis from diverse specialist communities, and the potential harms may still not be immediately apparent, definable, or quantifiable (Prasad, 2019).

Therefore, in the case of AI regulation it is best to not look at the regulation of high-risk and low-risk use cases as a zero-sum game. Efforts need to be made to increase regulatory capacity to keep up with the ever-expanding size and number of regulated entities, and even utilise AI for regulation. Until the regulator develops technology that can deal with the challenge of regulating at scale, a risk-based framework can help regulators in identifying high-risk use cases. It is worth emphasising that any quantitative risk-based framework such as the one presented in this note will be severely reductionist in nature and will not be able to completely capture the full extent of risks in AI. Therefore, these risk-based matrices should be used as tools to complement the qualitative assessments of the regulator.

3. A proposed AI risk matrix

This section introduces an AI risk matrix that can help in assessing the risks from AI systems. The objective of this matrix is to identify those high-risk use-cases of AI in which a potential malfunction can have an adverse bearing on the rights and quality of life of individuals on a large scale. Identifying these high-risk applications will help the regulators allocate greater attention and regulatory capacity in overseeing these use-cases. This matrix can be applied horizontally across all use-cases of AI regardless of the sector they belong to. Further, the matrix is designed to be sensitive to qualitative differences in the nature of harm that can arise from a potential malfunction of the AI system.

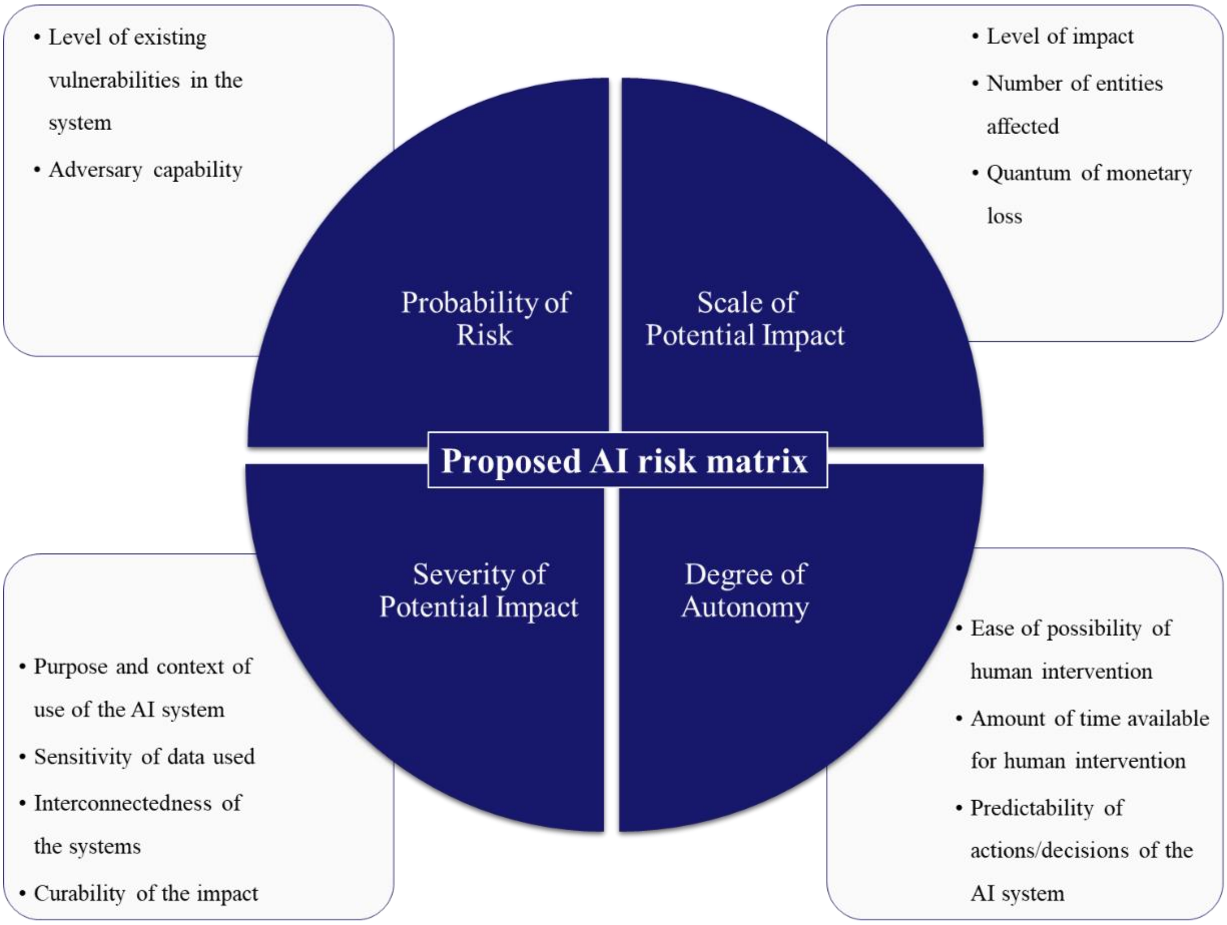
The important criteria for any risk matrix include (a) the probability of occurrence of risk (b) the scale of risk and (c) the severity of risk (ScienceDirect, 2016). The likelihood of risk and its consequences are important parts of risk analysis (United Nations Economic Commission for Europe, 2012). Two additional dimensions become salient in gauging the risk from AI i.e., the ability to control the AI system and the ability to predict the outcome (Buiten, 2019). **Degree of autonomy of the algorithm** is a widely accepted indicator of ability to control (International Committee of the Red Cross, 2019). Further, the **ability to predict** the adverse impact of algorithms is dependent on the **context** that they are used in, the **sensitivity** of the data they use, the **interconnectedness** of other systems that are dependent on the decisions made by one algorithm and the **curability of the impact**. We utilise these indicators to build a matrix that can capture the riskiness of AI use cases, across all sectors. The four main criteria for the risk-based matrix are set out below.

i. the **probability** that a harm can materialise from the malfunction of an AI system;
ii. the **scale** of the harm;
iii. the **autonomy** of the AI system (which captures the ability to control the emergence of the harm), and
iv. the **severity** of the harm.

Figure 3 provides a graphical representation of the proposed AI risk-matrix. The blue quadrants in the figure set out the main criteria for the risk-matrix. The adjoining rectangles present the potential variables that can be used to measure the criteria.

The merits of each of these indicators and how they measure the riskiness of AI applications is set out in detail in the following sections.

Figure 3: AI Risk matrix


• Level of existing vulnerabilities in the system
• Adversary capability
• Level of impact
• Number of entities affected
• Quantum of monetary loss
Probability of Risk
Scale of Potential Impact
Proposed AI risk matrix
Severity of Potential Impact
Degree of Autonomy
• Purpose and context of use of the AI system
• Sensitivity of data used
• Interconnectedness of the systems
• Curability of the impact
• Ease of possibility of human intervention
• Amount of time available for human intervention
• Predictability of actions/decisions of the AI system

3.1. The probability of risk from the AI system

Risk can be described as a function of two factors: the likelihood of a risk event occurring, and the impact or consequences of that risk event (Council of Europe, 2013). This indicator deals with the estimation of the likelihood or probability of the occurrence of risk (United Nations Economic Commission for Europe, 2012). To reiterate, risk is defined as a perverse behaviour from the AI system.

AI Systems are vulnerable to both data security threats due to the data they hold and process, and to the threats emerging from the autonomous thinking of AI Systems. An estimation of the likelihood of risks must borrow from how the likelihood of risk is calculated in three domains i.e., information security, data privacy and autonomous systems. Therefore, we undertook a survey of regulatory practices in different jurisdictions to understand how the probability of risk is computed in the realm of information security and privacy. We found that the estimation of the likelihood of occurrence of risk is different in different jurisdictions.

The French data protection authority (CNIL): Likelihood of occurrence of risk refers to the possibility of a risk occurring. It primarily depends on:

i. *"the levels of vulnerabilities that supporting assets exhibit under threat, and*
ii. *the capacity of risk sources to misuse them"* (Commission Nationale Informatique & Libertés, 2018).

The level of vulnerabilities that supporting assets exhibit under threat, refers the extent to which the existing properties of supporting assets can be exploited to execute a threat. The capacity of risk sources to misuse the existing vulnerabilities seeks to capture the technical abilities of the source of risk to exploit the vulnerabilities and cause the risk (Commission Nationale Informatique & Libertés, 2012).

The US Department of Commerce's National Institute of Standard and Technology (NIST) sets out a special publication 'NIST 800-30' explaining the estimation of the likelihood of information security threats (Romine, 2018). The factors considered for the estimation of likelihood are:

i. *"adversary intent,*
ii. *adversary capability, and*
iii. *adversary targeting."*

It appears that the likelihood of occurrence of risk in the domain of data security tries to measure the level of inherent vulnerabilities in the system, and the intent and ability of an adversarial threat to exploit them.

The European Parliament uses the following factors to estimate the probability of harm or damage, with specific reference to AI (European Parliament, 2020):

i. *"the role played by algorithms in the decision-making process,*
ii. *the intricacy of the decision taken, and*
iii. *if the effects are reversible."*.

The likelihood of occurrence of risk in the realm of AI is relatively new and the EU's thinking on risk-based regulation of AI is still evolving. Currently the relation between these indicators and the likelihood of occurrence of risk appears nebulous. For instance, establishing the role played by the AI in decision-making is helpful for courts to establish the connection between the act of the algorithm and the harm inflicted, and thereby apportion liability (Buiten, 2019; Kingston, 2016). However, it is unclear how this indicator can help regulators to *a priori* estimate the likelihood of occurrence of perverse behaviour of AI. Therefore, we use these variables to measure the riskiness of AI under the fourth variable, i.e., *severity of impact.*

We propose the following sub-indicators drawing from the French and American regulators for measuring the indicator of likelihood of occurrence of risk:

i **The levels of vulnerabilities that exist in the system**. A vulnerability is understood as a flaw or weakness that can be accidentally triggered or intentionally exploited, resulting in a security breach or violation of policy (Health Insurance Portability & Accountability Act Collaborative of Wisconsin, n.d.). Organisations can gauge the vulnerabilities in their systems in different ways. Some of them include (a) defining threat scenarios which allows organisations to anticipate how threats can translate into harms (Romine, 2018) (b) tracking the frequency of risky events that are reported to regulators (c) tracking information gathered during inspections and site visits, independent field audits, grievances raised with regulators, information submitted by the regulated entities including financial information of regulated entities and information or assistance requests made to the regulator and (d) supervising the collection of data and sample data (State of New South Wales, 2016). A high score on this sub-indicator reflects higher riskiness in the AI system.

ii **Adversary capability:** Where the adversaries are identifiable, their ability to cause harm must be considered to estimate the likelihood of occurrence of harm. For instance, the adversary's scale of operation, technical prowess, ability to further sell the proprietary algorithm or personal data etc. should be used as factors for assessing the adversary's capability and likelihood to cause harm. Other fundamental inputs to risk assessments like quantitative and qualitative data in combination with other intelligence can also be used to determine the likelihood and impact of risk. A high score on this sub-indicator reflects a higher riskiness in the AI system.

3.2. The scale of risk emanating from an AI system

AI and advanced analytics have many positive benefits, but they can lead to severe unintended (or malicious) consequences for individuals, organisations and society. Issues can emerge ranging from digital safety, data breaches and bias to national security and systemic concerns (Cheatham et al., 2019). Therefore, the scale of impact is an essential parameter in the risk matrix to assess the riskiness of AI systems.

In terms of scale, risks can potentially have consequences at the individual level, at the community level and at the systemic level (Cheatham et al., 2019). In addition to the three levels identified, the magnitude of loss in terms of the number of people/communities affected, and the monetary costs that they will have to incur due to malfunction of AI are also indicators of the scale of risk emanating from AI (European Parliament, 2020). Therefore, we propose the following sub-indicators to measure the indicator of scale of risk:

i. **The level of impact**, i.e., if the impact is at the individual, community or at the systemic level. A high score on this sub-indicator denotes a higher risk score.

ii. **The number of entities affected** at the individual or community level. A high score on this sub-indicator denotes a higher risk score.

iii. **The quantum of monetary loss suffered** by the individuals or communities. This refers to the cost of damages caused by a malfunction of the AI or the amount of money that will be needed to undo it. A high score on this sub-indicator denotes a higher risk score.

3.3. The degrees of autonomy of the AI system

The use of AI systems that show human-like intelligence in narrow domains is becoming common (Marda, 2018). These systems exhibit varying levels of autonomy in performing different tasks with minimal or no human involvement (Walch, 2020). In essence, the level of autonomy of an AI system can be estimated based on the complexity of the algorithm and the level of human involvement (Walch, 2020). Autonomous AI systems that are designed for less human intervention has created accountability safety concerns in the past. Fatal road accidents by autonomous vehicles and racist chatbots have are some examples of concerns emerging from autonomous AI systems (Deamer, 2016; Perez, 2016). These levels of autonomy are an important dimension for regulators and for the consideration of riskiness of AI (Buiten, 2019).

Autonomy in AI systems results "*from the delegation of a decision to an authorised entity or system capable of taking action within specific boundaries*" and away from humans (Haddal & Frazar, 2018). Such a system (a) *"receives information from its environment through sensors ("sense")* (b) *"processes these data with control software ("think")* and (c) *based on its analysis, performs an action ("act") without further human intervention*" (International Committee of the Red Cross, 2019).

The sense-think-act model of AI systems make them independent of human intervention. While this may have benefits, a majority of concerns and risks arise from the **unpredictability** in the actions and decisions of the systems. In order to minimise the risks from such AI systems (however autonomous they may be), regulators around the world indicate that predictability of an AI system should be a crucial principle involved in the designing of AI systems. Separately, direct human control either over the entire system or over some specific functions is being used as a preventative measure. Such a human-on-the-loop supervisory mechanism must consider three features:

i. *"**situational awareness,*** i.e., awareness and sufficient knowledge of the human operator or supervisor regarding the state of the system at the time of the intervention;
ii. *enough **time to intervene,*** **i.e.,** time available for the human operator or supervisor to intervene in the event of an unfavourable outcome or to stop an ongoing process;
iii. *a **mechanism through which to intervene*** (a communication link or physical controls) in order to take back control, or to deactivate the system should circumstances require" (International Committee of the Red Cross, 2019).

Keeping in mind the various dimensions of autonomy in the application of AI systems, we recommend measuring:

i. The **ease of human intervention** present in the various processes of an AI system i.e. the ease with which a human can intervene in the processes of an AI system. Typically, a high score on this sub-indicator reflects lower riskiness in the AI system;
ii. The **amount of time available for human intervention** in case of an unfavourable action or decision by the AI system. Typically, a high score on this sub-indicator reflects lower levels of riskiness in the AI system.
iii. **Predictability of the actions or decisions of the AI system** with respect to the purpose of its deployment. Typically, a high score on this sub-indicator reflects a lower level of riskiness in the AI system (Walch, 2020; International Committee of the Red Cross, 2019; SAE International, 2018).

3.4. The severity of potential impact from risks emerging from an AI system

This indicator seeks to gauge the severity i.e., the depth and quality of risk that can arise from a perverse behaviour in the AI system. Discrimination, exclusion, exploitation, infringement of rights, privacy risks, monetary risks, physical safety, and manipulated political discourse are some examples of the different kinds of risks that can arise (NITI Aayog, 2020a; AI Now Institute, 2020; OECD, 2019; Brundage, 2018; Elish, 2019).

In terms of severity, each of these risks can have a different impact on individuals and communities depending on the nature of risk. For instance, the potential impact of a malfunctioning news aggregator algorithm could be milder than a malfunctioning predictive policing algorithm that could target particular communities and severely infringe individuals' rights (OECD, 2019). Further, an AI system could affect a small section of the population, but at the same time have a deep impact. For example, an AI system could impact only one neighbourhood in a city or a niche segment of women who use a specialised product, but the impact could heavily entrench exclusion or discrimination against that section (Ingold & Soper, 2016; Vigdor, 2019). Therefore, a severity parameter that captures significant risks that can have a deep (and potentially invisible) impact that might not otherwise seem significant in terms of scale is also crucial for an AI risk matrix.

However, the severity parameter would only signal how severe a potential impact could be. The true severity of impact could be difficult to measure given the nature of data-related harms. Data related harms are hard to anticipate, manifest in unfamiliar ways and are often not easily quantifiable (Prasad, 2019). Given the diversity of harms and the difficulty in quantifying and comparing them, we need a framework that can guide the assessment of severity. To provide a scale for reliably assessing the severity of impact from risk, the framework would have to justify why some kinds of impact should be considered more severe than others. A set of parameters that could *collectively* justify this ranking of the different kinds of impact include:

i. **Purpose of the AI system and context of use:** The purpose for which the AI system is being deployed and the kind of decisions it is trusted to make, influence the severity of impact from a failure or malfunction in the AI system.
To put it graphically, predictive analytics and automated AI decision support systems run on input data, with the output being a decision. These systems are deployed to produce a range of outputs, from relatively trivial decisions such as matching films and restaurants to individuals' preferences, to helping them purchase goods they have been searching for a while, to more significant ones such as sorting credit card applications and admission applications in universities and potentially life changing decisions such as determining which individual must receive an organ (Stanford Encyclopedia of Philosophy, 2020). If a film or restaurant sorting algorithm malfunctions, individuals may be exposed to material or food they may dislike. However, the stakes are much higher when an algorithm designed to assess admission applications malfunctions or is inherently biased against a particular community. Therefore, the purpose of the AI and the context of its use matter.This parameter is also used in conducting data protection impact assessments (DPIA) under the General Data Protection Regulation (GDPR) in the European Union (EU) (Information Commissioner's Office, 2020).
Purposes that have a bearing on the civil and economic well-being of individuals attract a higher score. A high score on this sub-indicator reflects higher riskiness of the AI system.

ii. **Sensitivity of data used by the AI system:** The sensitivity of personal data processed by AI systems can have implications for the kind of impact a system can have. For instance, AI systems could have more severe ramifications on individuals when they process highly sensitive data like biometrics or health data than less sensitive data like individuals' e-commerce purchase histories (Information Commissioner's Office, 2020). Processing personal information through AI and ML can help in revealing further and potentially more sensitive information about users, thus compromising their privacy and also exposing them to additional harms such as those of discrimination (Kerry, 2020). A higher score in this sub-indicator reflects greater sensitivity of data and also higher riskiness of the AI system.

iii. **Interconnectedness of systems:** The modern governance and service delivery design include complex interconnected technology systems that interact with each other through APIs, interconnected databases etc. This can be referred to as "coupling" between systems (Perrow, 1999). Tightly coupled systems are highly interconnected, almost seamless, and show cascading effects instantly. Loosely coupled systems on the other hand are less interconnected, have more friction and have slow cascading effects. In a highly coupled system, a shortcoming in any one technology system or interconnecting program could have severe cascading effects on the ultimate output (Perrow, 1999). Risks emerging from tightly coupled AI systems, similarly, could lead to severe end-results. For instance, when several administrative services rely on the same algorithm to base their decisions and the underlying algorithm is faulty, then all decisions based in separate departments will tend to have the same flaws. The National Institute of Science and Technology (NIST) in the USA found that a majority of facial recognition algorithms in the industry were biased against non-Caucasian faces. For one-to-one matching, NIST established that false positives were higher by a factor of 10 to 100 for Asian and African-American faces relative to images of Caucasians. This implies that when presented with pictures of suspected criminals, the algorithm may wrongly match an innocent Asian or African-American and expose them to legal proceedings (National Institute of Standards and Technology, 2019). Needless to say, all departments and private facilities that avail of these algorithms could further perpetuate this malfunction of the algorithm. Interconnected algorithms thus have a cascading effect by affecting other systems that are built on top of them. A high score on this sub-indicator reflects higher riskiness of the AI system.

iv. **Curability of the impact:** One of the steps in conducting a DPIA includes identifying measures to reduce risks caused due to data processing activities (Information Commissioner's Office, 2020). A similar parameter could be used to assess the severity of potential impact from AI systems. The degree to which the potential impact can be cured *ex post* once the risks from AI systems materialise could become one of the ways to rank impact. Impact that can be cured meaningfully could be ranked lower than impact that cannot be cured meaningfully. For instance, monetary loss that can be cured by repayments could be ranked lower than an infringement of rights that cannot be remedied.

To be clear, the curability of impact does not absolve AI users from the responsibility of considering overall risk. Curability is only one of the parameters to assess the severity of harm. A high score on this sub-indicator reflects lower riskiness of the AI system.

The proposed risk matrix is merely a suggested tool that could help regulators identify entities that pose a higher risk to individuals/ societies when their AI applications malfunction. The field of AI is relatively nascent, and the risk-based approach of regulation has not been tested. The risk-based approach of AI governance has its limitations. The most considerable limitation being that while this could be a useful tool for regulators, it might not be so helpful to the consumers, it may cause supervisory bodies to overlook less risky activities that may further snowball into higher risks.

The table on the adjacent page summarises this discussion.

Table 1: Summary of the AI Risk Matrix

Indicator	*Variables*	*Direction of Impact*
Probability of Risk	(i) level of vulnerabilities that exist in the system	**Higher** score reflects **higher** risk
	(ii) adversary capability	**Higher** score reflects **higher** risk
The Scope of Risk Emanating from AI	(i) level of impact	**Higher** score reflects **higher** risk
	(ii) number of entities affected	**Higher** score reflects **higher** risk
	(iii) quantum of monetary loss suffered	**Higher** score reflects **higher** risk
Degrees of Autonomy in the System	(i) ease of possibility of human intervention	**Higher** score reflects **lower** risk
	(ii) amount of time available for human intervention	**Higher** score reflects **lower** risk
	(iii) predictability of actions/decisions of the AI system	**Higher** score reflects **lower** risk
Potential Severity of Impact	(i) purpose of AI system and context of use	**Higher** score reflects **higher** risk
	(ii) sensitivity of data used	**Higher** score reflects **higher** risk
	(iii) interconnectedness of systems	**Higher** score reflects **higher** risk
	(iv) curability of the impact	**Higher** score reflects **lower** risk

Section II: Feedback on the roles of the Oversight Body

This section provides feedback on the seven roles set out for the Working Body, in the Working Document. The discussion in this section focusses on identifying **seventeen** specific functions that the Oversight Body will need to discharge to perform the wider roles set out for it in the Working Document. A finer understanding of the roles that need to be performed by the Oversight Body would help in estimating the resource requirements, the nature of enforcement powers and the institutional structure best suited for the Oversight Body for carrying out its functions and performing these tasks.

Further, performing these seventeen functions will also help the Oversight Body to operationalise the Principles of Responsible AI as set out in the *Working Document: Towards Responsible #AIforAll* and thus comprise a complete implementation strategy. These seventeen functions also map onto globally, well-established Principles of Responsible AI and bring the working of the Oversight Body at par with global standards[1]. **Table 2**, titled "*Key objectives and functions for the Oversight Body*" maps these functions to the roles identified in the Working Document.

In this table, the first column titled "*Role of the Oversight Body*" reproduces the role of the Oversight Body identified in the Working Document together with a short summary of the role. The second column titled "*Key functions for the Oversight Body*" presents the functions that the Oversight Body can perform, in addition to those mentioned in the Working Document, to actively perform the roles identified for it.

Table 2: Key functions for the Oversight Body

Role of the Oversight Body	**Key objectives and functions for the Oversight Body**
Manage and update principles for responsible AI in India. This role requires the Oversight Body to– a. continuously monitor & update the Principles of Responsible AI use based on updates in use cases and technology, and	Our research on principles and best practices from other jurisdictions suggest that performing the following functions can help the Oversight Body discharge this role. These include: **i. Calling for regular reports from implementing entities.** The Oversight Body could ask implementing entities to disclose information including how AI systems function, how they reach an output and how negative impact from the AI systems are mitigated (Access Now, 2018b). Such information would help the Oversight Body in identifying vulnerabilities and stay updated on the working of algorithms, which is critical for managing and updating the Principles for Responsible AI. The Oversight Body, while

[1] *See*, Appendix 1 for a relationship between the global recognised principles of responsible AI, the functions set out in this research note and the roles of the Oversight Body set out in the Working Document.

b. design specific mechanisms in collaboration with various bodies to translate the principles into practice.	undertaking this exercise, must ensure that it upholds and preserves the intellectual property rights related to the algorithms. ii. **Calling for technical audit assessments to manage principles.** The Oversight Body could call for technical audits of algorithms, data and design processes used in an AI system by internal and external auditors depending on the risks the system may pose. The assessments and evaluations from these audits could provide key learnings to the Oversight Body which it can use to fine tune the Principles for Responsible AI and the mechanisms that can help entities better adhere to the principles (Villani, 2018; Amnesty International & Access Now, 2018; Smart Dubai, 2019). Such evaluation reports could also assist the Oversight Body in building trust and confidence in the technology (High-Level Expert Group on Artificial Intelligence, 2020). iii. **Calling for AI impact assessments to manage principles based on impact of AI systems.** The Oversight Body could call for AI impact assessments both prior to and during the development, deployment and use of AI systems. The impact assessments could evaluate the purpose and objectives of the AI system, and the benefits and risks of using the AI system (The Public Voice Coalition, 2018). These assessments could also incorporate due diligence measures like consultations with relevant stakeholders (affected groups, human rights organisations, AI experts etc.) to understand their potential impact (Villani, 2018; Amnesty International & Access Now, 2018). Such assessments can help the Oversight Body in fine tuning (a) principles for certain use-cases based on the potential impact and (b) designing mechanisms for implementing principles for responsible AI. iv. **Leveraging feedback loops between implementing entities and individual.** Feedback loops such as those created through grievance redress channels could provide valuable information to the Oversight Body about the ground-level impact of AI systems. Feedback channels could therefore serve as important sources of information for managing and updating principles for responsible AI (Access Now, 2018a). This will also help in achieving the objective of "*monitoring the impact of AI technologies at the consumer level*" set out for the Oversight Body in the National Strategy for AI #AIForAll (NITI Aayog, 2018).
Research technical, legal, policy, societal issues of AI This role requires the Oversight Body to–	Our research on principles and best practices from other jurisdictions presents certain key functions that the Oversight Body can perform to discharge this role. These include:

a. support multi-disciplinary research into AI ethics for advancing the field, identifying issues and addressing concerns around AI and to inform policy decision and guidelines; and b. Study and monitor impact of AI deployments on the ground.	**i. Engaging directly with impacted individuals at the grassroots.** The Oversight Body could gain valuable insights about the context and values in communities where AI systems are deployed to understand how AI systems interact with the community. This could help the Oversight Body in identifying issues and concerns arising from AI systems deployed on the ground (Fjeld et al., 2020; Access Now, 2018). **ii. Leveraging impact assessments.** The Oversight Body could leverage AI impact assessments during the design, development, deployment and use of AI systems to nurture deeper understanding about AI and study impact on the ground. The assessments could be undertaken with the help of relevant stakeholders including affected groups, human rights organisations, AI experts etc. to understand their potential impact (Villani, 2018; Amnesty International & Access Now, 2018). Such assessments can help the Oversight Body in making informed policy decisions and guidelines about AI systems.
Provide clarity on responsible behaviour through design structures, standards, guidelines etc. This role requires the Oversight Body to identify design standards, guidelines and benchmarks for responsible AI.	Our research on principles and best practices from other jurisdictions presents certain key elements that the Oversight Body can adapt into its guidelines to discharge this role: **i.** The Oversight Body can set out guidelines for responsible behaviour in AI. Some **elements of responsible AI** include- a. **Accuracy.** AI systems should be designed with high levels of (a) accuracy in the AI decision-making process and (b) attention to detail in the design and development phase of the AI system (Leslie, 2019). b. **Reliability.** AI systems should be designed in a manner where they operate dependably and as intended, even in cases where there may be changes in the operating environment (Fjeld et al., 2020). c. **Explainability.** AI system developers and implementing entities should be able to explain technical concepts involved in the decisions made by AI systems in a simple and coherent manner. Individuals should be able to require explanation by right (Fjeld et al., 2020). Further, information regarding the degree to which an AI system influences and shapes the implementing entity's decision-making process, the design choices for the AI system, and the rationale for deploying the system should also be available to individuals subject to the AI system (European Commission, 2019). d. **Verifiability and replicability.** The standards and guidelines developed should provide for mechanisms that ensure that an AI system presents similar results under similar conditions. The decisions made by the AI system should be documented to facilitate verification of the AI systems.

Further, the AI system should provide adequate information about its operations so that its decision-making process is verifiable and can be reviewed (The Federal Government, Germany, 2018).

e. **Universal design principles** that help in designing individual-centric AI systems that provide equitable access to AI products or services for all individuals regardless of the various barriers they may face (High-Level Expert Group on Artificial Intelligence, 2020)**.**

f. **Safety and security.** AI systems must have mechanisms that make it safe and secure, and auditable and transparent to uphold public trust of citizens or individuals on whom the applications are based (European Commission, 2019).

g. **Robust grievance redress.** Individuals who are subject to a decision made by an AI system should be entitled to challenge a decision made by the system (Fjeld et al., 2020). Entities should demarcate independent and visible processes for individuals to seek timely redress against adverse individual or societal effects of automated decisions (Amnesty International & Access Now, 2018). Further, the Oversight Body and the entities must create robust feedback loops with individuals who are impacted by AI systems to understand the on-ground impact post deployment of the AI systems (Access Now, 2018a).

ii. **Issuing guidelines for Preserving individuals' autonomy in AI systems,** can also help the Oversight Body in clarifying responsible behaviour, expected of the entities using AI**.** The guidelines issued by the Oversight Body can ensure–

a. **Transparency and explainability of AI systems:** Transparency refers to some level of accessibility to the data or algorithm, while explainability of AI refers to the ability to explain why or how a conclusion was reached (The Royal Society, 2019). Together, transparency and explainability are crucial instruments of accountability and instil confidence in consumers. They also ensure that important decisions about people are not made arbitrarily (The Royal Society, 2019). Regulators from other jurisdictions are issuing guidelines to help deployers of AI make their algorithms transparent and explainable. Some practices include designing data collection in a manner that renders itself explainable, extracting most relevant explanations in line with the differentiated skill sets of the audience and the domain in which AI is being used (Information Commissioner's Office & The Alan Turing Institute, n.a.).

b. **Contestability of AI system outputs:** Contestability, implies that users have the information they need to argue against a decision. Contestability provides people with agency to participate in

	automated societies (The Royal Society, 2019). It helps them adjudge if the decisions made about them are fair and the corrective actions they need to take if the decisions are unfavorable to them (The Alan Turing Institute, n.a.). c. **Robust data protection measures to protect individuals' right to privacy** (Fjeld et al., 2020; High-Level Expert Group on Artificial Intelligence, 2020; University of Montreal, 2017). In particular, individuals should have choice and control over how their personal data is used in AI systems (Fjeld et al., 2020). This implies that (a) personal data should not be used without the individual's informed consent (b) individuals must have the ability to opt-out of using services or products relying on AI systems and stop or limit the use of their personal data in AI systems (European Commission, 2019) (c) individuals should be allowed to update and rectify personal data records to ensure data quality (d) individuals must have the ability to remove their personal data completely from the processing cycle of AI systems and (e) privacy-by-default principles should be incorporated into AI systems to promote technical safety and security and to enhance the privacy afforded to individuals (Cavoukian, 2009; Agrawal et al., 2020).
Enable access to Responsible AI tools and techniques. This role requires the Oversight Body to– a. support projects for developing tools and technologies to enable access to responsible AI practices, b. enable data availability and sharing, c. promote research into data generation and identifying proxies, and d. create and adopt safe data sharing protocols.	Our research on principles and best practices from other jurisdictions presents certain key functions that the Oversight Body can perform to enable access to responsible AI tools and techniques. The Oversight Body could pursue the following functions for fulfilling this role: **i. Promoting equal opportunities.** The Oversight Body could promote equal opportunities by using AI systems to tackle power relationships and reduce socio-economic inequalities. In this way, the Oversight Body could ensure fair and equal access to technology and its benefits without discriminating between individuals or communities and prevent further widening of inequalities (Fjeld et al., 2020). Further, the Oversight Body can leverage technical audits by internal and external auditors to detect disparate impact from the AI system (High-Level Expert Group on Artificial Intelligence, 2020). Similarly, the Oversight Body can leverage algorithmic impact assessments to understand the impact of AI systems (Access Now, 2018b). **ii. Ensuring representativeness & fairness for AI systems**. AI systems require representative and high-quality data to provide safe and reliable outputs. The Oversight Body must take measures to ensure such datasets to reduce bias and improve accuracy (Fjeld et al., 2020). Further, some frameworks like the European Charter on AI in judicial systems have suggested specific protections for AI systems that

	process sensitive data on marginalised groups (caste, race, religion, genetic data etc.) (European Commission for the Efficiency of Justice, 2018; High-Level Expert Group on Artificial Intelligence, 2020). Further, AI systems could be trained to detect unfairness in input data and training data to systemically assess the representativeness and quality of data processed. **iii. Safeguarding human rights and core human values.** The standards and guidelines envisaged by the Oversight Body should ensure: a. **Protection of human rights and fundamental rights** that have been established and deemed inviolable under human rights law and the Constitution (Fjeld et al., 2020). b. **Protection of core human values** enshrined in human rights, fundamental rights, internationally recognised labour rights and other key instruments which seek to uphold human dignity and autonomy, promote human well-being, and pursue planetary well-being (Fjeld et al., 2020; G20, 2019). c. **Promotion of equal opportunities** by using AI systems to tackle power relationships and reduce socio-economic inequalities.
Education and Awareness on Responsible AI This role requires the Oversight Body to– a. engage with various stakeholders such as local communities, regional social organisations, academic institutions, public and private sectors for purposes including studying the impact of AI, reducing knowledge gaps and increasing awareness. b. leverage stakeholders to create open knowledge resources, case studies, needs assessment etc.	Our research on principles and best practices from other jurisdictions presents certain key functions that the Oversight Body can perform to promote education and awareness on responsible AI. The Oversight Body could pursue the following functions for fulfilling this role: **i. Creating feedback loops and open dialogue.** The creation of feedback loops and open dialogue with individuals who are impacted by AI systems would help to understand biases and other challenges on the ground post-deployment of the system (Access Now, 2018a). These learnings could inform research efforts by various entities to improve the use of Responsible AI. **ii. Regular reporting to disclose information.** The Oversight Body can help improve the trust and awareness of the public. To this end, it should ensure that organisations which make use of AI regularly report how outputs are reached, and the measures taken to minimise the impact of these decisions on the rights of individuals (Access Now, 2018b). This can help to improve awareness and reduce knowledge gaps.

Coordinate with various sectoral AI regulators, identify gaps and harmonise policies across sectors This role requires the Oversight Body to– a. identify risks with respect to AI use cases in co-ordination with various regulators; b. monitor existing policies and regulations gaps, inconsistencies, and other issues, and **c.** design policies, benchmarks, or ratify standards, and provide recommendations to regulators to address these risks.	Our research on principles and best practices from other jurisdictions presents certain key functions that the Oversight Body can perform to coordinate with various sectoral AI regulators, identify gaps and harmonise policies across sectors. The Oversight Body could pursue the following functions for fulfilling this role: **i. Disclosure of purpose, effects and impact of AI systems:** The Oversight Body should ensure that all information regarding the purpose of an AI system, the effects and impact it can have, and the decisions that it takes are disclosed by entities making use of AI (Access Now, 2018b). This will help the Body to monitor any inconsistencies and other issues that may arise and help provide recommendations to various sectoral regulators to address them. **ii. Use of Impact Assessment Frameworks:** The Oversight Body should mandate the use of impact assessments, both prior to and during the development, deployment and use of AI systems, which can provide insights on the level of risks that an AI system poses (The Public Voice Coalition, 2018). This can inform policy papers, and recommendations to sectoral regulators to address gaps and inconsistencies in the functioning of AI systems.
Represent India (and other emerging economies) in International AI dialogue This role requires the Oversight Body to– a. identify avenues for collaboration, such as at international forums and between universities, and present the perspective of India and other emerging economies, and b. facilitate collaborative research by developing cross-border data sharing protocols with relevant ministries.	Our research on principles and best practices from other jurisdictions presents certain key functions that the Oversight Body can perform to discharge this role. Different countries have prioritised different international forums for furthering international collaboration. For instance, China has helped in the development of international AI standards predominantly through bodies such as the International Organisation for Standardisation (ISO) and the International Electrotechnical Commission (IEC). This channel for standard setting has empowered countries and private entities to collaborate and promote AI standards and practices internationally. On the other hand, the United States has looked to develop governance frameworks and principles through organisations such as the G7, G 20 and the OECD. These efforts have helped in establishing widely accepted conventions in relation to AI despite not influencing AI standards or regulation (European Commission, 2019).

Appendix: Describing and operationalising the Principles for Responsible AI

This Annexure comprises two tables. Together these tables show that by performing the seventeen functions fleshed out in Section II, the Oversight Body can comply with the principles set out by the NITI Aayog in their Working Document towards Responsible #AIforAll.

The first table (Table A.1) sets out the **principles** laid out by the NITI Aayog in their Working Document towards Responsible #AIforAll, provides a **description of the principle** based on the literature survey of several jurisdictions including Singapore, UK, the EU and Australia. Based on this description of the principle and academic research, we highlight **functions** that the Oversight Body must perform to conform to the principle.

The second table (Table A.2) maps the **functions** identified in Table A.1 to the roles set out for the Oversight Body in the Working Document.

Table A.1: Principles for Responsible AI and their key dimensions

Principle in the Working Document	Description on the Principle based on global AI ethics frameworks	Functions needed to perform to conform to the Principle
Safety & reliability	The principle of “Safety” generally refers to the proper internal functioning of an AI system and the avoidance of unintended harms. Some documents also use a related term ‘reliability’ which means “*a system that is reliable is safe, in that it performs as intended, and also secure, in that it is not vulnerable to being compromised by unauthorised third parties”* (Fjeld et al., 2020, p. 4). The guide prepared by The Alan Turing Institute of United Kingdoms for responsible design and implementation of AI systems titled, *Understanding artificial intelligence ethics and safety*, emphasises on the need of thinking about designing AI systems in a manner that accurately and dependably operate as per the designers’ or the programmers’ expectations and intentions even	The key dimensions of the principles of “safety & reliability” appear to be: i. **Accuracy**: Often targeted at developers and programmers, accuracy promotes careful attention to detail on their part at the point of designing of AI solutions as well as while considering the validity of decisions. ii. **Reliability** on AI systems: AI systems must be designed in a manner that dependably operate, as intended, even in cases of anomalies and perturbations to the operating environment (Fjeld et al., 2020). iii. **Promote public trust**: AI systems must have all the mechanisms that allow for it for be safe and secure, auditable and transparent in a manner that does not erode public trust of citizens or users on whom the applications are based (Leslie, 2019).

	when confronted by anomalies or perturbations. Building an AI system that prioritises the technical objectives of accuracy, safety, reliability and robustness aid in preventing harmful outcomes and undermining public trust and reliance on such AI systems (Leslie, 2019). The *AI Ethics Principles* given by the Australian Government's Department of Industry, Science, Energy and Resources stipulates that AI systems reliably "*operate in accordance with their intended purpose*" although their operational lifecycle. For AI systems to be safe and reliable, all such safety measures that are proportionate to any potential risks must be adopted (Department of Industry, Science, Energy and Resources, Australian Government, 2019). This emphasis is similar to that placed on the principle of 'Safety and Security' in the Memorandum for the Heads of Executive Departments and Agencies on the subject, *Guidance for Regulations of Artificial Intelligence Applications* (Vought, 2020) and the *Ethical Principles for Artificial Intelligence* adopted by the Department of Defense (DOD) in the United States of America (U.S.A.) (U.S. Department of Defense, 2020).	

Privacy & security	“Privacy" is a dominant theme that occurs in several AI use ethics frameworks considered globally. This principle states that AI systems must respect the privacy of the individuals, both in cases of using their personal data and in providing agency to individuals over the decisions made that impact them. An exercise in the analysis of global ethics frameworks of AI use finds that ‘privacy’ appears as a principle in 97% of the frameworks considered for the analysis (Fjeld et. al, 2020, p. 4). Privacy preservation within AI systems and showing users and relevant stakeholders that the right processes and mechanism are in place to protect their personal data is an important aspect in establishing trust with them. Similarly, while designing AI systems, designers and programmers must be cognisant, and prepared for security risks through explicit efforts, such as by training and educating relevant personnel on potential harms and establishing processes to resolve the same, and by assessing any possibilities of adversarial attacks (The Institute for Ethical AI & Machine Learning, n.a.). The *AI Ethics Principles* given by the Australian Government’s Department of Industry, Science, Energy and Resources stipulates that AI systems respect and uphold privacy rights and data protection and ensure security of data throughout their life cycle. The principles further encourage	The key dimensions of the principles of “privacy & security” appear to be (Fjeld et al., 2020): i. **Consent**: Across frameworks, this notion broadly intends that individuals’ personal data is not used without their knowledge or permission. ii. **Control over the use of data**: Following from consent, in addition to not using individuals’ personal data without their permission, AI systems must not be designed in a manner that leaves individuals with no choice or control in how their personal data is used. iii. **Ability to restrict processing**: In connection to usage of data by AI systems and personal data of individuals, individuals must have the ability to stop or limit the usage of their data by an AI system. iv. **Right to rectification**: To promote the principle of privacy and security, AI systems must be built in a manner that allows individuals to complete personal data that is incomplete and/or correct data that is incorrectly recorded and being used to avoid the possibility of harm. v. **Right to erasure**: This dimension allows individuals the ability to remove their personal data completely from the processing cycle of AI systems. However, across several frameworks, the right to erasure includes providing individuals the option to completely remove their personal data from the public domain. vi. **Privacy-by-design**: This dimension pertains to promoting privacy within the AI systems by default in order to promote not just the technological safety and security of these systems, but also also to enhance the privacy afforded to individuals to whom these systems pertain. While this concept was first introduced as a set of 7 principles that were regarded as industry-grade best practices, following discourse has built upon these principles in order to operationalise them (Cavoukian, 2009; Agrawal et al., 2020).

	the use of technical methods such as proper data governance and management, appropriate data anonymisation, sound data analysis, identification of security vulnerabilities, resilience to attacks etc. by the designers to fulfil this principle (Department of Industry, Science, Energy and Resources, Australian Government, 2019). The principle of Security concerns an AI system's ability to resist external threats. In the *Principles of Artificial Intelligence Ethics for the Intelligence Community* given by the Office of the Director of National Intelligence, U.S.A., the principle of AI systems being 'secure and resilient' stipulates that best practices for maximizing reliability, security, and accuracy of AI systems be developed and employed to build resilience of AI systems in use against adversarial influence and attack (Director of National Intelligence, U.S.A., 2020).	
Transparency	The principle of "Transparency" requires that the design and implementation of AI systems is undertaken in a manner which permits oversight. This includes translating the operations of an AI system into comprehensible outputs, and providing information regarding where, when and how they are being used (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020).	The complexity and opacity of the technology related to AI poses the greatest challenge from a governance perspective. As such, the various global ethics frameworks have identified various dimensions under the principle of Transparency to respond to these challenges. These include: i. **Explainability**, where technical concepts and the decisions of outputs can be converted into a coherent format which can be evaluated (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020)

	The guidelines issued by the European Commission's *High Level Expert Group* highlight that transparency should be around "*the data, the system, and the business models*" (High-level Expert Group on Artificial Intelligence, 2019). IEEE's *Ethically Aligned Design* goes further to recommend that levels of transparency should be measured and tested through the creation of new standards, which will help to objectively assess the systems and determine its level of compliance (IEEE, 2019).	ii. The **right to information** which entitles individuals to know how automated and machine learning decision-making processes are achieved (Amnesty International & Access Now, 2018). iii. **Regular reporting**, an implementation mechanism where organisation should disclose information regarding how outputs are reached, and the steps taken to lessen the impact that such decisions may have on the rights of individuals (Access Now, 2018b). iv. **Notifying individuals** when they are interacting with an AI, or when an AI system makes a decision about them, so that individuals can experience the benefits of AI but are also provided with the choice of opting out of the use of such products in case of concerns (UK House of Lords, Select Committee on Artificial Intelligence, 2018).
Accountability	The principle of "Accountability" is highlighted across AI documents as a means to improve the trust of the public in AI systems (The Federal Government, Germany, 2018). It is generally accepted that necessary mechanisms must be established so that the responsibility and accountability of AI systems can be allocated among those who design the system, develop, and deploy it (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020). The G20 *Ministerial Statement on Trade and Digital Economy* highlighted that AI actors are accountable "for the proper functioning of AI systems" and for "respect of the other AI principles" (G20, 2019).	As such, global ethics frameworks have identified various dimensions under the principle of Accountability that can be mapped across three essential stages of the lifecycle of an AI system, namely design, monitoring, and redress. The various dimensions include: i. Building technologies that are capable of being **audited** (Villani, 2018; Amnesty International & Access Now, 2018), and utilise the learnings from the assessment as a feedback into the system to optimise the AI model (Smart Dubai, 2019). ii. **Verifiability** and **replicability** provide for mechanisms that ensure that an AI system presents similar results when reiterated under similar conditions, along with adequate information about its operations which can be corroborated (The Federal Government, Germany, 2018). iii. **Ability to appeal** allows the decision made by an AI system to be challenged by an individual made subject to the decision (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020)

		iv. Entities should demarcate independent and visible processes for seeking timely **redress** against adverse individual or societal effects of automated decisions (Amnesty International & Access Now, 2018). v. **Assessing the impact** of AI systems by evaluating its particular purpose, objectives, benefits and risks (The Public Voice Coalition, 2018).
Equality	All individuals deserve the same opportunities and protections from the increasing use of AI systems (European Commission for the Efficiency of Justice, 2018). There appear to be three distinct ways of understanding the principle under global ethics frameworks (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020): i. *Human rights*: The principle seems to be referring to the challenges AI systems could pose for equality through discrimination and disparate impact on different individuals (Access Now, 2018). ii. *Access to technology*: The principle seems to be referring to the equal access to benefits of AI systems and technology (European Commission for the Efficiency of Justice, 2018). iii. *Guarantees of equal opportunity:* the principle seems to be referring to a greater role for AI systems to eliminate relationships of dominance between groups based on power, wealth, knowledge etc. Further, it seems to focus on increasing social and economic	The key dimensions of the principles of "Equality," "Inclusivity and non-discrimination" and "Protection and reinforcement of human values" appear to be: i. **Protecting human rights and fundamental rights** that have been established and deemed inviolable under human rights law and the Constitution (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020). ii. **Protecting core human values** enshrined in human rights, fundamental rights, internationally recognised labour rights and other key instruments which (a) seek to uphold human dignity and autonomy and (b) promote human well-being and (c) pursue planetary well-being (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020; G20, 2019). iii. **Promoting equal opportunities** by using AI systems to tackle power relationships and reduce socio-economic inequalities. iv. **Ensuring fair and equal access to technology and its benefits** without discriminating between individuals or communities in order to prevent further widening of inequalities.

	benefits for all individuals by reducing social inequalities and vulnerabilities (University of Montreal, 2017).	
Inclusivity & non-discrimination	The principle requires AI systems to be designed and used in a manner that is impartial, maximises fairness and inclusive so that the costs and benefits are equally and justly distributed. In particular, the principle requires just distribution to groups that have been historically discriminated against. In this regard, the principle calls for greater diversity in AI design and development teams and greater involvement from the various communities in society in designing and developing AI systems. Bias in AI systems, training data, design and deployment should be mitigated as far as possible. Individuals should not be treated unfairly, unjustly discriminated against or stigmatised (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020; High-level Expert Group on Artificial Intelligence, 2019). The principle has also been interpreted to enable AI to detect bias and address discriminatory practices. The Montreal Declaration also discourages using AI systems for profiling individuals or creating filter bubbles which could obstruct an individual's personal development (University of Montreal, 2017).	

Protection & reinforcement of human values	The principle encourages AI to (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020): i. *Promote human values and human flourishing*: The principle broadly states that the purposes for which AI is deployed and the means of implementing AI should (a) respect humanity’s core values that uphold human dignity and autonomy and (b) promote human well-being and planetary well-being. In doing so, AI system should not favour some entities over others. Further, the ends to which AI is devoted, and the means by which it is implemented, should correspond with and be strongly influenced by social norms (University of Montreal, 2017; Department of Industry, Science, Energy and Resources, Australian Government, 2019; G20, 2019). ii. *Promote access to technology*: AI technology and benefits arising out of AI technology should benefit all individuals and entities equally in order to prevent widening inequality because of AI (High-level Expert Group on Artificial Intelligence, 2019). iii. *Leverage for the benefit of society*: AI systems should be used for public-spirited goals including respecting human dignity, freedom, democracy, equality and non-discrimination etc (High-level Expert Group on Artificial Intelligence, 2019; Department of Industry,	

	Science, Energy and Resources, Australian Government, 2019).	

Table A.2: Broad functions for operationalising the AI ethics principles.

Principle in the Working Document	**Functions that need to be performed to conform to the principles.**	**Relevant Role of the Oversight Body as set out in Working Document**
Safety & reliability	The key dimensions of the principles of "safety & reliability" appear to be: i. **Accuracy**: Often targeted at developers and programmers, accuracy promotes careful attention to detail on their part at the point of designing of AI solutions as well as while considering the validity of decisions. ii. **Reliability** on AI systems: AI systems must be designed in a manner that dependably operate, as intended, even in cases of anomalies and perturbations to the operating environment (Fjeld et al., 2020). iii. **Promote public trust**: AI systems must have all the mechanisms that allow for it for be safe and secure, auditable and transparent in a manner that does not erode public trust of citizens or users on whom the applications are based (Leslie, 2019).	From the Working Document, we identify the following functions stated as roles of an Oversight Body, or any other principles-implementing organisation that may potentially aid in achieving the principle of Safety and Reliability: i. **Monitor and Update.** By continuously monitoring and updating the principles of responsible AI use based on updates in use cases and technology, the Oversight Body would ensure that AI systems continue to remain reliable for the public. ii. **Reduce trust issues and apprehension of AI systems** for the general public by means of providing functional training to implementing agencies or other relevant stakeholders in standards, guidelines and best practices of responsible AI use and grievance redressal mechanisms can contribute towards achieving the principle of creating an AI system that is safe and reliable.
Privacy & security	The key dimensions of the principles of "privacy & security" appear to be (Fjeld et al., 2020):	From the Working Document, we identify the following functions stated as roles of an Oversight Body, or any other principles-implementing organisation that may potentially aid in achieving the principle of Privacy and Security:

	iii. Consent: Across frameworks, this notion broadly intends that individuals’ personal data is not used without their knowledge or permission. **iv. Control over the use of data**: Following from consent, in addition to not using individuals’ personal data without their permission, AI systems must not be designed in a manner that leaves individuals with no choice or control in how their personal data is used. **v. Ability to restrict processing**: In connection to usage of data by AI systems and personal data of individuals, individuals must have the ability to stop or limit the usage of their data by an AI system. **vi. Right to rectification**: To promote the principle of privacy and security, AI systems must be built in a manner that allows individuals to complete personal data that is incomplete and/or correct data that is incorrectly recorded and being used to avoid the possibility of harm. **vii. Right to erasure**: This dimension allows individuals the ability to remove their personal data completely from the processing cycle of AI systems. However, across several frameworks, the right to erasure includes providing individuals the option to completely remove their personal data from the public domain.	i. **Clarifying responsible behaviour.** By learning about and implementing standards and guidelines that are being developed around the world on responsible ways of managing technologies under specific contexts, the Oversight Body can stay on par with global operating standards of AI use. Separately, NITI Aayog’s working document titled, *Towards Responsible #AIforAll* identifies certain technical best practices to ensure privacy protecting and secure AI systems to (a) interpret AI decisions to instil trust and adoption, (b) allow data processing in a privacy protecting manner and (c) assess datasets for representation and "fairness" (NITI Aayog, 2020a). These include: i. To interpret AI decisions to instil trust and adoption: a. *“‘Pre hoc’ techniques such as Exploratory Data Analysis (EDA), concept extraction, dataset summarization, distillation techniques;* b. *‘Post hoc’ techniques for model explanation through input attribution (LIME, SHAP, DeepLift) and example influence matching (MMD critic, influence function, etc.”* ii. To allow data processing in a privacy protecting manner: *“Usage of methods such as federated learning, differential privacy, Zero Knowledge Protocols or Homomorphic Encryption.”* iii. To assess datasets for representation and "fairness": “*Tools such as IBM ‘AI Fairness 360’, Google ‘What-If’ Tool, Fairlearn and open source frameworks such as FairML.*”

	viii. **Privacy-by-design**: This dimension pertains to promoting privacy within the AI systems by default in order to promote not just the technological safety and security of these systems, but also to enhance the privacy afforded to individuals to whom these systems pertain. While this concept was first introduced as a set of 7 principles that were regarded as industry-grade best practices, following discourse has built upon these principles in order to operationalise them (Cavoukian, 2009; Agrawal et al., 2020).	
Transparency	The complexity and opacity of the technology related to AI poses the greatest challenge from a governance perspective. As such, the various global ethics frameworks have identified various dimensions under the principle of Transparency to respond to these challenges. These include: i. **Explainability**, where technical concepts and the decisions of outputs can be converted into a coherent format which can be evaluated (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020) ii. The **right to information** which entitles individuals to know how automated and machine learning decision-making processes are achieved (Amnesty International & Access Now, 2018).	From the Working Document, we identify the following functions stated as roles of an Oversight Body, or any other principles-implementing organisation that may potentially aid in achieving the principle of Transparency: i. **Best standards of data classification**. Ensure the documentation of the data sets and the processes that produce the AI system's decision, including the gathering and labelling of data and the algorithms used are at the best possible standard (European Commission, 2019). ii. **Ensure documentation of the decisions** made by the AI system to the best possible standards (European Commission, 2019). iii. **Information regarding the purpose of an AI system, and the effects and impacts that it can have, and the decisions it takes should be disclosed**. This will help to assess if laws regarding labour, workplace safety, privacy, liability, competition etc. are being maintained. This does not have to include the full disclosure of the algorithm, but rather allowing the effect of the algorithms to be independently assessed (Think 20, 2018; UK House of Lords, Select Committee on Artificial Intelligence, 2018).

	iii. **Regular reporting**, an implementation mechanism where organisation should disclose information regarding how outputs are reached, and the steps taken to lessen the impact that such decisions may have on the rights of individuals (Access Now, 2018b). iv. **Notifying individuals** when they are interacting with an AI, or when an AI system makes a decision about them, so that individuals can experience the benefits of AI but are also provided with the choice of opting out of the use of such products in case of concerns (UK House of Lords, Select Committee on Artificial Intelligence, 2018).	However, the code for decision-making algorithms used by public authorities must be made accessible to all (University of Montreal, 2017). iv. **Explainability of decisions made by the AI systems**. Explanation should be available about the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it. Therefore, ensure business model transparency (European Commission, 2019). v. **Transparency in the identification of AI systems and their uses.** There should be transparency in communication as well. AI systems should identify as themselves, and not as humans, to users. This allows the users to have the choice to interact with an AI system or a human counterpart. The level of accuracy provided by the AI system, as well as its limitations, must also be communicated depending on the use case (Access Now, 2018b; European Commission, 2019).
Accountability	As such, global ethics frameworks have identified various dimensions under the principle of Accountability that can be mapped across three essential stages of the lifecycle of an AI system, namely design, monitoring, and redress. The various dimensions include: i. Building technologies that are capable of being **audited** (Amnesty International & Access Now, 2018; Villani, 2018), and utilise the learnings from the assessment as a feedback into the system to optimise the AI model (Smart Dubai, 2019).	From the Working Document, we identify the following functions stated as roles of an Oversight Body, or any other principles-implementing organisation that may potentially aid in achieving the principle of Accountability: i. **Assessment and auditability of AI system processes.** Enable the assessment of algorithms, data and design processes by internal and external auditors, and make such evaluation reports available to improve trustworthiness of the technology (High-Level Expert Group on Artificial Intelligence, 2020). ii. **Auditing for check on fundamental rights.** Independent auditing should be conducted for those applications affecting fundamental rights, such as safety-critical applications[2] (European Commission, 2019).

[2] Safety-critical systems are those systems whose failure could result in loss of life, significant property damage or damage to the environment. There are many well-known examples in application areas such as medical devices, aircraft flight control, weapons and nuclear systems (IEEE, 2019).

	ii. **Verifiability** and **replicability** provide for mechanisms that ensure that an AI system presents similar results when reiterated under similar conditions, along with adequate information about its operations which can be corroborated (The Federal Government, Germany, 2018). iii. **Ability to appeal** allows the decision made by an AI system to be challenged by an individual made subject to the decision (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020) iv. Entities should demarcate independent and visible processes for seeking timely **redress** against adverse individual or societal effects of automated decisions (Amnesty International & Access Now, 2018). v. **Assessing the impact** of AI systems by evaluating its particular purpose, objectives, benefits and risks (The Public Voice Coalition, 2018).	iii. **Responsiveness of AI systems to reviews.** Ensure the ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome. The decisions made by an AI system must also be capable of being reviewed (and challenged) by individuals who are subject to the decisions made by the system. This requires the availability of accessible mechanisms for adequate redress against any unjust adverse impact (Think 20, 2018). iv. **Protection of human entities impacted by AI systems.** Provide due protection to whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI system (Think 20, 2018). v. **Impact Assessments.** Use impact assessments both prior to and during the development, deployment and use of AI systems to minimise negative impact. Assessments should be proportionate to the risks that the AI systems pose. Due diligence can also be conducted by consulting with relevant stakeholders such as any affected groups, human rights organisations, AI experts etc. (Amnesty International & Access Now, 2018; Villani, 2018). vi. **Professional responsibility of deploying agencies.** The architects of the digital society, such as the researchers, engineers, developers, and other professionals involved in the design, development, and deployment of AI systems must be conscientious of the influence that the technology can have on the wider society. This requires collaboration among the various actors to understand the diverse set of human norms and existing values that should be embedded in the AI systems. They should be guided by established professional values and practices. They should also take a long-term view while designing and developing the AI system, and anticipate future risks and impacts (Think 20, 2018; Villani, 2018).

		The Working Document also identifies some functions and best practices that can help entities in implementing the accountability principle. These include: i. Ensure provision for public auditing without opening up the system for unwarranted manipulation. ii. Assess the potential social impact of the system by evaluating error rates across sub population groups. iii. Auditing the algorithm by engaging with the open source, academic, and research community.
Equality	The key dimensions of the principles of "Equality," "Inclusivity and non-discrimination" and "Protection and reinforcement of human values" appear to be: i. **Protecting human rights and fundamental rights** that have been established and deemed inviolable under human rights law and the Constitution (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020). ii. **Protecting core human values** enshrined in human rights, fundamental rights, internationally recognised labour rights and other key instruments which (a) seek to uphold human dignity and autonomy and (b) promote human well-being and (c) pursue planetary well-being (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020; G20, 2019). iii. **Promoting equal opportunities** by using AI systems to tackle power relationships and reduce socio-economic inequalities. iv. **Ensuring fair and equal access to technology and its benefits** without discriminating between individuals or	The Working Document identifies some functions and best practices that can help entities in implementing the equality and inclusivity and non-discrimination principles. These include (NITI Aayog, 2020a; NITI Aayog, 2020b): i. preempting harms emerging from an AI system; ii. identifying and documenting goals for equality, non-discrimination and inclusion; iii. assessing the fairness and representativeness of datasets; iv. ensuring fairness goals are reflected when training the AI system; v. promoting collaboration with sectoral experts, social scientists and stakeholder community representatives; vi. enabling access to responsible AI tools and techniques through open technology projects, enabling data availability and sharing, and vii. establishing a grievance redressal mechanism. The Working Document could also consider the following functions and practices drawn towards implementing the principles: i. **Ensuring representative & high quality data** is available for input into the AI system to reduce bias and improve accuracy (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020). Some frameworks like the European Charter on AI in judicial systems have called for particular protections

	communities in order to prevent further widening of inequalities.	when sensitive data on marginalised groups (caste, race, religion, genetic data etc.) is processed (European Commission for the Efficiency of Justice, 2018; High-Level Expert Group on Artificial Intelligence, 2020). ii. **Training AI systems to detect unfairness** in input data and training data to systemically assess the representativeness and quality of data processed. iii. **Engaging directly with impacted stakeholders** at the grassroots before designing and developing AI systems to understand the context and values in those communities (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020). iv. **Creating universal design principles** that help in designing user-centric AI systems that provide equitable access AI products or services for all users regardless of the various barriers they may face (High-Level Expert Group on Artificial Intelligence, 2020). v. **Upskilling AI system administrators and users** for the safe deployment of AI systems and for ensuring their proper use by users (High-Level Expert Group on Artificial Intelligence, 2020). vi. **Preserving individuals' autonomy in AI systems** by ensuring (a) transparency and explainability of AI systems (b) contestability of AI system outputs and (c) robust data protection measures to protect privacy including consent, data rectification, erasure, privacy-by-design, ability to restrict processing etc. (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020; High-Level Expert Group on Artificial Intelligence, 2020; University of Montreal, 2017). vii. **Creating feedback loops and open dialogue** with individuals who are impacted by AI systems to understand biases and other challenges on the ground post-deployment of the system (Access Now, 2018).

References

Access Now. (2018, November). *Human Rights in the Age of Artificial Intelligence.* Retrieved from Access Now: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

Access Now. (2018, May). *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems.* Retrieved from Access Now: https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/

Access Now. (2018a, May). *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems.* Retrieved from Access Now: https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/

Access Now. (2018b, November). *Human Rights in the Age of Artificial Intelligence*. Retrieved from Access Now: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

AI Now Institute. (2020, June 14). *Submission to the European Commission on "White Paper on AI - A European Approach"* . Retrieved from AI Now Institute: https://ainowinstitute.org/ai-now-comments-to-eu-whitepaper-on-ai.pdf

Amnesty International & Access Now. (2018). *Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning Systems*. Retrieved from Access Now: https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf

Australian Prudential Regulatory Authority. (2018). *Probability and Impact Rating System.* Retrieved from https://www.apra.gov.au/sites/default/files/2018-02-pairs-guide-ud-external_1.pdf

Baldwin, R., & Black, J. (2016). Driving priorities in risk-based regulation: what’s the problem? *Journal of Law & Society*, 565-595.

Black, J., & Robert, B. (2012). *When risk-based regulation aims low: approaches and challenges.* Retrieved from LSE Ressearch Online: http://eprints.lse.ac.uk/43339/

Bradley, P. (2019, April 11). *Risk management standards and the active management of malicious intent in artificial superintelligence*. Retrieved from SpringerLink: https://link.springer.com/article/10.1007/s00146-019-00890-2

Brundage, M. (2018, February). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation.* Retrieved from Electronic Frontier Foundation: https://www.eff.org/files/2018/02/20/malicious_ai_report_final.pdf

Buiten, M. C. (2019, April 29). *Towards Intelligent Regulation of Artificial Intelligence*. Retrieved from European Journal of Risk Regulation: https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/towards-intelligent-regulation-of-artificial-intelligence/AF1AD1940B70DB88D2B24202EE933F1B

Cheatham, B., Javanmardian, K., & Samandari, H. (2019, April 26). Retrieved from mckinsey.com: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence

Commission Nationale Informatique & Libertés. (2018). *Privacy Impact Assessment Methodology*. Retrieved from https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-1-en-methodology.pdf

Commission Nationale Informatique & Libertés. (2012). *Methodology for Privacy Risk Management*. Retrieved from CNIL: https://www.cnil.fr/sites/default/files/typo/document/CNIL-ManagingPrivacyRisks-Methodology.pdfCouncil of Europe. (2013). Guidance on Risk-based Supervision and Risk Assessments. *(Project Against Money Laundering and Terrorist Financing in Serbia)*. Retrieved from https://rm.coe.int/16806de43c

Council of Europe. (2019). Retrieved from https://rm.coe.int/responsability-and-ai-en/168097d9c5

Deamer, K. (2016, July 01). What the First Driverless Car Fatality Means for Self-Driving Tech. *Scientific American*. Retrieved from https://www.scientificamerican.com/article/what-the-first-driverless-car-fatality-means-for-self-driving-tech/

Department of Industry, Science, Energy and Resources, Australian Government. (2019). *AI Ethics Principles*. Retrieved from Department of Industry, Science, Energy and Resources, Australian Government: https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles

Director of National Intelligence, U.S.A. (2020). *Principles of Artificial Intelligence Ethics for theIntelligence Community*. Retrieved from Director of National Intelligence: https://www.dni.gov/files/ODNI/documents/Principles_of_AI_Ethics_for_the_Intelligence_Community.pdf

Dvara Research. (2018, February). *THE DATA PROTECTION BILL, 2018*. Retrieved from Dvara Research Blog: https://www.dvara.com/blog/wp-content/uploads/2018/02/Data-Protection-Bill-Draft-Dvara-Research.pdf

Elish, M. (2019, May 15). *When Humans Attack*. Retrieved from Data and Society: https://points.datasociety.net/when-humans-attack-re-thinking-safety-security-and-ai-b7a15506a115

European Commission. (2019). *Ethics Guidelines for Trustworthy AI*. Retrieved from European Commission: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

European Commission for the Efficiency of Justice. (2018, December). *European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment.* Retrieved from Council of Europe: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c

European Parliament. (2020). *Report with Recommendations to the Commission on a Civil Liability Regime for Artificial Intelligence.* Retrieved from https://www.europarl.europa.eu/doceo/document/A-9-2020-0178_EN.pdf

Federal Deposit Insurance Corporation. (2019, June). *Consumer Compliance Examinations - Evaluating Impact of Consumer Harm*. Retrieved from Federal Deposit Insurance Corporation: https://www.fdic.gov/resources/supervision-and-examinations/consumer-compliance-examination-manual/documents/2/ii-2-1.pdf

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., & Srikumar, M. (2020, January 15). Principles Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. *Berkman Klein Center for Internet and Society Research Publication Series*. Retrieved from Berk: https://dash.harvard.edu/bitstream/handle/1/42160420/HLS%20White%20Paper%20Final_v3.pdf?sequence=1&isAllowed=y

G20. (2019). *G20 Ministerial Statement on Trade and Digital Economy*. Retrieved from OECD: https://www.mofa.go.jp/files/000486596.pdf

Haddal, H. N., & Frazar, S. (2018). *AUTONOMOUS SYSTEMS, ARTIFICIAL INTELLIGENCE AND SAFEGUARDS*. Retrieved from U.S. Department of Energy: Office of Scientific and Technical Information: https://www.osti.gov/servlets/purl/1561151

Health Insurance Portability & Accountability Act Collaborative of Wisconsin. (n.d.). *Risk Toolkit*. Retrieved January 15, 2021, from HIPAA: https://hipaacow.org/resources/hipaa-cow-documents/risk-toolkit/

Heijden, J. (2019, June 20). *Risk Governance and Risk-Based Regulation: A Review of the International Academic Literature*. Retrieved from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3406998

Henriques-Gomes, L. (2020, November 20). *'Robodebt-related trauma:' the victims still paying for Australia's unlawful welfare crackdown*. Retrieved from The Guardian: https://www.theguardian.com/australia-news/2020/nov/21/robodebt-related-trauma-the-victims-still-paying-for-australias-unlawful-welfare-crackdown

High-level Expert Group on Artificial Intelligence. (2019, April 8). *Ethics Guidelines for Trustworthy AI.* Retrieved from European Commission: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

High-Level Expert Group on Artificial Intelligence. (2020, July). *The Assessment List for Trustworthy Artificial Intelligence for self assessment.* Retrieved from European Commission: https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

IEEE. (2019). *Ethically Aligned Design: : A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*. Retrieved from IEEE: https://ethicsinaction.ieee.org/

Information Commissioner's Office & The Alan Turing Institute. (n.a.). *Summary of the tasks to undertake* . Retrieved from Information Commissioner's Office: https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence/part-2-explaining-ai-in-practice/summary-of-the-tasks-to-undertake/

Information Commissioner's Office. (2020). *Sample DPIA template*. Retrieved from General Data Protection Regulation: https://gdpr.eu/wp-content/uploads/2019/03/dpia-template-v1.pdf

Ingold, D., & Soper, S. (2016, April 21). *Amazon Doesn't Consider the Race of Its Customers. Should it?* Retrieved from Bloomberg: https://www.bloomberg.com/graphics/2016-amazon-same-day/

International Committee of the Red Cross. (2019, August 20). *Autonomy, artificial intelligence and robotics: Technical aspects of human control*. Retrieved from International Committee of the Red Cross: https://www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control

International Telecommunication Union . (2018, September). Retrieved from https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-ISSUEPAPER-2018-1-PDF-E.pdf

Kerry, C. (2020, February 10). Protecting Privacy in an AI-driven World. The Brookings Institution. Retrieved from https://www.brookings.edu/research/protecting-privacy-in-an-ai-driven-world/

Kingston, J. (2016). Artificial Intelligence and Legal Liability. In M. Bramer, & M. Petridis (Eds.), *Research and Development in Intelligent Systems XXXIII.* Springer. Retrieved from https://arxiv.org/abs/1802.07782

Leslie, D. (2019, June). *Understanding artificial intelligence ethics and safety*. Retrieved from The Alan Turing Institute: https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

Lines, R. (2018, April 14). Retrieved from https://www.linkedin.com/pulse/risk-impact-scale-vs-achievement-objectives-roger-lines/

Marda, V. (2018, September 10). *Artificial Intelligence Policy in India: A Framework for Engaging the Limits of Data-Driven Decision-Making*. Retrieved from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3240384

National Institute of Standards and Technology. (2019, December 19). NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software. Retrieved from https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software

NITI Aayog. (2018, June). *National Strategy for Artificial Intelligence*. Retrieved from NITI Aayog: https://niti.gov.in/sites/default/files/2019-01/NationalStrategy-for-AI-Discussion-Paper.pdf

NITI Aayog. (2020a, July). *Working Document: Towards Responsible #AIforAll*. Retrieved from NITI Aayog: https://niti.gov.in/sites/default/files/2020-11/Towards_Responsible_AIforAll_Part1.pdf

NITI Aayog. (2020b, November). *Working Document: Enforcement Mechanisms for Responsible #AIforAll*. Retrieved from NITI Aayog: https://niti.gov.in/sites/default/files/2020-11/Towards-Responsible-AI-Enforcement-of-Principles.pdf

OECD. (2019). *Artificial Intelligence in Society*. Retrieved from European Commission: https://ec.europa.eu/jrc/communities/sites/jrccties/files/eedfee77-en.pdf

Perez, S. (2016, March 24). Microsoft Silences its New AI Bot Tay, after Twitter Users Teach it Racism. *TechCrunch*. Retrieved from https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/

Perrow, C. (1999). *Normal Accidents: Living with High Risk Technologies*. Princeton University Press.

Prasad, S. (2019, March). *Defining "Harm" in the digital ecosystem*. Retrieved from Dvara Research : https://www.dvara.com/blog/2019/05/06/defining-harm-in-the-digital-ecosystem/

Romine, C. (2018, June 05). Fundamental and Applied Research and Standards for AI. National Institute of Standards and Technology, US Department of Commerce. Retrieved from https://www.nist.gov/system/files/documents/2018/06/04/iii-c_cr_ai_itl.pdf

ScienceDirect. (2016). *Risk Matrix*. Retrieved from ScienceDirect: https://www.sciencedirect.com/topics/engineering/risk-matrix

Singh, A., & Prasad, S. (2020, April 13). *Artificial Intelligence in Digital Credit in India*. Retrieved from Dvara Research Blog: https://www.dvara.com/blog/2020/04/13/artificial-intelligence-in-digital-credit-in-india/

Smart Dubai. (2019). *Artificial Intelligence Principles and Ethics*. Retrieved from https://www.smartdubai.ae/initiatives/ai-principles-ethics

Stanford Encyclopedia of Philosophy. (2020, April 30). *Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/entries/ethics-ai/#BiasDeciSyst

State of New South Wales. (2016). *Guidance for Regulators to Implement Outcomes and Risk-based Regulation*. Department of Finance, Services and Innovation. Retrieved from http://productivity.nsw.gov.au/sites/default/files/2018-05/Guidance_for_regulators_to_implement_outcomes_and_risk-based_regulation-October_2016.pdf

The Alan Turing Institute. (n.a.). *A right to explanation*. Retrieved from The Alan Turing Institute: https://www.turing.ac.uk/research/impact-stories/a-right-to-explanation

The Federal Government, Germany. (2018). *Artificial Intelligence Strategy*. Retrieved from Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, the Federal Ministry of Labour and Social Affairs: https://www.ki-strategie-deutschland.de/home.html

The Federal Government, Germany. (2018). *Artificial Intelligence Strategy* . Retrieved from Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, the Federal Ministry of Labour and Social Affairs: https://www.ki-strategie-deutschland.de/home.html

The Institute for Ethical AI & Machine Learning. (n.a.). *The Responsible Machine Learning Principles*. Retrieved from The Institute for Ethical AI & Machine Learning: https://ethical.institute/principles.html

The Public Voice Coalition. (2018). *Universal Guidelines for Artificial Intelligence*. Retrieved from The Public Voice Coalition: https://thepublicvoice.org/ai-universal-guidelines/

The Royal Society. (2019, November). *Explainable AI: the basics*. Retrieved from The Royal Society: https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf

Think 20. (2018). *Future of Work and Education for the Digital Age*. Retrieved from G20 Insights: https://www.g20-insights.org/wp-content/uploads/2018/07/TF1-1-11-Policy-Briefs_T20ARG_Towards-a-G20-Framework-For-Artificial-Intelligence-in-the-Workplace.pdf

Toh, A. (2020, September 29). *Automated Hardship: How the Tech-Driven Overhaul of the UK's Social Security System worsens poverty*. Retrieved from Human Rights Watch: https://www.hrw.org/report/2020/09/29/automated-hardship/how-tech-driven-overhaul-uks-social-security-system-worsens

U.S. Department of Defense. (2020, February). *DOD Adopts Ethical Principles for Artificial Intelligence* . Retrieved from U.S. Department of Defense: https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/

UK House of Lords, Select Committee on Artificial Intelligence. (2018, April 16). *Report of Session 2017–19: AI in the UK: ready, willing and able?* Retrieved from Publications of the UK Parliament: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

UN General Assembly. (2015). *Universal Declaration of Human Rights, 1947*. Retrieved from United Nations: https://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf

United Nations Economic Commission for Europe. (2012). Risk Managment in Regulatory Frameworks: Towards a Better Management of Risks. Retrieved from https://unece.org/fileadmin/DAM/trade/Publications/WP6_ECE_TRADE_390.pdf

University of Montreal. (2017). *Montreal Declaration for a responsbile development of artificial intelligence*. Retrieved from Montreal Declaraton Responsible AI: https://www.montrealdeclaration-responsibleai.com/the-declaration

Victoria Legal Aid. (2020, June 10). *Robo-debts*. Retrieved from Victoria Legal Aid: https://www.legalaid.vic.gov.au/find-legal-answers/centrelink/robo-debts

Vigdor, N. (2019, November 10). *Apple Card Investigated After Gender Discrimination Complaints*. Retrieved from New York Times: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

Villani, C. (2018). *For a Meaningful Artificial Intelligence Towards a French and European Strategy*. Retrieved from AI for Humanity: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

Vought, R. T. (2020, January). *Guidance for Regulation of Artificial Intelligence Applications*. Retrieved from Federal Register: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiEtN_6qJ7uAhVn_XMBHVuCCAgQFjAAegQIAhAC&url=https%3A%2F%2Fcstools.asme.org%2Fcsconnect%2FFileUpload.cfm%3FView%3Dyes%26ID%3D60580&usg=AOvVaw16bYadVJNfaduH7SsxpCkx

Walch, K. (2020, May 30). *The Autonomous Systems Pattern Of AI*. Retrieved from Forbes: https://www.forbes.com/sites/cognitiveworld/2020/05/30/the-autonomous-systems-pattern-of-ai/?sh=77e9551b6a6b