Building a Responsible Future.

Incorporating AI Regulations in AI Systems

AUTHORS

Ivana Bartoletti, Anindito De, Sanghamitra Singi,
Dipojjwal Ghosh & Siladitya Sen

Volume I

Contents

Introduction

Artificial intelligence is revolutionizing all industries, enhancing productivity, efficiency, and innovation. AI will contribute to economic improvement by streamlining processes, reducing costs, and driving business growth. However, it is crucial to acknowledge that, with the potential for good, AI brings risks as well. For example, biased AI systems may absorb existing inequalities and perpetuate them. We have seen how algorithmic decision-making may lock people out of services and rights. Recent developments in Generative AI have introduced additional security and disinformation risks. Governments worldwide are grappling with the challenge of how to govern AI systems.

The purpose of this paper is twofold. First, we will examine the AI-related policies, regulations, and guidelines emerging in major economies and distill some common denominators. Organizations need to understand AI regulations before implementing their governance plans. Second, we will delve into technical solutions that can translate the legal and normative requirements into computation and AI operations.

AI and the law

AI does not emerge into a legal vacuum. Many existing regulations, from privacy to consumer, human rights, and liability law, already implicitly apply to AI systems. It would be misleading to assert that AI is unregulated. In fact, over recent years, privacy and data protection laws are playing a key role in upholding people's rights in confronting automated decision-making and black box systems, as well as pervasive facial recognition technology. One could argue that privacy and data protection regulators have become de facto AI regulators.

However, AI is more complex than privacy law because AI systems may not involve personal data in the first place (and yet have an impact on people). Also, AI systems profoundly affect the labor market, transforming how we make decisions, analyze situations, and coexist with machines.

Governments and international organizations are deploying new sets of tools to rein in the risks of AI while harnessing its potential. It could be argued that regulating AI is not the target here: What matters is to regulate the behavior of people, businesses, and industries around AI – ensuring that AI-enabled products are safe, robust, and secure before they hit the market.

> “Although these tools are novel, they are not exempt from existing rules, and the FTC will vigorously enforce the laws we are charged with administering, even in this new market.”
>
> Lina Khan, Chairwoman,
> Federal Trade Commission [1]

Overview of global AI legislation

European Union

The EU Artificial Intelligence Act was approved by European Parliament 13 March 2024 and will take effect twenty days after its publication in the Official Journal. The AI Act is a product-based legislation and controls are wrapped around AI products based on the risks they may pose to society and individuals. The Act takes a horizontal approach and applies regardless of sector. AI systems are classified based on the risks they may pose to safety, health, and the fundamental rights of citizens.

Subliminal manipulation techniques or biometric categorisation are prohibited practices

AI products which are safety components, or the AI itself is a product, of an item which is covered by the EU harmonisation legislation, are high risks. This is the case, for example, of aviation or medical equipment.

AI products that when used in specific context may cause detriment individuals´ fundamental rights. For example, AI systems used in employment or benefit allocation become high risk, as an error could lock people out of opportunities, rights, and dignity.

High-risk AI systems must perform a conformity assessment and draw up technical documentation. The assessment requires organizations to demonstrate compliance related to transparency, human oversight, safety, sound data governance, privacy, fairness, and non-discrimination. General purpose AI (GPAI) – meaning AI systems with broad use cases, including many use cases that cannot be anticipated by the developer – is governed through a separate regime, with its own framework of systemic risks. GPAI that carries systemic risks is subject to tighter accountability controls.

Prohibitions on AI with "unacceptable" levels of risk will kick off six months after the act is published. It will be a full year before the rules around how GPAI take effect, and another two years after that before all rules of the act and obligations for high-risk systems apply.

United Kingdom

The UK government has initiated an AI-focused consultation process, aiming to establish a framework for regulating AI that is proportionate, future-proof, and supportive of innovation. Existing regulators will be tasked with interpreting and implementing the core AI principles of safety, transparency, fairness, accountability, and contestability. The UK government's minister for AI and intellectual property has stated that there are no immediate plans to pass any laws regarding AI regulation. [2]

United States

The US government is opting for a sector-specific approach, characterized by the Biden administration’s Executive Order on Safe, Secure, and Trustworthy Development and Use of AI, published in October 2023. Under this initiative, AI developers must evaluate and inform about any potential threats to national security posed by their algorithms. Notable aspects of this order encompass the creation of safety standards for AI, prioritizing consumer privacy, addressing discriminatory AI algorithms to promote equity and civil rights, and implementing measures to shield consumers from potential harm arising from AI-related healthcare practices.

The 2023 legislative session has seen a surge in state AI laws proposed across the U.S., surpassing the number of AI laws proposed or passed in prior legislative sessions. Ten states included AI regulations as part of larger consumer privacy laws passed or going into effect in 2023, and even more states have proposed similar bills. Several states proposed task forces to investigate AI, and others expressed concern about AI’s impact on services like healthcare, insurance, and employment. The US has continued to grapple with the intersection of AI and privacy laws, with states like California implementing the California Consumer Privacy Act (CCPA) and passing the California Privacy Rights Act (CPRA) to regulate the collection and use of personal data, including data used in AI systems.

On April 29, 2024, the NIST (National Institute of Standards and Technology) released a slew of new draft documents on artificial intelligence guidance and deployment, spanning topics from synthetic content risks to international standards development. NIST released four draft publications intended to help improve the safety, security and trustworthiness of artificial intelligence (AI) systems in support of President Biden's Executive Order. NIST's four new documents are, (a) the AI RMF Generative AI Profile, (b) Secure Software Development Practices for Generative AI and Dual-Use Foundation Models, (c) Reducing Risks Posed by Synthetic Content, and (d) A Plan for Global Engagement on AI Standards.

NIST released a draft publication based on the AI Risk Management Framework (AI RMF) to help manage the risk of Generative AI. The first, the AI RMF Generative AI Profile works to guide organizations to identify risks generative AI softwares can pose in their digital networks and helps create a set of actions relatively tailored to individual organizations' needs.

Canada

Canada's proposed Artificial Intelligence and Data Act (AIDA), introduced as part of the Digital Charter Implementation Act of 2022, would set the foundation for the responsible design, development, and deployment of AI systems. [5, 6] The Act would seek to ensure that AI systems deployed in Canada are safe and non-discriminatory and would hold businesses accountable for how they develop and use these technologies. Canada's regulatory framework for AI focuses on issues such as bias and discrimination, privacy protection, and accountability. Canada has also been addressing issues related to data governance in the context of AI, including data ownership, consent, and access.

China

On July 13, 2023, China's government finalized regulations for Generative AI, known as the Interim Measures for the "Management of Generative Artificial Intelligence Services." The objective is to establish guidelines to regulate Generative AI by making developers accountable for any harm caused by the AI, and also making GenAI equitable by imbibing core values of socialism. [2] While a comprehensive AI law is not yet in place, existing laws address certain aspects of AI development, deployment, and use.

Brazil

Brazil created a group of experts to prepare a proposal to regulate artificial intelligence in March 2022. The final draft of the AI law was published on December 6, 2022. [3] This draft aims to establish principles, rules, guidelines, and foundations to regulate the development and application of artificial intelligence in the country. Aligned with European Union guidelines, this project is based on similar principles, emphasizing concepts such as inclusive growth, sustainable development, transparency and explainability, robustness, and safety. This project provides guidelines for categorizing AI based on the potential risks AI products pose to society, requiring AI developers to perform risk assessments before introducing their AI products to the market. AI systems belonging to the "Highest" risk category are strictly prohibited, and developers are held accountable for any damages caused by AI. The law focuses on empowering users through notifications about AI usage and provisions to challenge the decisions of AI. The Legislative process to convert the draft into law began in May 2023 and is being discussed by Congress. The Final working needs to be debated and approved by Congress before it is converted into law. [4]

Comparative Analysis: EU AI Act and US Executive Order (EO)

For many organizations, distinctions between the EU and US approaches to AI regulation will be particularly germane.

	EU AI ACT	US EO
Regulatory Reach	**Risk-Based Approach:** The EU AI Act is a product-based legislation. High-risk AI systems are those used in products falling under the EU's product safety legislation and those for the purpose listed in Annex III of the EU AI Act. AI products can pose unacceptable risks. In that case, they need to be banned. Lighter controls are wrapped around other AI use cases that present limited risks.	**Sector-Based Approach:** Common standards, guidelines, practices, and rules for AI in various sectors should be implemented soon. AI risk management follows a sector-specific framework involving federal agencies. The U.S. Executive Order requires AI developers to share safety test results and related information with the US government.
Standards & Guidelines vs. Binding Regulation	The EU AI Act will be enforceable within 6 / 12 - 24 months starting formal adoption of the EU AI Act. Requirements around prohibited AI will be applicable in six months while GPAI requirements in 12 months and high risk and transparency requirements in 24 months. Breaches can lead to fines and penalties.	The EO focuses on standards and guidelines, mandating NIST and the Department of Commerce to create two sets of guidelines for AI systems by end of July 2024. The first set focuses on safe and secure development; the second outlines standards for red-teaming tests. Privacy-enhancing technology guidelines for agencies must be issued by October 2024.

How to Comply	1. Conduct a risk assessment to evaluate risk level. 2. Comply with the EU AI Act through self-assessment or third-party evaluation. 3. Maintain technical documentation and records. 4. Provide transparency and disclosure about AI systems: a. Remove unacceptable risk systems from the market. b. Register high-risk systems on the EU database before market placement. c. Comply with the Transparency requirements. Inform for limited-risk systems. Disclose and label AI-generated content.	The Responsible Use and Development of Generative AI policy follows the NIST AI RMF. The NIST AI Risk Management Framework is a valuable reference in constructing AI governance and conducting an AI impact assessment for high-risk assets.
Foundation Models/General Purpose AI systems	GPAI systems— such as models behind the viral boom in Generative AI tools like OpenAI's ChatGPT — are regulated under a separate regime that includes a systemic risk tier. The trigger for high-risk rules to apply to Generative AI technologies is determined by an initial threshold set in the law. At this present time, GPAI carries systemic risks when the cumulative amount of computation used for its training measured in floating point operations is greater than 10(^25), although this may change overtime in view of technology advancements.	The EO focuses on "dual-use foundation models" with exceptional capabilities that may threaten security, the economy, health, or safety. This includes cyber risks, CBRN weapons, deception, and manipulation. AI developers must assess these using red-teaming testing standards from the NIST and report to the government. The Department of Homeland Security will establish an AI Safety & Security Board to assess these results for critical infrastructure sectors.
System Testing & Monitoring	Businesses are required by the EU AI Act to ensure adherence to regulations involving thorough pre-market testing procedures, the methods and benchmarks employed for testing before launch, and a post-market surveillance approach focusing on the developer continuously monitoring the system's performance.	The EO emphasizes assessing and monitoring AI systems to ensure they function as intended, are secure and ethical, and comply with laws. It highlights the need for infrastructure to evaluate and supervise AI-enabled healthcare technology. In the pre-market testing phase, organizations must submit evidence of safety and effectiveness to the FDA. After deployment, organizations must adhere to post-market requirements, including tracking and reporting malfunctions.
Compliance With International Standards/ Cybersecurity Standards	1. Adhering to cybersecurity standards, like promoting "security by design," is mandatory. 2. The EU AI Act mandates EU standards setting organisations to develop the appropriate standards. When the standards will be published, organisations that will adhere to them will be lifted from the obligations to compile a conformity assessment.	1. Adherence to NIST cybersecurity standards, including "security by design," is mandatory. 2. The EO aims to prevent malicious international cyber entities from misusing advanced AI models, safeguarding them from exploitation by foreign cyber threats.

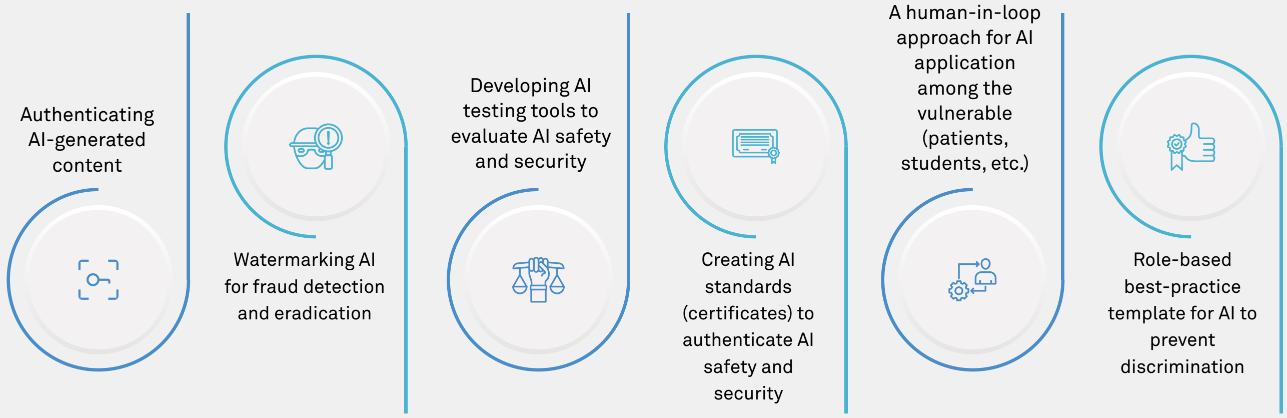
Penalties	• Up to 7% of global annual turnover or €35M for prohibited AI violations. • Up to 3% of global annual turnover or €15M for most other violations. • Up to 1.5% of global annual turnover or €7.5M for supplying incorrect information, capping fines for SMEs and start-ups.	The Executive Order lacks specific penalties for violations.

A deep dive into the US Executive Order

The EO is a directive issued by the President of the United States that applies to the federal government's operations. The main goal is a unified approach to AI use across different departments. In the realm of AI governance, the EO in the USA stands out as a significant step toward shaping the future of AI policies. It underscores the government's commitment to fostering innovation while addressing potential risks associated with AI deployment. The order emphasizes the importance of transparency, public trust, and accountability in developing and using AI technologies. Some key points of interest for AI governance (from the US's perspective) include establishing safety standards for AI, a focus on protecting consumer privacy, efforts to advance equity and civil rights by curbing discriminatory AI algorithms, and initiatives to shield consumers from potential harm caused by AI-related healthcare practices. The EO considers a wide range of risk factors, including economic, environmental, cybersecurity, and national security risks, among many others. Moreover, the EO weighs various facets of ethics – equity, privacy, safety, etc. The EO emphasizes equity, highlighting the importance of providing equal opportunities and fair treatment to each individual. These orders tackle systemic inequalities and foster inclusivity in various domains, including education, employment, and housing. The EO aims to protect citizens from the potential risks related to misuse of personal data by AI systems. To accomplish this goal, the EO drives enterprises to fortify their AI systems with different privacy-preserving strategies.

For safety, the EO addresses the potential hazards or malfunctions in AI systems, emphasizing the importance of establishing clear safety standards to minimize risks to users and the public.

The EO suggests several avenues for enterprises to mitigate and govern AI risks:



Authenticating AI-generated content

Watermarking AI for fraud detection and eradication

Developing AI testing tools to evaluate AI safety and security

Creating AI standards (certificates) to authenticate AI safety and security

A human-in-loop approach for AI application among the vulnerable (patients, students, etc.)

Role-based best-practice template for AI to prevent discrimination

One of the greatest objectives of the EO is to prioritize the protection and support of vulnerable individuals and groups like medical patients through responsible usage of AI. The objective is to guarantee that AI technologies positively impact patient care and safety in the healthcare sector. To achieve this, the federal government is developing programs to evaluate and track the influence of AI. Furthermore, the government has is creating released/issued created resources and guidelines to aid healthcare professionals with responsible AI integration tools and solutions.
The government emphasizes the significance of ethical and effective implementation of AI in healthcare settings.

In the climate domain, to emphasize the importance of AI research in addressing and mitigating climate change, the government aims to leverage AI advancements to develop innovative solutions that contribute to environmental sustainability and resilience.

Furthermore, the EO aims to devise best practices for inhibiting AI algorithms from perpetuating biases and intensifying discrimination across different domains, such as housing and the justice system.
By highlighting the significance of impartial and fair AI implementation, the EO emphasizes the necessity of creating frameworks that foster inclusivity and address discriminatory practices that may emerge from AI technologies. It aims to guarantee that AI systems are developed and executed in a way that upholds fundamental principles of fairness, equity, and justice.

Summary of global AI legislation

By 2024, the impact of AI regulation on global technological advancements and the ethical use of AI is anticipated to be substantial. In addition to country-specific regulations, it is worth noting that international organizations such as the Council of Europe, the OECD, and others have already produced AI guidelines and conventions.

While each jurisdiction is forging ahead with its frameworks and strategies, they are also intensifying their collaborative endeavours to synchronize and harmonize their diverse approaches. This transition signifies a crucial milestone in guaranteeing that the swift progress of AI technology is in line with ethical norms, transparency, and the public's well-being.

How can organizations enable AI governance in the face of evolving legal requirements?

There is little doubt that we are only at the beginning of the discussion around AI governance and regulation. While countries are equipping themselves with strategies and legal solutions, all actors with a stake in AI innovation should seek clear alignment in some crucial areas, namely robustness, safety, privacy, non-discrimination, and a commitment to tackle deep fakes, especially in election contexts.

This alignment will be critical as global organizations seek common denominators that they can use to build global AI governance strategies. The key question that every data/AI and governance leader is asking is: How can we efficiently translate these global common denominators of effective AI governance into manageable and accountable computation and technical strategies?

In other words, with existing, evolving, and upcoming regulations, how can companies devise an AI strategy accounting for the emerging alignment on legal/responsible AI requirements?

To answer this question, we suggest a model focused on developing AI systems based on two key concepts: responsible-by-design and responsible-in-design. Responsible-by-design emphasizes incorporating ethical considerations into the design phase of any AI system. AI system developers need to communicate with various stakeholders and domain experts to understand the potential risks and impacts on consumers. The concept of responsible-in-design focuses on mitigating the risks identified in the responsible-by-design phase through multi-modal processes, human-in-loop, best practices, etc.

Key Concepts

Before defining and describing the proposed methodologies of responsible-by-design and responsible-in-design, we need to highlight some key concepts that will play a significant role in the later sections.

Risk Pillars

A responsible AI framework should identify and classify risks around impact areas. We term these areas risk pillars. We propose four main risk pillars that any AI system should consider: [5]

Individual

AI tools must prioritize privacy, security, and non-discrimination to uphold human dignity. Transparency and openness about data use and due diligence processes are essential to maintain trust and empower citizens to scrutinize AI, which will be crucial for AI advancement.

Social

The development of AI systems should prioritize enhancing society and reinforcing the shared values that bind us together as a cohesive unit. Thus, AI systems must promote inclusivity, equality, and ethical decision-making.

Technical

Ensuring the resilience and security of AI systems is of utmost importance, as they must withstand various challenges and potential threats. Moreover, these systems should prioritize safeguarding personal and corporate data, guaranteeing confidentiality and integrity.

Environmental

To achieve environmental sustainability in the field of artificial intelligence, it is imperative for compute-intensive AI systems to minimize their impact on the environment. This means that these systems need to find ways to reduce their carbon footprint and overall energy consumption. AI systems need to contribute to a more sustainable future.

Dimensions of Responsible AI

The concept of ethics and responsibility in AI was first outlined in the “Asilomar AI Principles” (2017) by a group of AI researchers, scientists, and ethicists [6, 7]. However, the first formal definition of Responsible AI was provided by Vincent C. Müller in his paper “Ethics of Artificial Intelligence and Robotics.” He defined “Responsible AI” as the ethical consideration and implementation of artificial intelligence systems. The paper emphasizes the need for AI development that aligns with societal values and addresses potential ethical challenges. Responsible AI, according to Müller, involves the careful examination and mitigation of biases, transparency in decision-making processes, accountability for the consequences of AI systems, and an overall commitment to human well-being [7]. According to Müller's proposition, any Responsible AI framework should encompass the following dimensions: "Fair," "Safe," “Robust,” and “Explainable.”

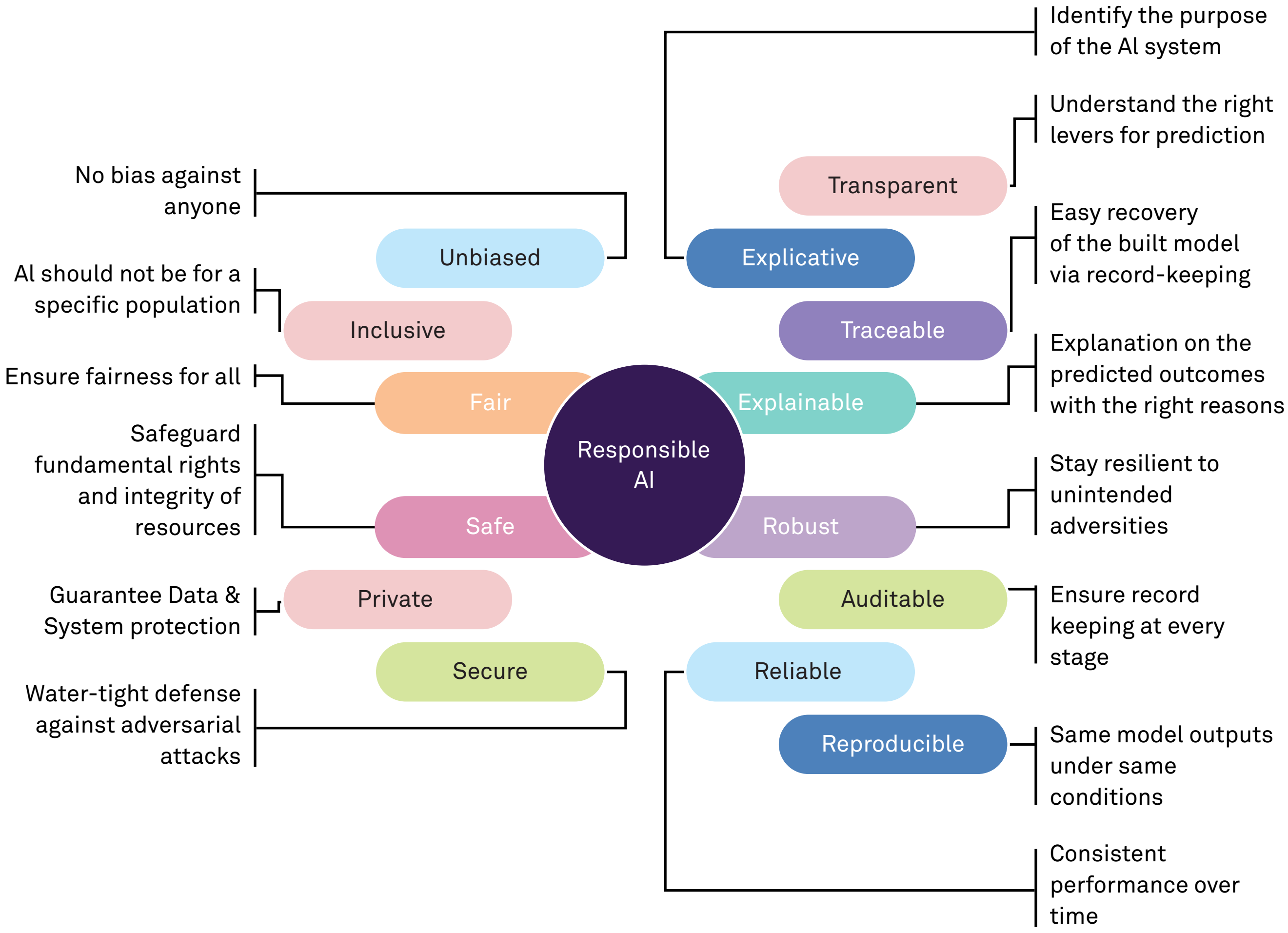

Responsible AI
Fair
Safe
Explainable
Robust
Unbiased
Inclusive
Private
Secure
Transparent
Explicative
Traceable
Auditable
Reliable
Reproducible
No bias against anyone
AI should not be for a specific population
Ensure fairness for all
Safeguard fundamental rights and integrity of resources
Guarantee Data & System protection
Water-tight defense against adversarial attacks
Identify the purpose of the AI system
Understand the right levers for prediction
Easy recovery of the built model via record-keeping
Explanation on the predicted outcomes with the right reasons
Stay resilient to unintended adversities
Ensure record keeping at every stage
Same model outputs under same conditions
Consistent performance over time

Figure 1: Dimensions of Responsible AI

These dimensions of responsible AI remain highly relevant, and they also capture many of the common denominators described in the previously discussed AI regulations. An illustrative example of this alignment can be observed in the convergence of the core principles highlighted in the European Union AI Act and the Executive Order issued by the Government of the United States (see Figure 2).

AI ACTS

DIMENSIONS	EU AI Act	Executive Order US
Fair	• Measures to address biases in training data to enhance faimess and prevent discrimination in AI system outcomes • Compliance with fundamental rights obligations	• Frameworks to be designed to tackle systemic inequalities and foster inclusivity in various domains • Role-based best-practice template for AI to prevent discrimination
Explainable	• AI systems to provide clear, comprehensible information about their functioning and purpose • Should feature user friendly interfaces to facilitate AI systems operations and enable them to make informed decision • High-risk AI system systems are required to provide explanations for the decision they make • Enabling users to understand the rationale behind the AI generated outcomes and ensuring accountability	• It mandates clear disclosure of AI system's functionality and decision-making processes • Encourages AI to provide explanations in a contextually relevant manner • Promotes the idea that AI systems should not only explain their decisions but also actively learn from the feedback
Robust	• Human in the loop in system for all deployments of AI system to ensure decisions align with ethical and legal standards • AI system must be designed and developed to withstand both intentional and unintentional attempts to manipulate them • High-risk AI system must incorporate fallback mechanisms to ensure they can respond in-case of failures	• Human-in-loop approach for application of AI among vulnerable • Authentication of AI generated contents • Watermarking of AI for fraud detection and eradication
Safe	• Identification & qualification of risk possess by AI system • A proper data governance need to be implemented to ensure the quality, integrity and security and reliability of the data used by AI systems • Emergency stop mechanism to enable quickly deactivation or overridden in case of bodies emergencies	• Proper Risk identification and mitigation framework need to be implemented • Development of AI testing tools to evaluate AI safety • Share the safety test results and other critical information with governing bodies • Creation of AI standards (certificates) to authenticate AI safety and security

Figure 2: Mapping AI acts with dimensions of Responsible AI

Responsible-by-Design: Guided approach to designing a Responsible AI system

As discussed earlier, responsible-by-design aims to proactively address issues like safety, privacy, and non-discrimination, and ensure that the AI systems and services are designed with the well-being of individuals and society in mind.

The first step toward incorporating responsible behavior in the AI system design phase is to identify the nature of risks that can be encountered according to the different risk pillars and the dimensions of responsible AI. This risk identification exercise should be carried out according to the regulatory standards set by the countries where the AI system would be operational. For example, the risk identification for countries within the European Union will be based on the conformity assessment criterion set by the European Union.

Once the nature of risks is discovered, the AI developers, business stakeholders, and domain experts need to gauge the exposure and potential effect of the identified risks on the consumers to provide a risk quantification.

Risk Identification

The significance of risk identification in AI cannot be overstated. It plays a crucial role in the safety and reliability of AI systems. Risk identification allows us to understand the impact on the population. This analysis helps us make informed decisions and take appropriate measures to minimize risks.

Enterprises should prioritize the identification of the domain in which the AI use case will be implemented. It is crucial to have a comprehensive understanding of the domain and the potential risks that may arise within it.

Once the domain has been determined, the next step is to assess risk pillars – Individual, Social, Technical, and Environmental. Organizations should utilize a reliable risk evaluation checklist (details in **Appendix**) during the early stages of discussions to assess the suitability of a use case. This checklist can also aid in gaining a deeper understanding of the relevant AI landscape. It will consist of questionnaires designed to identify the specific risks associated with an AI implementation. This process will involve a human-in-loop system with a scoring mechanism to categorize the risks into the risk pillars.

These risks should then be aligned with the Responsible AI dimensions such as Fairness, Explainability, Robustness, and Safety. It is essential to design specific test cases to determine the Responsible AI dimensions under which the identified risks will fall. This step is crucial as the mitigation strategy relies heavily on this information.

	Individual	Social	Technical	Environmental
Fair	Unbiased	Inclusive	Trustworthy	Equitable
Explainable	Transparency	Explicative	Traceable	Awareness
Robust	Safeguarding	Auditable	Reproducible	Sustainable
Safe	Privacy	Reliable	Secure	Controllable

Figure 3: Recipe to deliver Responsible AI values

This table is a guideline for creating risk mitigation recipes based on the cross-relationship between Responsible AI dimensions and Risk pillars defined earlier.

Risk Measurement

Risk measurement is essential for organizations as it allows them to assess and evaluate the identified risks. When measuring risks, two crucial criteria come into play: Coverage and Impact. These criteria help organizations determine how risk may affect their operations and the potential consequences on overall objectives.

Coverage

The number of stakeholders and users impacted by the risk is indicated. This metric primarily addresses the number of users/stakeholders of the system affected by the risk. Subsequently, the coverage is categorized as High, Medium, or Low based on the count of affected users.

Impact

This indicates the potential severity of the threat and the level of risk to an organization. Furthermore, it will discuss whether this risk is confined solely to the specific organization or if it can also affect society or the environment.
The impact is categorized as Severe, Moderate,or Low.

Responsible-in-Design: Enabling responsible behavior throughout the AI lifecycle

Responsible-in-design focuses on fostering responsible behavior at every stage of the AI lifecycle and making the AI system compliant with regulatory requirements. In the previous stages, the team of AI developers, business stakeholders, and domain experts have determined the nature and impact of the inherent risks of the proposed AI system. During this phase, the AI developers need to exercise best practice approaches to mitigate these identified risks and ensure compliance in the AI system.

Risk Mitigation

Much has been discussed about the methods that enterprises need to employ to identify and measure risk associated with an AI implementation. Following this, the subsequent crucial aspect is the mitigation of identified risks. To diminish these risks, the organization must implement appropriate strategies and safeguards in the identified areas.

To mitigate these risks, the organization should implement the right strategies and guardrails at the right places. The three fundamental levers of any organization — namely **People, Process,** and **Technology** — can be utilized to quell the identified risks.

People

People are crucial in mitigating the identified risks throughout various stages of the AI model development phase. During experimentation, a thorough evaluation of the business problem is required. Teams should determine if an AI solution is necessary and discuss the different approaches required to achieve the expected performance of the AI system. Risk consultants need to assess risks through conformity assessments. Deep-level data explorations should be conducted to uncover sensitive variables (like PIIs, race, gender, etc.) and proxies present in the data to ensure inclusivity in the trained AI system. Moreover, continuous human-in-loop monitoring is necessary to detect data and model drifts along with adversarial situations.

Process

Developing a best practice process template is necessary to create a robust Responsible AI framework and to identify and eradicate any bias that might have infiltrated the AI development process. Moreover, automated test cases designed on best practices can be used to gauge AI's performance and issue certificates such as the Responsible AI certificate, the Quality certificate, the Standardization certificate, etc. This certification process will ensure that the organization can implement AI in its system without undue risks.

Technology

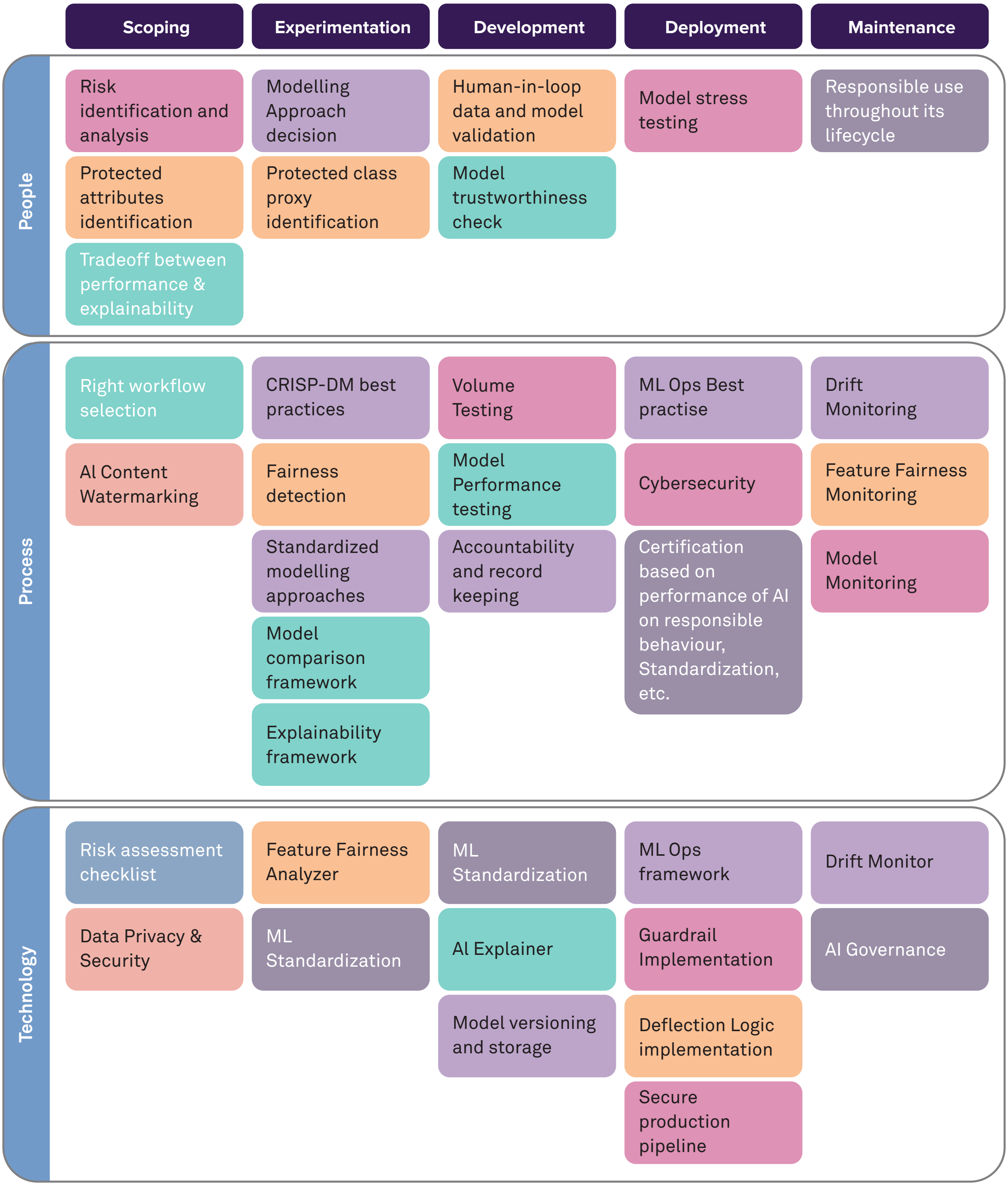
During development, developers must select the most effective technological templates to prevent the AI from learning from negative events or existing biases. Additionally, developers can establish domain and best practices-based ethical guardrails to ensure that the AI systems are restricted from perpetuating and amplifying biases and unfairness in the current data. These guardrails will encompass three key aspects: Domain, Safety, and Security.

Domain-based guardrails will establish boundaries and standards for AI applications within a specific domain, ensuring they operate within defined limits and adhere to the ethical principles of AI. This will make the AI accountable and fair.

Implementing **safety-based guardrails** will enable the AI system to proactively identify and mitigate risks, ensuring the reliability and trustworthiness of the AI system.

On the other hand, **security-based guardrails** will ensure that the AI system adheres to security protocols, safeguarding sensitive data and preventing potential breaches.

These actions make the AI systems “Responsible” by acting in conformance with the four different dimensions of Responsible AI through direct or mixed influences as seen in Figure 4.



	Scoping	Experimentation	Development	Deployment	Maintenance
People	🟥 Risk identification and analysis 🟧 Protected attributes identification 🟩 Tradeoff between performance & explainability	🟪 Modelling Approach decision 🟧 Protected class proxy identification	🟧 Human-in-loop data and model validation 🟩 Model trustworthiness check	🟥 Model stress testing	⬛ Responsible use throughout its lifecycle
Process	🟩 Right workflow selection AI Content Watermarking	🟪 CRISP-DM best practices 🟧 Fairness detection 🟪 Standardized modelling approaches 🟩 Model comparison framework 🟩 Explainability framework	🟥 Volume Testing 🟩 Model Performance testing 🟪 Accountability and record keeping	🟪 ML Ops Best practise 🟥 Cybersecurity ⬛ Certification based on performance of AI on responsible behaviour, Standardization, etc.	🟪 Drift Monitoring 🟧 Feature Fairness Monitoring 🟥 Model Monitoring
Technology	Risk assessment checklist Data Privacy & Security	🟧 Feature Fairness Analyzer ⬛ ML Standardization	⬛ ML Standardization 🟩 AI Explainer 🟪 Model versioning and storage	🟪 ML Ops framework 🟥 Guardrail Implementation 🟧 Deflection Logic implementation 🟥 Secure production pipeline	🟪 Drift Monitor ⬛ AI Governance

🟧 Fair

🟪 Robust

🟥 Safe

🟩 Explainable

⬛ Fair, Safe, Explainable & Robust

Figure 4: Impact of "Responsible-In-Design" actions with respect to Dimensions of Responsible AI

Illustration of the mechanism

To illustrate these concepts above, let's delve into a practical example. A healthcare service provider employs an AI-powered recommender system to help doctors offer treatment and medication for different health conditions. A panel of AI consultants was set up to evaluate, estimate, and mitigate risks that might be present in the AI recommendation system to ensure compliance with the relevant country's AI regulations.

The AI consultants understood that they were dealing with a business problem in the healthcare domain and used the questionnaires in the Risk Identification Checklist to understand the potential risks associated with the healthcare domain. During this time, it was discovered that a specific group of individuals was receiving insufficient pain management treatments or medications. Based on this information, the risk scoring system and the human experts categorized that the issue belongs to the risk pillars of Individual and Societal since the observed discrepancy is a direct consequence of both underserving individual patients and contributing to bias on a population level. By conducting test cases, it was determined that individuals belonging to a particular racial profile are being prescribed lower doses of pain medication compared to the recommended dosage outlined in the healthcare protocol established by central authorities. This discrepancy highlights the need to address potential biases and ensure unbiased and inclusive healthcare practices for all patients [8].

The identified risk of this scenario will look like:

Domain	**Healthcare**
Risk Pillar	**Individual, Social**
RAI Dimension	**Fair**
RAI Values	**Unbiased, Inclusive**

After identifying the risk, it is crucial to assess the extent and scope of the risk and develop effective strategies to mitigate it. The primary goal should be to ensure fair and inclusive healthcare practices for all patients. This incident has a medium coverage as it affects a certain segment of the population due to the flawed implementation of AI. The impact is high, as it can greatly affect the underserved population. Consequently, the utilization of this AI system in healthcare carries a higher level of risk, considering the potential impact on population health.

The classification will look like this:

Coverage	Medium
Impact	High
Measured Risk	High

Once the risk has been effectively identified and assessed, enterprises can minimizing the exposure to risk by implementing a set of meticulously crafted, standardized risk mitigation strategy templates developed by a team of AI experts. Enterprises must exercise caution when choosing datasets for retraining a flawed AI system. It is imperative to thoroughly examine the data and eliminate any possible sensitive information and proxy variables that may indicate the protected class, such as gender, racial profile, PIIs, and so on. Thorough feature imputation methods are necessary to minimize occurrences of extreme events and biases. This de-biased data can be employed to retrain the AI system. During the model development phase, it is crucial to exercise extreme caution and implement a rigorous feature selection strategy.

This is necessary to minimize the risk of bias and the inclusion of unwanted sensitive information in the model. Appropriate algorithms and approaches should be utilized to re-create the AI system, preserving the transparency and explainability of the recommendations with the inclusion of safety and security-based guardrails. These steps will prevent adverse results, making the revised AI system reliable and trustworthy. Automated test cases must be created to consistently oversee and regulate the real-time results, identifying any potential infiltration of new biases within the AI system. If an unjust outcome is detected, the medical practitioner will be promptly notified through pre-established alerts. Furthermore, the perpetual administration is used to issue certifications that validate the excellence of execution and commitment to the principles of Responsible AI.

Conclusion

The USA's Executive Order on AI is but one of many recent catalysts promoting responsible AI practices. The growing number of active and emerging AI regulations around the world underscores the significance of risk identification, measurement, and mitigation strategies that enterprises should adopt. By prioritizing these strategies, companies can effectively navigate the complex landscape of AI and contribute to developing a responsible and accountable AI framework. This approach safeguards enterprises against potential risks and fosters public trust and confidence in AI technologies.

References

1. L. Feiner, "FTC Chair Lina Khan says she's on alert for abusive A.I. use," CNBC, 3 May 2023. [Online]. Available: https://www.cnbc.com/2023/05/03/ftc-chair-lina-khan-says-shes-on-alert-for-abusive-ai-use.html. [Accessed 25 January 2024].

2. M. Klimentov, "From China to Brazil, here's how AI is regulated around the world," 3 September 2023. [Online]. Available: https://www.washingtonpost.com/world/2023/09/03/ai-regulation-law-china-israel-eu/. [Accessed 16 January 2024].

3. "Comissão De Juristas Responsável Por Subsidiar Elaboração De Substitutivo Sobre Inteligência Artificial No Brasil," 06 12 2022. [Online]. Available: https://legis.senado.leg.br/comissoes/mnas?codcol=2504&tp=4&_gl=1*1exxtct*_ga*MTk3NDE2NDE1LjE3MDY2MjA1NTk.*_ga_CW3ZH25XMK*MTcwNjYyMDU1OS4xLjAuMTcwNjYyMDU1OS4wLjAuMA... [Accessed 01 February 2024].

4. D. A. S. "Comissão conclui texto sobre regulação da inteligência artificial no Brasil," Senado Notícias, 06 12 2022. [Online]. Available: https://www12.senado.leg.br/noticias/materias/2022/12/06/comissao-conclui-texto-sobre-regulacao-da-inteligencia-artificial-no-brasil. [Accessed 01 February 2024].

5. G. o. C. "Innovation, Science and Economic Development Canada," Government of Canada, 18 August 2022. [Online]. Available: https://ised-isde.canada.ca/site/innovation-better-canada/en/canadas-digital-charter/bill-summary-digital-charter-implementation-act-2020. [Accessed 01 February 2024].

6. G. o. C. "Department of Justice," Government of Canada, 04 November 2022. [Online]. Available: https://www.justice.gc.ca/eng/csj-sjc/pl/charter-charte/c27_1.html. [Accessed 01 February 2024].

7. A. S. "Wipro Blogs," Wipro, [Online]. Available: https://www.wipro.com/ai/solutions/four-dimensions-of-responsible-ai/. [Accessed 07 February 2024].

8. "A Principled AI Discussion in Asilomar," 5-8 January 2017. [Online]. Available: https://futureoflife.org/principles/principled-ai-discussion-asilomar/. [Accessed 01 February 2024].

9. "AAIP: Asilomar AI Principles," 5-8 January 2017. [Online]. Available: ttps://futureoflife.org/ai-principles/. [Accessed 01 February 2024].

10. V. C. Müller, "Ethics of artificial intelligence and robotics," Stanford Encyclopaedia of Philosophy, 2020.

11. T. N. Santoro and J. D. Santoro, "Racial Bias in the US Opioid Epidemic: A Review of the History of Systemic Bias and Implications for Care," Cureus, vol. 10, no. 12, 14 December 2018.

Appendix

Risk identification checklist

The risk identification checklist should question the AI systems on its approach across the different dimensions of Responsible AI and assess its level of impact concerning the risk pillars. The checklist should probe how the AI application collects, processes, and secures PIIs and other sensitive information. Moreover, the checklist should address the quality of input data safety mechanisms for data leakage or poisoning issues. Furthermore, the checklist should identify any fallback options in case of AI malfunctions and evaluate the human experts' opinion on the AI system in terms of fundamental rights violation (like unethical surveillance or loss of autonomy).

High-risk AI systems

Here are some examples of high-risk AI systems as per the EU AI Act:

Internal or Third-Party CA according to the high-risk AI systems (AI System that belong to the use cases of Annex III of the EU AI Act)

Type of CA		
1. Biometrics (and biometrics-based systems)	If harmonized standards or common specifications have been applied	**Internal CA Or Third-Party CA**
	If the provider has not applied harmonized standards/ common specifications or has applied them only in part	**Third-Party CA**
2. Critical infrastructure 3. Education and vocational training 4. Employment, workers management. and access to self-employment 5. Access to and enjoyment of essential private services and public services and benefits 6. OW enforcement 7. Migration, asylum, and border control management 8. Administration of justice and democratic processes		**Internal CA** (The Commission may amend this rule and require third party CA, through delegated acts.)

Ambitions Realized.

Wipro Limited
Doddakannelli
Sarjapur Road
Bengaluru – 560 035
India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading technology services and consulting company focused on building innovative solutions that address clients' most complex digital transformation needs.

Leveraging our holistic portfolio of capabilities in consulting, design, engineering, and operations, we help clients realize their boldest ambitions and build future-ready, sustainable businesses. With over 230,000 employees and business partners across 65 countries, we deliver on the promise of helping our clients, colleagues, and communities thrive in an ever-changing world.

For additional information, visit us at **www.wipro.com**

●●●●● GEO/AGENCY CODE/VALIDITY