

# OECD Due Diligence Guidance for Responsible AI





# **OECD Due Diligence Guidance for Responsible AI**

This work was approved and declassified by the Digital Policy Committee and the Investment Committee on 26 January 2026.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

**Please cite this publication as:**

OECD (2026), *OECD Due Diligence Guidance for Responsible AI*, OECD Publishing, Paris, <https://doi.org/10.1787/41671712-en>.

ISBN 978-92-64-31703-1 (print)  
ISBN 978-92-64-31822-9 (PDF)  
ISBN 978-92-64-44865-0 (HTML)

**Photo credits:** Cover © Vladimir\_Timofeev/Getty Images.

Corrigenda to OECD publications may be found at: <https://www.oecd.org/en/publications/support/corrigenda.html>.

© OECD 2026



**Attribution 4.0 International (CC BY 4.0)**

This work is made available under the Creative Commons Attribution 4.0 International licence. By using this work, you accept to be bound by the terms of this licence (<https://creativecommons.org/licenses/by/4.0/>).

**Attribution** – you must cite the work.

**Translations** – you must cite the original work, identify changes to the original and add the following text: *In the event of any discrepancy between the original work and the translation, only the text of the original work should be considered valid.*

**Adaptations** – you must cite the original work and add the following text: *This is an adaptation of an original work by the OECD. The opinions expressed and arguments employed in this adaptation should not be reported as representing the official views of the OECD or of its Member countries.*

**Third-party material** – the licence does not apply to third-party material in the work. If using such material, you are responsible for obtaining permission from the third party and for any claims of infringement.

You must not use the OECD logo, visual identity or cover image without express permission or suggest the OECD endorses your use of the work.

Any dispute arising under this licence shall be settled by arbitration in accordance with the Permanent Court of Arbitration (PCA) Arbitration Rules 2012. The seat of arbitration shall be Paris (France). The number of arbitrators shall be one.

# Foreword

AI development has the potential to transform society in ways comparable to the industrial revolution or the advent of the internet. AI represents not merely an incremental advance but a transformative technology with the capacity to enhance productivity, create economic value, and solve complex challenges across a variety of sectors such as healthcare, manufacturing, logistics, and public administration. To harness this positive potential, the OECD sets out a balanced approach to responsible AI that enhances the opportunities of AI while addressing risks of adverse impacts.

Following the May 2024 adoption of the revised Recommendation of the Council on Artificial Intelligence (“AI Principles”), the OECD Council meeting at Ministerial level instructed the Digital Policy Committee (“DPC”) through its Working Party on AI Governance (“AIGO”) to “continue its important work on artificial intelligence building on this Recommendation” and “to develop and iterate further practical guidance on the implementation of this Recommendation.” The AI Principles recognise the positive potential of AI to support beneficial outcomes for people and the planet. The AI Principles promote an ecosystem for reliable AI systems by establishing principles and policy guidelines that foster innovation while addressing risks.

Further, the OECD Guidelines for Multinational Enterprises on Responsible Business Conduct (“MNE Guidelines”), updated in 2023, explicitly note that “technological innovation [has] driven productivity in all sectors, as well as the ability of enterprises to conduct due diligence and contribute to sustainable development”. The MNE Guidelines recognise the importance of realising “the economy-wide effects of technological progress, including productivity growth and job creation”. The MNE Guidelines also call on enterprises to carry out risk-based RBC due diligence with respect to actual and potential adverse impacts related to science, technology and innovation.

This guidance is intended to assist enterprises in implementing both the MNE Guidelines and the AI Principles. It serves as a tool for multinational enterprises engaged in the AI system value chain – those supplying inputs for AI development, actively participating in the AI system lifecycle, or utilizing AI systems in their operations, products, and services across all sectors.

This project is jointly overseen by the DPC through AIGO and the Investment Committee (“IC”) through the Working Party on Responsible Business Conduct (“WPRBC”). This guidance is based on the work of the OECD.AI Expert Group on Risk & Accountability. The Expert Group is made up of over 100 representatives from government, civil society, workers’ representatives, and large and small enterprises from across the AI supply chain. This guidance also benefitted significantly from review and feedback from the Civil Society Information Society Advisory (“CSISAC”), OECD Watch, the Trade Union Advisory Committee (“TUAC”) and Business at OECD (“BIAC”).

# Table of contents

Foreword	3
Executive summary	6
1 Introduction to RBC due diligence and key considerations for AI	7
Introduction to responsible AI	8
Purpose of this guidance	8
Target audience	9
Understanding the risks related to the development and use of AI	12
Basics of RBC due diligence	13
How to use this guidance	17
2 Due diligence framework and practical examples for identifying and addressing risks	18
Step 1 - Embed RBC into policies and management systems	19
Step 2 - Identify and assess actual and potential adverse impacts	23
Step 3 - Cease, prevent and mitigate adverse impacts	33
Step 4 - Track implementation and results of due diligence activities	46
Step 5 - Communicate actions to address impacts	47
Step 6 - Provide for or co-operate in remediation when appropriate	48
References	50
Glossary	53
Notes	57
<b>FIGURES</b>	
Figure 1.1. Graphical representation of the RBC due diligence framework	14
Figure 2.1. Due diligence expectations based on involvement with the adverse impact	30
<b>TABLES</b>	
Table 2.1. Step 1: Roadmap of related provisions in existing frameworks	19
Table 2.2. Step 2: Roadmap of related provisions in existing frameworks	23
Table 2.3. Factors to consider when prioritising risk	32

Table 2.4. Step 3: Roadmap of related provisions in existing frameworks	33
Table 2.5. Step 4: Roadmap of related provisions in existing frameworks	46
Table 2.6. Step 5: Roadmap of related provisions in existing frameworks	47
Table 2.7. Step 6: Roadmap of related provisions in existing frameworks	48

## BOXES

Box 1.1. Considerations for Small and Medium Sized Enterprises (SMEs)	11
Box 2.1. Examples of high-risk uses of AI systems drawn from different frameworks	25
Box 2.2. Understanding the risk-based approach	26
Box 2.3. Identifying risks to data quality, interoperability and access throughout the AI system lifecycle	27
Box 2.4. Understanding involvement with the risk	29
Box 2.5. Scenarios illustrating the involvement framework	31
Box 2.6. Tailoring risk management to the enterprise's circumstances	33
Box 2.7. Using AI to support RBC due diligence	34
Box 2.8. Enabling transparency, explainability and traceability throughout the AI system lifecycle	36
Box 2.9. Content authentication and provenance mechanisms	37
Box 2.10. Pre-deployment response plan	38
Box 2.11. Preventing or mitigating risks when deploying AI systems	39
Box 2.12. Deployment in contexts where laws are inconsistent with international standards on RBC	40
Box 2.13. Temporary or permanent suspension of the functioning of the AI system	40
Box 2.14. Special considerations for enterprises engaging with 'control points'	43
Box 2.15. Understanding disengagement from business relationships in the context of risks	44
Box 2.16. Practical examples of due diligence for investors and financial institutions investing in the development of AI systems	45
Box 2.17. Potential options for remedying adverse impacts	49

# Executive summary

This guidance aims at supporting enterprises in their implementation of the MNE Guidelines and the AI Principles. This guidance is intended to be used as a tool for multinational enterprises involved in the AI system value chain.

Chapter 1 introduces the concept of RBC due diligence and provides an overview of the broader AI risk management policy landscape. It also describes the target audience and how to use this guidance as a tool to navigate risk management frameworks

The OECD due diligence framework described in the MNE Guidelines and elaborated on in the OECD Due Diligence Guidance for Responsible Business Conduct serves as the foundation for this guidance. The MNE Guidelines and related OECD RBC standards provide voluntary principles for responsible business conduct. The due diligence framework outlines the following measures:

- Step 1: Embed RBC into policies and management systems
- Step 2: Identify and assess actual and potential adverse impacts
- Step 3: Cease, prevent, and mitigate adverse impacts
- Step 4: Track implementation and results of due diligence activities
- Step 5: Communicate actions to address impact
- Step 6: Provide for or cooperate in remediation when appropriate.

Chapter 2 lays out the RBC due diligence framework and practical implementation examples for enterprises involved in the development and use of AI systems. The due diligence framework presented in this guidance also features a roadmap of related provisions in existing frameworks at the beginning of each step indicating how each step of the due diligence framework is complemented by and relate to relevant provisions from related AI risk management frameworks.

“Practical examples for implementation” are included in each step to further illustrate ways to implement and adapt as needed, the supporting measures and due diligence process. The practical examples have been selected to fit this context and also draw on leading AI risk management frameworks as well as desk research and consultations with experts. The practical examples are not meant to represent an exhaustive check list. Not every practical example will be appropriate for every situation. Likewise, enterprises may find additional examples or implementation measures useful in some situations.

By meaningfully implementing the recommendations of the existing AI risk management frameworks, including those described in this guidance enterprises can observe many of the expectations of the RBC due diligence approach. In some cases, the RBC framework provides additional clarity and closes gaps in other frameworks particularly with respect to stakeholder engagement and remediation, which are less comprehensively addressed in existing frameworks.

# **1** Introduction to RBC due diligence and key considerations for AI

---

This chapter introduces the concept of RBC due diligence and provides an overview of the broader AI risk management policy landscape. It also describes the target audience and how to use this guidance as a tool to navigate risk management frameworks.

---

## Introduction to responsible AI

AI development has the potential to transform society in ways comparable to the industrial revolution or the advent of the internet. AI represents not merely an incremental advance but a transformative technology with the capacity to enhance productivity, create economic value, and solve complex challenges across a variety of sectors such as healthcare, manufacturing, logistics, and public administration. To harness this positive potential, the OECD sets out a balanced approach to responsible AI that enhances the opportunities of AI and establishes conditions for AI to be more profitable, innovative and competitive, while addressing risks of adverse impacts.

Responsible AI also depends on data suppliers, finance, and physical infrastructure as much as digital innovation. The OECD Guidelines for Multinational Enterprises on Responsible Business Conduct (“MNE Guidelines”) (OECD, 2023<sup>[1]</sup>) and the OECD Recommendation on Artificial Intelligence (“OECD Recommendation on AI” or “AI Principles” as relevant) (OECD, 2024<sup>[2]</sup>) also emphasise the role of all enterprises involved in the development and use of AI systems. This “whole-of-value chain” approach will support secure and resilient AI value chains, more resistant supply chain shocks and interference.

Critically, responsible AI development and use should proceed with stakeholder engagement and workers at the centre of consideration, viewing the technology as an enhancement to human capability rather than a replacement for human labour. By ensuring meaningful engagement with workers and other stakeholders, enterprises can guide AI toward applications that supplement human work rather than automate it away.

Proactively addressing potential harms associated with AI systems creates a foundation of trustworthiness that can significantly accelerate market growth and investment. When enterprises demonstrate commitment to responsible AI development and use – through the best practices described in this guidance and other OECD instruments – they build confidence among investors, customers, regulators, and policymakers. This trustworthiness translates into competitive advantage. Enterprises that establish reputations for responsible practices can more readily access capital markets, attract premium business relationships, and navigate regulatory landscapes with greater ease. Far from hindering innovation, responsible AI can actually accelerate growth by reducing friction and preventing costly reputational and societal damage that might otherwise occur.

Responsible and trustworthy AI is becoming increasingly crucial for accessing global markets as international regulatory frameworks continue to evolve. Companies that proactively integrate harm prevention into their AI development and adoption processes, position themselves advantageously for cross-border expansion, potentially avoiding the substantial costs of retrofitting systems to meet various regional requirements. This forward-looking approach can transform what might otherwise be viewed as compliance costs into strategic investments that yield returns through expanded market access.

The economic case for responsible and trustworthy AI becomes particularly compelling when considering that enterprise customers increasingly include AI risk management in their procurement processes, making trustworthiness not merely an ethical consideration but a business requirement. In this way, addressing AI harms serves both ethical imperatives and business interests simultaneously, creating a virtuous cycle where responsible innovation drives both societal benefit and commercial success.

## Purpose of this guidance

This guidance aims at supporting enterprises in their implementation of the MNE Guidelines and the AI Principles<sup>1</sup>. This guidance is intended to be used as a tool for multinational<sup>2</sup> enterprises involved in the AI system value chain (i.e., the supply inputs for the development of AI systems, play an active role in the AI system lifecycle or that use AI systems in their operations, products and services, across all sectors).

The objectives of this guidance are to:

- support innovation, investment and growth of enterprises in the AI value chain by providing clarity on how enterprises can proactively identify and address actual and potential adverse impacts (i.e., risks) that they may cause, contribute to or be directly linked to and harness the positive contributions of AI to society related to topics covered in the MNE Guidelines and AI Principles
- help enterprises navigate existing international, national, multi-stakeholder or industry-led AI risk management and governance frameworks
- promote policy coherence, and where possible interoperability, between the MNE Guidelines, AI Principles and other national or international AI risk management and governance frameworks
- serve as a common reference point for AI risk management frameworks across different jurisdictions.

The OECD due diligence framework described in the MNE Guidelines and elaborated on in the OECD Due Diligence Guidance for Responsible Business Conduct (“RBC Guidance”) (OECD, 2018<sup>[3]</sup>) serves as the foundation for this guidance. The MNE Guidelines and related OECD RBC standards provide voluntary principles for responsible business conduct. Matters covered by the MNE Guidelines may be the subject of domestic law and international commitments. Importantly, OECD RBC standards are aligned and complementary of the UN Guiding Principles on Business and Human Rights (“UNGPs”) (United Nations Office of the High Commissioner on Human Rights, 2012<sup>[4]</sup>) and the ILO Tripartite Declaration of Principles concerning Multinational Enterprises and Social Policy (ILO, 2023<sup>[5]</sup>). This guidance seeks to translate the high-level framework contained in the RBC Guidance into concrete and practical actions for enterprises to identify, prevent, mitigate and remedy actual and potential adverse impacts related to the development and use of AI systems.

To support global cooperation and policy coherence for trustworthy AI, and contribute to interoperability where appropriate, this guidance draws on existing international and national AI-specific risk management frameworks, regulations and other initiatives, to offer practical examples for implementing RBC guidance in the AI context.

By providing a resource for enterprises on how to practically implement the MNE Guidelines while demonstrating consistency and coherence with the AI Principles, the guidance aims to support enterprises operating across multiple jurisdictions and subject to multiple regulatory requirements or engaged in multiple voluntary initiatives to respond to such expectations. This will help companies remain trusted by consumers and will ensure they have the freedom to innovate and be competitive in the global market.

## Target audience

The MNE Guidelines recommend that enterprises carry out due diligence to identify and address any actual and potential adverse impacts (i.e., risks) that they: (1) might cause or contribute to through their own operations; or (2) might contribute to or be directly linked to through their business relationships.

The primary audience of this guidance is composed of enterprises in three groups described in more detail below. They include enterprises involved in the AI system lifecycle<sup>3</sup> as described in the AI Principles (i.e., the planning and design; data collection and processing; model building and/or adaptation; testing, evaluation, verification, and validation; deployment; and operation and monitoring AI systems). It is also addressed to enterprises involved in supplying digital, physical and financial inputs for the development of AI systems (e.g., data annotation services, compute providers, cloud service providers, hardware manufacturers and investors), as well as the sale, licensing, trade, and use of AI systems, as described in the MNE Guidelines and in line with the revised Recommendation on AI. This includes enterprises outside of the “technology sector” that use AI systems in their operations, products and services.

While the framework in this guidance might be relevant for the development and use of other software and technologies, this guidance specifically focuses on AI systems. AI systems are clearly defined by the Recommendation on AI and explained in detail in an accompanying memorandum (OECD, 2024<sup>[6]</sup>).

Roles in the AI value chain and business relationships between different actors are non-linear and overlapping. Likewise, the process for AI development does not progress linearly. For example, some developers of AI systems may also be the same companies that design and manufacture hardware or might be involved in gathering data and annotating datasets. To understand RBC due diligence in the context of the development and use of AI systems, this guidance describes enterprises' due diligence responsibilities by categorising them into different groups according to the activity that they perform.<sup>4</sup>

The three groupings below are therefore not rigid nor exclusive, but intended to inform how enterprises performing different activities should approach due diligence. Enterprises may conduct activities that would place them in multiple groups and should tailor their due diligence approach accordingly. Likewise, some enterprises' due diligence efforts will prioritise risks of adverse arising out of their own operations, while others will prioritise risks arising from their business relationships.

Adverse impacts may be related to the activities described in each of the groups. While most of the examples contained in this guidance cover how to address adverse impacts related to activities listed in Group 2, enterprises in Groups 1 and 3 are still expected to conduct due diligence to address adverse impacts that they may cause through their own operations, products and services.

### **Group 1: Suppliers of AI inputs**

Enterprises in this group are suppliers of AI inputs. They provide inputs into the development of the AI system and are generally considered to be the 'upstream' segment of the AI system value chain. This includes activities pertaining to the provision of inputs in the AI ecosystem, (i.e., the skills and resources, such as data, code, algorithms, models, research, know-how, training programmes, governance, processes, and best practices required to understand and participate in the AI system lifecycle, including managing risks). They include enterprises involved in:

- data provision and data annotation
- dataset creation and curation
- developing, adapting, or providing code for third-party use, including contributions to open-source libraries and software components for AI development
- development of metrics and evaluation measures

It also includes activities related to the provision of financial, logistical, administrative, and hardware inputs needed to support the development of the AI system. They include enterprises involved in:

- the provision of capital (e.g., financial institutions, venture capital, and other providers of capital)
- the provision of digital infrastructure and administrative services (e.g., compute providers, cloud service providers, digital payment platforms, digital labour platforms, operating systems, app stores, security software providers, and enterprise software providers)
- the provision of hardware (e.g., semiconductor manufacturers and distributors, network equipment vendors, and other hardware manufacturers).

This guidance does not cover supply chains of hardware inputs (e.g., mining of raw materials and manufacturing of hardware components) which is the subject of separate RBC due diligence guidance.<sup>5</sup>

## **Group 2: Enterprises actively involved in the design, development, deployment, and operation of AI systems**

Enterprises in this group include enterprises involved in the AI system lifecycle activities listed below. Understanding the AI system lifecycle can help all enterprises better identify risks and interact with business relationships in Group 2. These include enterprises involved in:

- planning and design of the system
- building the model and/or adapting an existing model
- testing, evaluating, verifying and validating the model and systems
- deploying<sup>6</sup> the system, regardless of the distribution channel (including the distribution of open-source software)
- operating the system for customers and monitoring the system.

Group 2 could also include enterprises that modify and re-deploy existing AI models for enterprise-specific use cases.

## **Group 3: Users of the AI system**

Enterprises in this group use AI systems in their operations, products and services, are generally considered to be the ‘downstream’ segment of the AI system value chain. These include financial institutions and enterprises in the ‘real economy’ (i.e., manufacturers and sellers of goods and services, including unrelated to AI systems or technology).

Enterprises in this group should consider due diligence on the AI systems that they use as part of their broader due diligence process across their operations and business relationships. This means prioritising the risks of adverse impacts presented by the AI system in relation to other risks the enterprise might be causing, contributing to or linked to in their specific sectors. For example, if the AI system is not related to significant risks, then the enterprise might prioritise other RBC topics for action, including those not related to AI systems.

### **Box 1.1. Considerations for Small and Medium Sized Enterprises (SMEs)**

AI systems have the potential to deliver vast economic benefits to SMEs, including through access to tools that could allow for increased efficiency at lower costs. SMEs also play a critical role at multiple stages in the development of AI systems. Under RBC standards, SMEs are expected, like other enterprises, to carry out due diligence.

RBC standards also recognise, however, that SMEs may not have the same capacity to implement due diligence expectations as larger enterprises. SMEs that are in initial stages of research and development, proof-of-concept and funding function with limited resources, and tend to allocate such resources to more immediate, practical needs for commercialisation of their products or services. Thus, generally speaking, SMEs may face RBC implementation challenges relating to engaging stakeholders, exercising leverage over business relationships, and bearing the costs necessary to take risk prevention and mitigation measures.

To address these challenges, and to promote implementation of standards on RBC, the guidance particularly encourages SMEs, where possible, to use collaborative approaches and engage in industry initiatives to pool resources (in line with competition law), reduce the costs of due diligence and facilitate access to and harmonisation of information on risks of adverse impacts. RBC standards also recognise that the nature and extent of due diligence should be proportionate to the size of the enterprise, its

involvement with an adverse impact and the severity of the adverse impact. In acknowledging these realities, the MNE Guidelines seek to ensure that SMEs are not subject to undue burdens, allowing them to focus on the most relevant risks within their capacity. Furthermore, they encourage larger enterprises to prioritise engaging with SME business relationships to support their due diligence processes.

In addition to meeting international expectations, implementing RBC standards may open new markets or enable better access to finance for SMEs. It may help in acquiring or retaining staff. Similarly, it may prove pivotal in integrating into value chains as larger business relationships increasingly face RBC due diligence practices.

SMEs can leverage cooperation networks (e.g., regional AI and digital transformation initiatives<sup>1</sup> and regional RBC initiatives<sup>2</sup>) to seek further technical support and clarity when implementing RBC and AI standards.

Notes:

1. See e.g., the OECD-African Union AI Dialogue on AI
2. See e.g., the OECD global engagement programmes in Asia, MENA and Latin America.

### ***Other relevant audiences***

The MNE Guidelines also have a unique promotion and remedy mechanism that relies on National Contact Points for RBC (“NCPs”) <sup>7</sup>. This guidance can also be a useful resource for NCPs in promoting the MNE Guidelines and informing decisions related to accountability for alleged violations of the MNE Guidelines.

This guidance may also be useful for those developing standards related to responsible AI such as policymakers, regulators, industry-led and multi-stakeholder initiatives that seek to support alignment with international standards.

Other relevant audiences may include civil society organisations, workers, workers’ representatives, trade unions, industry associations, and national regulatory authorities, including data protection authorities and sectoral oversight bodies.

Special attention should be given to the implementation challenges faced by enterprises and governments in developing countries, including the need for capacity-building, technical assistance, and differentiated guidance adapted to local regulatory and institutional realities.

Finally, this guidance is also relevant for individuals and groups and their representatives that have been or may be adversely impacted by an AI system.

## **Understanding the risks related to the development and use of AI**

The development and use of AI systems have the potential to positively impact matters covered by the MNE Guidelines. For example, use of AI systems can unlock significant improvements in occupational health and safety through automation of dangerous tasks. In public administration, the use of AI in smart grids, smart cities and connected devices can help predict infrastructure maintenance requirements and direct traffic flows to reduce road congestion. The ability of AI to quickly analyse enormous amounts of data, recognise patterns, and build predictive models make it an important tool to detect financial crime, to combat kidnapping and human trafficking, to identify situations of bonded or child labour, and to analyse crime scenes. More broadly, the use of AI systems presents opportunities for innovation, economic growth, and the promotion of human rights.

In order to achieve these positive benefits, it is important that risks of adverse impacts associated with AI systems are effectively managed.<sup>8</sup> When considering the entire AI system value chain, a larger scope of risks may be relevant. For example, the significant computing power used to train and use some types of AI systems have had a demonstrated impact (OECD, 2022<sup>[7]</sup>), and some services performed by humans, such as data enrichment services, have resulted in harmful labour practices (Partnership on AI, 2021<sup>[8]</sup>). Likewise, as with many new technologies, public and private malign actors may find ways to exploit AI systems. The significant dual-use potential of AI systems and ability to repurpose AI systems can lead to harmful uses even when their design was intended to be innocuous.

The MNE Guidelines recognise that enterprises should carry out risk-based due diligence with respect to actual and potential adverse impacts of their activities related to technological innovation. They also recognise that enterprises involved in the development of new technology or new applications of existing tools should anticipate adverse impacts and challenges raised by technologies, while promoting responsible innovation.

Multiple frameworks exist at the international, regional and national level that describe risks related to the development and use of AI systems and recommend actions companies should take to address those risks. The scope of risks covered in these frameworks varies. While the list of frameworks can inform a range of potentially relevant risks for an enterprise's due diligence efforts, it is non-exhaustive and many of the risks overlap and may be linked to each other. Likewise, not all frameworks are relevant for every enterprise. Each enterprise is expected to identify its priority risk areas based on its individual circumstances, including additional risks not listed here. This guidance takes a risk-agnostic approach to remain evergreen. Future research can be used to complement this guidance as policy views and understanding about risks related to AI systems continue to evolve.

## Characteristics of trustworthy AI

This guidance is also intended to enable the responsible stewardship and development of trustworthy AI systems. In the context of this guidance, the term “trustworthy AI” refers to AI systems that embody the OECD's AI Principles, updated in 2024 (OECD, 2024<sup>[9]</sup>). These are not just principles, but outcomes against which to both assess risk and identify mitigation/prevention responsibilities.

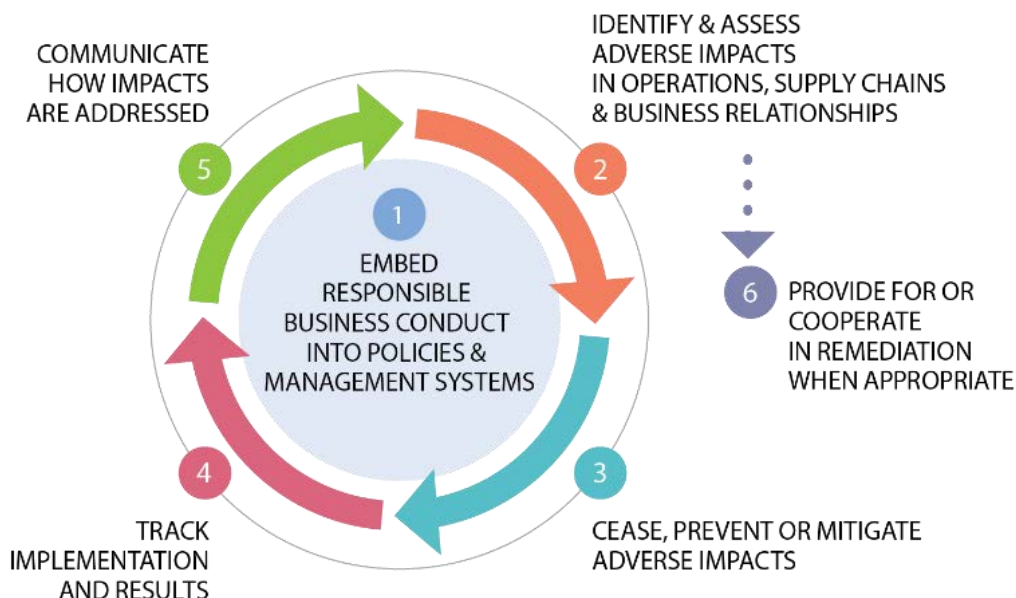
## Basics of RBC due diligence

### *The RBC due diligence framework*

The MNE Guidelines sets out a voluntary due diligence framework for enterprises that governments have committed to actively promote and implement. It outlines the following measures:

1. embedding responsible business conduct into policies and management systems
2. identifying and assessing actual and potential adverse impacts associated with the enterprise's operations, products or services
3. ceasing, preventing and mitigating adverse impacts
4. tracking implementation and results
5. communicating how impacts are addressed
6. providing for or co-operating in remediation when appropriate.

Figure 1.1. Graphical representation of the RBC due diligence framework



Source: OECD (2018<sub>[3]</sub>), *OECD Due Diligence Guidance for Responsible Business Conduct*, <https://doi.org/10.1787/15f5f4b3-en>.

These steps are meant to be simultaneous and iterative, as due diligence is an ongoing process that is both proactive and reactive. The steps are described in more detail and contextualised for the development and use of AI in Chapter 2 of this guidance.

The RBC due diligence framework is broadly aligned with other AI risk management frameworks (OECD, 2023<sub>[10]</sub>) and there is significant overlap across many of the frameworks on key issues. Each of the steps of this guidance points directly to related requirements in existing AI risk management frameworks and can therefore support cross-referencing and coherent implementation of requirements across frameworks and jurisdictions. By meaningfully implementing the recommendations of the other existing AI risk management frameworks enterprises can observe many of the expectations of the RBC due diligence approach. In some cases, the RBC framework provides additional clarity and closes gaps in other frameworks particularly with respect to stakeholder engagement and remediation, which are less comprehensively addressed in existing frameworks.

### **Relationship with legal obligations**

The MNE Guidelines provide voluntary principles and standards for RBC consistent with applicable laws and internationally recognised standards. Matters covered by the MNE Guidelines may be the subject of domestic law and international commitments. The MNE Guidelines outline recommendations on RBC that may go beyond what enterprises are legally required to comply with. The recommendation from governments that enterprises observe the MNE Guidelines is distinct from matters of legal liability and enforcement (see MNE Guidelines, Preface, Paragraph 5 (OECD, 2023<sub>[11]</sub>)).

The MNE Guidelines provide that obeying domestic laws in the jurisdictions in which the enterprise operates and/or where they are domiciled is the first obligation of enterprises (see MNE Guidelines, Ch.1, Paragraph 2 (OECD, 2023<sub>[11]</sub>)). Due diligence can help enterprises observe their legal obligations on

matters pertaining to the RBC. In jurisdictions where domestic laws and regulations conflict with the principles and standards of the MNE Guidelines, due diligence can also help enterprises implement the MNE Guidelines to the fullest extent. Domestic law may also in some instances require an enterprise to take action on a specific RBC issue, (e.g., laws pertaining to specific RBC issues such as online risks to minors).

Due diligence expectations derived from or referencing the MNE Guidelines are increasingly being integrated into legal requirements. While this guidance may be helpful to enterprises and governments in better understanding how they could implement some of these legal requirements, it should not be relied on as a blueprint for compliance.

### ***Business confidentiality***

When implementing RBC due diligence, sufficient attention should be paid to commercial confidentiality, commercial secrets, commercially sensitive information and possible competition law prohibitions relating to the sharing of such information as well as information protected through intellectual property laws. While these are legitimate barriers to some aspects of disclosure, transparency and stakeholder engagement, enterprises are still expected to make good faith efforts to communicate and engage meaningfully with stakeholders while appropriately taking into account confidentiality, competition law, other relevant legal concerns, and in view of legitimate confidentiality considerations.

### ***Implementing RBC in line with competition law***

Collaborating with competitors or business relationships to support the implementation of RBC, including as part of sustainability initiatives, is subject to competition law (OECD, 2015<sup>[11]</sup>).

The MNE Guidelines affirm that “while enterprises and the collaborative initiatives in which they are involved should take proactive steps to understand competition law issues in their jurisdiction and avoid activities which could represent a breach of competition law, credible responsible business conduct initiatives are not inherently in tension with the purposes of competition law and typically collaboration in such initiatives will not be in breach of such laws” (see MNE Guidelines, Ch. X, para 121 (OECD, 2023<sup>[11]</sup>)).

There are three broad practical actions that enterprises can consider in understanding issues related to cooperative activity and competition law:

- Seeking the advice of competition authorities: enterprises can seek the advice of competition authorities if they are in doubt as to whether a particular conduct or cooperative activity can be viewed as contrary to competition law and therefore raise regulatory risks.
- Practicing transparency: Authorities tend to be more sceptical of initiatives or agreements amongst competitors if conduct is completely private. Therefore, transparency regarding RBC initiatives can be a useful way of mitigating competition concerns. Importantly the simple fact that an agreement is overt or that there is transparency around an initiative does not shield it from the application of law if it is indeed anticompetitive. However, transparency can help bring to light potentially problematic issues and thus ensure they are addressed quickly.
- Integrating RBC initiatives with compliance programmes: As enterprises have the responsibility to self-assess whether their conduct poses concerns under competition law they are encouraged to develop and implement compliance programs to ensure there is awareness of the risks and an understanding of how they should be managed at an organisational level. Most large enterprises will likely already have established anti-trust compliance programs in place which can be reference or adapted for the purpose of specific collaborative initiatives regarding RBC.

## ***Meaningful stakeholder engagement***

Meaningful<sup>9</sup> stakeholder engagement, especially with workers, workers' representatives and trade unions, affected communities, or other stakeholders who are most vulnerable to risks of adverse impacts, is essential for effective due diligence. Such engagement supports the development of trustworthy AI systems. Stakeholder engagement is an integral part of all of the steps of the due diligence framework. In some jurisdictions, stakeholder engagement may also be a right in and of itself (e.g., patient consent when applying AI in medical contexts, see also EU AI Act Article 61).

Meaningful stakeholder engagement can also have numerous benefits for enterprises, including through building trust and resilience to crises, and also stronger alignment with market and societal expectations. When stakeholders are involved throughout the due diligence process, they understand not just what decisions were made but why, creating procedural trust even when they might not agree with every choice. Early external feedback helps pivot approaches before significant resources are invested to address less significant risks, potentially reducing costs.

External perspectives might help identify underserved market segments or overlooked use cases. Direct engagement with end users of AI systems might also reveal actual needs rather than assumed ones. For example, enterprises developing and using AI systems in healthcare might consult with medical professionals and representatives of certain patient groups to discover that interpretability is more important than marginal accuracy improvements. AI systems developed with stakeholder input also face fewer barriers to adoption since key concerns have been addressed proactively.

Stakeholders can be impacted at multiple stages across the development and use of AI systems. This includes, for example, individuals whose private data or intellectual property (IP) are used to train AI systems, workers involved in data enrichment services in the development phase, and communities impacted by AI compute harms. Post-deployment, it can include, for example, workers being monitored by AI systems and individuals impacted by government services that use AI systems.

Stakeholder engagement involves interactive processes of engagement with relevant stakeholders through, for example, meetings, hearings or consultation proceedings. Relevant stakeholders are persons or groups, and/or their legitimate representatives, who have rights or interests related to the matters covered by this guidance that are or could be affected by adverse impacts associated with the enterprise's development, deployment, operation, financing, sale, licensing, trade, and/or use of AI systems.

To be meaningful, stakeholder engagement should be two-way, conducted in good faith and responsive to stakeholders' views. Stakeholders should be provided with timely, truthful and complete information and should be given an opportunity to provide input prior to major decisions being made that may affect them. Where appropriate, it is particularly important to have stakeholders actively participate in the identification of adverse impacts.

The rapid development and real-time modifications of AI systems may require enterprises to develop or adapt current practices to ensure meaningful stakeholder engagement. Stakeholder engagement should not be seen as a one-off event, but rather as a continuous process built into the AI system lifecycle and where relevant other aspects of the development and use of AI (e.g., when collecting data or during end-use). Practically, there are a number of ways in which enterprises may engage with stakeholders.<sup>10</sup> Stakeholders can be involved:

- as part of internal discussions about the product purposes and desired impact
- as part of product design
- as part of dataset curation and validation
- as part of training and testing
- ongoing during the use of the AI system

- as part of post deployment testing and evaluation of AI systems
- through multi-stakeholder initiatives and independent assessment processes
- as part of regular trainings for workers in contexts where AI systems are used to manage worker activity. It is critical that workers are updated and informed about the capabilities and risks of AI systems regularly, so that they can meaningfully engage in discussions about the due diligence process.

In addition to any other mechanisms they may implement regarding stakeholder engagement, enterprises should respect the right of workers to establish or join trade unions and representative organisations of their own choosing, including by avoiding interfering with workers' choice to establish or join a trade union or representative organisation of their own choosing (OECD, 2023, MNE Guidelines, Ch. V, para 1(a)).

In very large enterprises that may develop or use hundreds of AI systems, stakeholder engagement at multiple stages of development of each AI system might not be feasible. SMEs may also face resource and access challenges with engaging stakeholders. All enterprises may face challenges related to the speed of development of AI resulting in a constrained availability of relevant stakeholders to engage with enterprises. These challenges suggest that stakeholder engagement requires careful planning and that enterprises may choose different modalities of stakeholder engagement such as to approach it at a higher level, with the objective of transferring learnings to the product-level. Product-specific engagement (e.g., in the design or with impacted stakeholders) may only be practical for certain high-risk contexts.

Limited stakeholder literacy on AI and RBC issues might also present engagement challenges. When stakeholders understand an AI system's capabilities, limitations, and potential consequences, they can make informed decisions and provide meaningful input on risk management processes. This knowledge empowers them to identify potential adverse impacts before they occur and advocate for responsible development and deployment practices. By investing in education and transparent communication about AI systems, enterprises can foster trust and encourage collaborative problem-solving, ultimately leading to a more efficient due diligence process.

Together enterprises and stakeholders are encouraged to identify methods for engagement that are feasible and effective for them. Enterprises should prioritise engaging with stakeholders, or their interlocutors, who are most likely to be affected by the activities of the enterprise. Special efforts should be made to engage with stakeholders who are the most vulnerable to risks of adverse impacts.

## How to use this guidance

This guidance is intended to provide a framework to support implementation by enterprises of the MNE Guidelines, the RBC Guidance and OECD Recommendation on AI and should be used in conjunction with these standards, as well as other international, national and industry frameworks, initiatives and other sources of risk management information, particularly context-specific guidance that provide more detail on certain risks or use cases. Other sources of information are referenced throughout the document.

Acknowledging that relevant AI related regulation and voluntary frameworks differ between countries, enterprises can tailor their due diligence actions to their specific contexts and to the regulatory environments within which they operate. Additionally, compliance with these regulations or frameworks will often contribute towards observances of related provisions of this guidance.

Users of this guidance can first read through and understand the core framework provided and practical examples before turning to more context specific resources that are cited in the endnotes or modules accessible on the OECD.AI Catalogue of Tools & Metrics (OECD, n.d.<sup>[12]</sup>).

# **2** Due diligence framework and practical examples for identifying and addressing risks

---

This chapter lays out the RBC due diligence framework and practical implementation examples for enterprises involved in the development and use of AI systems. The due diligence framework presented in this guidance also features a roadmap of related provisions in existing frameworks at the beginning of each step indicating how each step of the due diligence framework is complemented by and relate to relevant provisions from related AI risk management frameworks.

---

This section of the guidance sets out the OECD due diligence framework for RBC for enterprises involved in the development and use of AI systems and provides practical implementation examples. The due diligence framework presented in this guidance also features a Roadmap of Related Provisions in Existing Frameworks at the beginning of each step indicating how each step of the OECD due diligence framework is complemented by and relates to relevant provisions from related AI risk management frameworks.

“Practical examples for implementation” are included in each step to further illustrate ways to implement and adapt as needed, the supporting measures and due diligence process. The practical examples have been selected to fit this context and also draw on leading AI risk management frameworks as well as desk research and consultations with experts. The practical examples are not meant to represent an exhaustive check list. Not every practical example will be appropriate for every situation (e.g., the Hiroshima Process International Code of Conduct is targeted at advanced AI systems). Likewise, enterprises may find additional examples or implementation measures useful in some situations.

The tables featured under each step demonstrate where similar requirements or expectations can be found in other national and international frameworks on AI risk management. These tables and the related practical examples aim to support enterprises in cross-jurisdictional implementation. Although the provisions are related to implementation of the RBC due diligence framework, this table is not an equivalency framework, as the scope and nature of the expectations in the other frameworks may vary.

## Step 1 - Embed RBC into policies and management systems

**Table 2.1. Step 1: Roadmap of related provisions in existing frameworks**

AI risk management framework	Related provisions
ASEAN Guide on AI Governance and Ethics (“ <b>ASEAN Guide</b> ”)	Section C.1: 1. Internal governance structures and measures and Annex A: 2 Internal governance structures and measures
<b>Australia Guidance for AI Adoption (Implementation Practices)</b>	Implementation Practice 1, 2, and 4
Canada Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems, and Implementation Guide for Managers of Artificial Intelligence Systems (“ <b>the Canada CoC</b> ”)	Accountability
Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Council of Europe, 2024) and the associated methodology for the Risk and Impact Assessment of AI Systems from the point of view of Human Rights, Democracy and Rule of Law (“ <b>CoE HUDERIA</b> ”)	HUDERIA Workflow
European Union AI Act (“ <b>EU AIA</b> ”)	Art. 9.1-3: Risk Management System, Art. 17.1 Quality Management System
European Union Digital Services Act (“ <b>EU DSA</b> ”)	Art. 14: Terms and Conditions and Art. 45: Code of Conduct
European Union Corporate Sustainability Due Diligence Directive (“ <b>EU CSDDD</b> ”)	Art. 7: Integrate due diligence into policies and risk management systems
Hiroshima Process International Code of Conduct for Advanced AI Systems and Reporting Framework (“ <b>Hiroshima Process CoC</b> ”)	Principle 4 (paragraph 2) and Principle 5 (paragraphs 1,3, and 4) and Principle 7
Institute of Electrical and Electronics Engineers (IEEE) Standard Model Process for Addressing Ethical Concerns during System Design 7000-2021 (“ <b>IEEE 7000</b> ”)	6: Key roles; 7: Concept of Operations (ConOps) and Context Exploration Process
International Standards Organisation / International Electrotechnical Commission Information technology — Artificial Intelligence — Guidance on risk management 23894 (“ <b>ISO/IEC 23894</b> ”)	5.1: General; 5.2: Leadership and Commitment; 5.3: Integration; 5.4: Design

AI risk management framework	Related provisions
ISO/IEC 38507:2022 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations (“ <b>ISO/IEC 38507</b> ”)	4: Governance implications of the organizational use of AI (includes governance and accountability considerations); 6: Policies to address use of AI (includes oversight, decision-making, data use, compliance, and culture and values).
ISO/IEC 42001 Information technology — Artificial intelligence — Management system (“ <b>ISO/IEC 42001</b> ”)	Broadly covered in: 4: Context; 5: Leadership; 6: Planning; 7: Support; 8: Operation; 9: Performance evaluation; 10: Improvement. Implement organizational and technical measures as necessary to address identified risks.  Covered in detail in Annex: A.2 (AI policies) and sub-controls A.2.2–A.2.4 and Annex A.5.2 (AI impact assessment) for risk management
Japan AI Guidelines for Business Version 1.1 (“ <b>Japan AI Guidelines for Business</b> ”)	Part 2 E. Building AI governance, Part 3, 4, 5; Appendix 2. “Section 2. E. Building AI Governance”, and Appendix 3, 4, 5.
<b>Korea AI Basic Act</b>	Art. 34
<b>Singapore AI Verify Testing Framework</b>	1.1.1 – 9.6.3
United Kingdom Department for Science, Innovation & Technology (DSIT) Guidance: Introduction to AI Assurance (“ <b>UK DSIT AI Assurance Framework</b> ”)	3.2: AI assurance and governance; 6.1: Steps to build AI assurance
UN Guiding Principles on Business and Human Rights (“ <b>UNGPs</b> ”)	Operational Principles 15 and 16: Policy Commitment
United States National Institute of Standards and Technology, AI Risk Management Framework (“ <b>US NIST RMF</b> ”)	Govern

### **Step 1.1 – RBC policies**

Devise, adopt and disseminate a combination of policies on RBC issues that articulates the enterprise’s commitments to the principles and standards contained in the OECD AI Principles and the MNE Guidelines. Policies should include plans for implementing due diligence, which will be relevant to the enterprise’s own operations and business relationships in the development and use of AI.

### **Practical examples for implementation (for all enterprises in the AI value chain, Groups 1-3)**

1. In addition to commitments to relevant RBC principles and standards, commit to implement the OECD AI Principles, as relevant, through the design, development, deployment, operation and use of AI systems (i.e., human-centred, fair, transparent, explainable, robust, secure, safe, and accountable).
2. Develop or review and update existing RBC policies, including risk management policies, with the active participation of stakeholders, including workers, workers’ representatives and trade unions, to align with principles from relevant international, regional and national frameworks.
3. Build on findings from the risk assessment (see Step 2) in order to update or more clearly define the enterprise’s approach to addressing the most significant risks identified (see Step 3).
4. Set risk-tolerance thresholds to help determine low, medium, high severity and likelihood of risks and guide appropriate responses or trigger deeper due diligence.
5. Make risk management policies publicly available as appropriate, (e.g., on the enterprise’s website, and when relevant, in the local languages of areas where the enterprise operates or maintains business relationships). In some cases, enterprises may wish to make more detailed policies and risk management information available to specific stakeholders and business relationships (see Step 5) or take specific commitments on certain risks and issues.

6. Consider policies that are sufficiently flexible and technology neutral to allow a margin for future developments while also being precise enough to provide guidance to operational teams.
7. Consider aligning policies with relevant national strategies to ensure coherence between RBC commitments and the host country's priorities.

### **Step 1.2 - Internal management systems**

Seek to embed RBC issues and trustworthy AI into the enterprise's policies, oversight bodies, structures, systems, processes, and teams, so that they are implemented as part of the regular processes; taking into account the potential independence, autonomy and legal structure of some of these bodies that may be foreseen in domestic law and regulations.

### **Practical examples for implementation (for all enterprises in the AI value chain, Groups 1-3)**

1. Assign oversight and responsibility for AI due diligence to relevant senior management and assign responsibilities to the board of directors for AI RBC more broadly.
2. Assign responsibility for implementing aspects of the policies across relevant departments with particular attention to those staff whose actions and decisions are most likely to increase or decrease risks of adverse impacts (e.g., development teams, staff involved in data gathering, data annotation and content moderation, system design, and/or service or product procurement, systems designed to make decisions of consequence). It is important that roles and responsibilities and lines of communication related to risk management are documented and communicated to individuals and teams throughout the enterprise.
3. Develop or adapt existing information and record-keeping systems to collect information on AI risk management processes for adverse impacts (e.g., processes for relevant teams to inventory AI systems, and document and communicate risks of the AI systems they design, develop, deploy, evaluate and use); such as to:
  - a. Develop policies and practices to collect, consider, prioritise, and integrate feedback from other units within the enterprise – external to the team that developed or deployed the AI system – regarding potential risks (e.g., procurement, sales, compliance, export control, marketing, and human resources).
  - b. Develop mechanisms to enable the team that developed or deployed AI systems to regularly incorporate external and internal feedback on risks into system design and implementation.
  - c. Develop transparency policy and procedures to clearly map and communicate, both internally and externally, the risks presented by AI systems across all business relationships.
4. Establish channels of communication, or utilise existing channels of communication, between relevant senior management and implementing departments for sharing and documenting risk and risk management decision-making.
5. Communicate the policies to the organisation's relevant staff (e.g., during staff orientation or training, during the design review process, for customer relationship management staff, as a standing item of board meetings, etc).
6. Develop policies, procedures, and training to ensure that staff are familiar with their duties related to risk management and the organisation's risk management practices.
7. Encourage alignment across teams and business units on relevant aspects of the enterprise's RBC policies for AI. This could be done for example by creating cross-functional groups or committees to share information and decision-making about risks, and including business units that can impact adoption of the RBC policies for AI in decision-making.

- a. Consider establishing a dedicated committee or working group with clear roles and responsibilities related to the implementation of AI due diligence. Consider appointing independent external expertise to the committee as part of broader stakeholder engagement efforts.
  - b. Develop incentives for staff and business units that are compatible with the enterprise's RBC policies for AI (e.g., including objectives or metrics in employee evaluations linked to implementation of RBC policies, such as energy consumption, conducting stakeholder engagement activities, and adopting RBC policies; and rewards-based challenges or hackathons linked to developing solutions to RBC challenges).
8. Review existing processes across IT, security, procurement, software development lifecycle (SDLC), etc., to identify how these will interoperate with policies and processes put in place in relation to AI due diligence.
  9. Promote broad decision-making related to AI risk management (e.g., relevant staff or stakeholders are made up from a diversity of disciplines, experience, expertise, and backgrounds).
  10. Develop incident monitoring and response systems.
  11. Develop, draw from or adapt existing complaint or whistleblower procedures for staff to raise issues or complaints related to RBC issues, in compliance with domestic regulations.
  12. Develop policies and practices to foster critical thinking mindset in the design, development, deployment and use of AI systems.
  13. Develop processes for upgrading, decommissioning and phasing out AI systems safely and in a manner that does not increase risks, create new risks or decrease the enterprise's trustworthiness.
  14. Develop contingency plans to handle failures, incidents or adverse impacts linked to AI systems.
  15. Develop a stakeholder engagement plan to allow stakeholders to assess and monitor the implementation of relevant RBC processes across the enterprise, ensuring regular participation of stakeholders in RBC processes, including workers, workers' representatives and trade unions.
  16. Establish or join collaborative initiatives to develop, advance, and adopt, where appropriate, shared standards, tools, mechanisms, and best practices for ensuring the safety, security, and trustworthiness of advanced AI systems, such as the OECD Catalogue of Tools for Trustworthy AI.

### **Step 1.3 - Expectations on business relationships**

Incorporate RBC expectations and policies into engagement with business relationships.

#### **Practical examples for implementation (for all enterprises in the AI value chain, Groups 1-3)**

1. Communicate key aspects of the RBC policies for AI to relevant business relationships, including suppliers of inputs, sales partners and users of AI systems.
2. Where enterprises rely on sales partners – business relationships who buy, distribute, integrate, and resell products and services to end customers – develop channels of communication across the sales channel and with relevant external stakeholders to ensure ongoing due diligence.
3. Develop and implement pre-qualification processes for suppliers or customers that takes into account due diligence on relevant RBC issues, where feasible, adapting such processes to the specific risk and context to focus on RBC issues that have been identified as relevant for the business relationships and their activities or area(s) of operation.
4. Where relevant, and particularly with regards to business relationships with SMEs, consider providing adequate resources to suppliers, customers, end-users and other business relationships for them to understand and apply the relevant RBC policies and implement due diligence (e.g.,

participating in targeted awareness raising, training, or capacity building with relevant business relationships and workers). Ideally and where appropriate, resources and additional guidance provided to customers and end users should be as specific and targeted as possible.

5. Seek to understand and address barriers arising from the enterprise's way of doing business that may impede the ability of suppliers, users and other business relationships to implement RBC polices, such as the enterprise's purchasing practices when acquiring algorithms, datasets, software or hardware.

## Step 2 - Identify and assess actual and potential adverse impacts

**Table 2.2. Step 2: Roadmap of related provisions in existing frameworks**

ASEAN Guide	Section C.2: Determining the level of human involvement in AI-augmented decision-making; C.3. Operations Management; and Annex A: 3: Determining the level of human involvement in AI-augmented decision-making
Australia Guidance for AI Adoption (Implementation Practices)	Implementation Practice 2 and 3.2
Canada CoC	Safety Measure 1
CoE HUDERIA	Context-based risk analysis (COBRA), Impact Assessment (IA)
EU AI ACT	Art. 9(2): Identify, analyse and evaluate known and foreseeable risks, Art. 55(1): Obligations for general purpose AI models with systemic risk
EU DSA	Art. 34: Risk assessment
EU CSDDD	Arts. 8 and 9: Identify and assess actual or potential adverse impacts, and, where necessary, prioritise potential and actual adverse impacts
Hiroshima Process CoC	Principles 1, 2, 6 and 7
IEEE 7000	8. Ethical Values Elicitation and Prioritization Process; 9. Ethical Requirements Definition Process
ISO 31000 & ISO/IEC 23894	6.3: Scope, context, criteria – 6.4: Risk Assessment
ISO/IEC 42001	6: Planning (6.1.1, 6.1.2, 6.1.4); 8: Operation (8.2, 8.4); Annex A.5 (Assessing AI system impacts); sub-controls A.5.2–A.5.5
ISO/IEC 42005	5.8 Actual and potential impacts
Japan AI Guidelines for Business	Appendix 1.B. AI's benefits and risks; Appendix 2.A. Building of AI governance and monitoring by management
Korea AI Basic Act	Arts. 32, 33, and 35
Singapore AI Verify Testing Framework	Safety 4.1.1 – 4.3.1
UK DSIT AI Assurance Framework	4.1.1: Measure; 4.1.2: Evaluate, 5.4: Risk Assessment, .5: Impact assessment; 5.6: Bias audit
UNGPs	Operational Principles 17, 18
US NIST AI RMF	Govern 1, 4, Map 1-5

### Step 2.1 – Initial scoping of risks

Carry out a scoping exercise to identify where risks may be present and where they may be most significant.

Multiple frameworks exist at the international, regional and national level that describe risks related to the development and use of AI systems and recommend actions companies should take to address those risks. One objective of this guidance is to support enterprises' implementation of other risk management frameworks. While the list of frameworks can inform a range of potentially relevant risks for an enterprise's due diligence efforts, it is non-exhaustive and many of the risks overlap and may be linked to each other. Likewise, not all frameworks are relevant for every enterprise. Each enterprise is expected to identify its priority risk areas based on its individual circumstances.

When prioritising the order in which risks are to be addressed (see Step 2.4), enterprises should take into account that certain risks are closely linked to or may enable others.

As with many new technologies, public and private malign actors may find ways to exploit AI systems. The significant dual-use potential of AI systems and ability to repurpose AI systems can lead to harmful uses even when their design was intended to be innocuous.

### ***Practical examples for implementation (for suppliers of AI inputs, Group 1)***

1. Identify business relationships that are actively involved in the development of AI systems (i.e., business relationships in Group 2).
2. Develop an initial understanding of the types of AI systems being developed by business relationships, including the risk information related to AI systems described in Step 2.1.

### ***Practical examples for implementation (for enterprises in the AI system lifecycle, Group 2)***

1. Develop an understanding of the enterprise's role in the AI system lifecycle and maintain an up-to-date registry of AI systems linked to the enterprise.
2. Develop an understanding of the risks of potential adverse impacts related to the development and/or use of the AI system(s). This can be done through desk research and consultation on which risks might be associated with the AI system and gathering and reviewing reporting on risks about the AI system or the enterprise that developed or modified the system. Internal sources of risk information include incident monitoring mechanisms, oversight bodies, and communication channels described in Step 1.2. External sources of risk information include reports from national human rights institutions, national AI observatories, regulatory agencies, sector-specific ministries, international and regional human rights accountability mechanisms, civil society organisations and workers, workers' representatives, trade unions, public incident data bases,<sup>11</sup> court cases, grievance mechanisms, and engagement with affected and at-risk communities. The OECD Framework for the Classification of AI Systems also provides a foundation that can be built upon to understand risks related to AI systems (see OECD Framework for the Classification of AI Systems (OECD, 2022<sub>[13]</sub>)).
3. Risk information can be related to:
  - a. *the interaction of AI systems*. While two AI systems may be benign in isolation, their capacity to cause harm can increase if deployed in a manner that they can interact without guardrails in place.
  - b. the nature of the AI system, use-case or product that it is used in (e.g., facial and emotion recognition technology, predictive policing, surveillance technology, marketing technology)
  - c. the type of user/use of the AI system and the business objectives of the user (e.g., media companies, healthcare providers, judiciary bodies, law enforcement, intelligence agencies, individuals)
  - d. sources of data inputs, software, physical extension, and human-in-the-loop aspects of the system (e.g., labour risks during data annotation and content moderation, sourcing private data or intellectual property (IP))
  - e. the geographic/socio-economic/political context where the AI system is deployed (e.g., areas with high levels of corruption, human rights and labour rights abuses, and conflict zones)
  - f. the competency and scientific validity of the AI system (e.g., risks of incompetently or inadequately performing important tasks)

- g. known or reasonably foreseeable circumstances related to the use of the AI system in accordance with its intended purpose, or under conditions of reasonably foreseeable improper use or misuse, which may give rise to adverse impacts (see Box 2.1) for indicators of uses of AI systems that potentially pose higher risks of adverse impacts).
4. Where it is not feasible or practical to conduct in-depth assessments of all AI systems, consider an escalation system (e.g., having a questionnaire for all new AI systems to assess the baseline risk level and anything with indicators of high risk is then escalated for further in-depth due diligence).
  5. Review the findings of the scoping exercise on a regular basis and when significant changes occur (e.g., operating in a new country, development of a new AI system or product, restructuring, engaging with new business relationships, when relevant laws undergo significant changes).
  6. Consider relevant laws, regulations, and standards in areas such as consumer protection or sector specific laws, regulations and standards (e.g., in healthcare, manufacturing, aviation, etc.) to help enterprises understand risks and define high-risk uses of AI systems.

### Box 2.1. Examples of high-risk uses of AI systems drawn from different frameworks

High-risk uses of AI systems will be context specific and criteria for identifying them varies by jurisdiction. The below applications of AI have been identified by some leading AI risk management frameworks as potentially being high-risk:

- uses posing chemical, biological, radiological, and nuclear risks, such as the ways in which barriers to entry can be lowered, including for weapons development, design or acquisition. In this context, it is important that dual use aspects are also considered
- uses in critical infrastructures that may pose risks to health and safety (e.g., transport)
- uses in educational or vocational training that may determine or strongly influence the access to education and professional course of someone's life (e.g., scoring of exams, assignment services)
- uses relating to the provision of mental health advice and companionship, particularly involving minors or other vulnerable individuals
- uses in employment, management of workers and access to employment (e.g., CV-sorting software for recruitment procedures)
- uses related to granting or revoking access to essential private and public services (e.g., access to healthcare or credit scoring informing citizens' opportunity to obtain a loan)
- uses related to law enforcement and the administration of justice (e.g., informing sentencing, parole and probation, pretrial release and detention, surveillance, crime forecasting and predictive policing, and forensic analysis)
- uses related to migration, asylum and border control management (e.g., verification of authenticity of travel documents or decision-making in relation to asylum claims).

Sources: AIA Article 6, and Annexes I and III European Union (2024<sup>[14]</sup>) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and [https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L\\_202401689;G7](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689;G7) (2023<sup>[15]</sup>) Hiroshima Process International Code of Conduct for Advanced AI Systems, [https://www.mofa.go.jp/ecm/ec/page5e\\_000076.html](https://www.mofa.go.jp/ecm/ec/page5e_000076.html).

### **Practical examples for implementation (for users of AI systems, Group 3)**

1. Develop an initial understanding of all uses of AI systems within the enterprise including the risk information related to AI systems described in Step 2.1, paragraph 54(b), (e.g., human resources, marketing, sales, customer service, procurement, due diligence, etc.) and consider which uses call for deeper due diligence. In cases of low risk, further in-depth due diligence activity may not be warranted.
2. Identify business relationships that develop and deploy AI systems used in operations, products and services.

### **Step 2.2 – In-depth assessment of most significant risks**

Starting with the most significant areas of risk identified, carry out iterative and increasingly in-depth assessments of prioritised risks related to (1) the enterprise’s own activities and (2) the enterprise’s business relationships (e.g., suppliers, customers and users).

#### **Box 2.2. Understanding the risk-based approach**

The MNE Guidelines recommend that enterprises take a risk-based approach to adverse impacts, recognising that it may not always be possible to respond to all adverse impacts immediately. This concept is essential for applying due diligence to AI, particularly general-purpose AI systems, where some enterprises might be linked to impacts caused by hundreds or thousands of business relationships using the AI system.

In this respect, the MNE Guidelines clarify that **where enterprises have a large number of business relationships, they are encouraged to identify general areas where the risk of adverse impacts is most significant and, based on this risk assessment, prioritise due diligence accordingly.** The risk-based approach should also take into account the known or reasonably foreseeable circumstances related to the use of the AI system in accordance with its intended purpose, or under conditions of reasonably foreseeable improper use or misuse.

The significance (sometimes referred to as ‘saliency’) of an adverse impact is understood as a function of its likelihood and severity (OECD, 2018<sup>[3]</sup>). Severity of impacts will be judged by their scale, scope and irremediable character (see Table 2.3). Scale refers to the gravity of the adverse impact. Scope concerns the reach of the impact, for example the number of individuals that are or will be affected. Irremediable character means any limits on the ability to restore the situation to the equivalent before the adverse impact. Severity is not an absolute concept and it is context specific. The concept of likelihood, which some frameworks describe as ‘probability’, is interlinked with and should also be considered alongside severity factors.

For example, if an AI system presents risks that are less likely, but very severe and also risks that are very likely, but less severe, the enterprise(s) involved in developing the AI system are expected to demonstrate that both risks are being considered and monitored and that processes are in place to respond to the risk that the enterprise considers most significant at that time.

Enterprises are expected to demonstrate a credible prioritisation process and progress against outcome-oriented and time-bound targets. Engaging with stakeholders on risk prioritisation and being transparent with their prioritisation criteria and rationale will help bolster the credibility of due diligence processes (see also Box 2.6).

## **Practical examples for implementation (for enterprises in the AI system lifecycle, Group 2) – identifying risks in own operations**

1. Catalogue the specific legal requirements and national/international/industry standards applicable to the AI system or AI actor being assessed, including relevant national and international RBC and labour standards.
2. Review TEVV information (Test and Evaluation, Verification and Validation), including those related to experimental design, data collection and selection (e.g., availability, accuracy, representativeness, suitability), system trustworthiness, and construct validation. Make efforts to ensure that tools or metrics that are used to measure, test, or mitigate AI system risk are themselves be tested, have proven, quantifiable utility and have quality assurance testing across all measures.
3. Where relevant to the risk, review evaluations of the AI system involving human subjects.
4. Review how system output is expected to be utilised and overseen by humans.
5. Consult domain experts, users, and enterprises external to the team that developed or deployed the AI system.
6. Consult and engage stakeholders, including workers, workers' representatives and trade unions, impacted and potentially impacted communities, independent experts, and civil society groups to gather information on significant risks, taking into account potential barriers to effective stakeholder engagement. Where directly consulting with stakeholders is not possible, consider reasonable alternatives such as consulting independent expert resources, including human rights defenders, trade unions and civil society groups.
  - a. Consult potentially impacted stakeholders both prior to and during projects or activities that may affect them.
  - b. Consider allowing some stakeholders to participate in or review AI system tests (e.g., A/B tests), with due regard for business confidentiality and IP rights.
7. Consider risks of adverse impacts at the pre-deployment stage and development stage (e.g., model theft, and misuse from internal use).
8. Identify risks to robustness and security of AI systems, for example through mathematical guarantees or adversarial robustness testing.
9. Identify risks to privacy and data governance at the data and model levels (see Box 2.3).
10. Identify risks of the AI system facilitating or advocating for outcomes that result in adverse impacts on human rights and harms to society and the public interest.

### **Box 2.3. Identifying risks to data quality, interoperability and access throughout the AI system lifecycle**

Data collection and processing in the context of AI come with several risks that, if not properly addressed, can undermine the credibility of the AI system and the accuracy of its outputs and results in adverse impacts. Impacts can be the result of illegally or inappropriately obtained data, manipulated data, and asymmetrical access to data. They can arise at the data and the model levels, at their intersection, as well as during the interaction between the human and the AI system.

- At the data level: Data protection impact assessments are the standard procedure to assess risks. This procedure is legally formalised in some jurisdictions. These assessments take into account risks of data poisoning, where the training data is maliciously manipulated to affect a model's behaviour.

- At the model level: The security of an AI model can be assessed based on:
  1. the access level a malicious actor might have, from “black box” (e.g., no knowledge about the model) to “full transparency” (e.g., full information about the model and its training data)
  2. the phases in which an attack might happen (e.g., during training or inference)
  3. whether passive (e.g., “honest but curious”) or active (e.g., fully malicious) attacks are likely given the threat profile
  4. whether the model relies on cross-border transfers of data (e.g., if the developer is a multinational enterprise and data is collected from multiple jurisdictions, or if some of the development of the model is outsourced to other entities).
- At the intersection of data and model levels: Risks include making inferences about certain members of the training dataset through its interactions with the model. Techniques to assess vulnerability levels include statistical disclosure, model inversion, inferring class representatives, and membership and property inference.
- At the human-AI interaction: Training, checklists and verification processes could help identify risks arising from the interaction between the human and the system (e.g., unintentional actions – or lack of action – by developers or users that compromise the privacy or data governance of an AI system).

The OECD Privacy Guidelines, adopted in 1980 and revised in 2013 (OECD, 2015<sub>[16]</sub>), are the cornerstone of the OECD’s work on privacy and are recognised as the global minimum standard for privacy and data protection. Additionally, the Implementation Guidance for the OECD Privacy Guidelines: Chapter on Accountability can assist stakeholders in better understanding and implementing the accountability principle outlined in the Privacy Guidelines through privacy management programmes. Such programs, with their risk-based approach, offer organisations a valuable tool to address evolving risks and challenges, such as those posed by emerging technologies.

Source: OECD (2023<sub>[17]</sub>), *OECD Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI*, <https://doi.org/10.1787/2448f04b-en>; OECD (2013<sub>[18]</sub>), *Guidelines governing the protection of privacy and transborder flows of personal data*, <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0188>.

### ***Practical examples for implementation (for enterprises in the AI system lifecycle, Group 2) – identifying risks related to business relationships using AI systems***

1. Map the organisation’s relevant operations or business relationships relevant to the prioritised risk.
2. Review whether relevant high-risk business relationships have due diligence policies and internal management systems in place (per Step 1).
3. For assessments of business relationships, notably data enrichment services, where available, use information from the organisation’s own, or third parties’ impact assessments, legal reviews, compliance management systems, financial audits, occupational, health and safety inspections; worker’s organisation, trade union, and/or civil society reporting; and any other relevant assessments carried out by the organisation or by industry and multi-stakeholder initiatives and “know your customer” (KYC) processes.
4. Consider an escalation system to flag potentially high- risk sales without unnecessarily encumbering low-risk sales. In some contexts, this process might be related to or can be integrated with existing compliance processes for export control and sanctions. This could include:
  - a. setting criteria to flag high-risk sales and conducting further due diligence before approving flagged sales
  - b. training relevant staff on end use risks and customer due diligence

- c. integrating RBC policies in performance incentives for sales teams and sales partners
- d. integrating RBC requirements in contracts with sales partners
- e. providing sales partners with training and guidance on RBC due diligence
- f. creating a risk information reporting mechanism for sales partners.

***Practical examples for implementation (for suppliers of inputs and users of AI systems, Groups 1 and 3)***

1. When engaging with enterprises involved in Group 2 activities that are flagged by due diligence processes as being high-risk, enterprises should engage directly with business relationships to better understand their due diligence efforts, and where appropriate, encourage them to take further steps to address actual and potential adverse impacts. Specific due diligence information to gather from business relationships includes:
  - a. commitments made by the AI actor to adhere to national or international standards on responsible AI (e.g., the Hiroshima Process Code of Conduct, the EU AI Pact, relevant Codes of Conduct arising out of the EU AIA and/or DSA)
  - b. significant adverse impacts or risks of an AI system identified, prioritised and assessed
  - c. criteria for prioritising which risks to address (see Step 2.4)
  - d. actions taken or planned to prevent or mitigate risks, including where possible estimated timelines and benchmarks for improvement and their outcomes (see Step 3)
  - e. measures to track implementation and results (see Step 4)
  - f. provision of or co-operation in any remediation (see Step 6).
2. Where due diligence information is not available or if business relationships do not provide sufficient detail to inform an enterprise's risk assessment, enterprises can refer to existing assessments – such as assessments shared through a collaborative industry-led or multi-stakeholder initiatives, academic studies, or AI safety ratings service providers, and NGO reports or other market research services – while continuing to engage with the business relationship to make the disclosures available.

***Step 2.3 – Assess involvement with the actual or potential impact (cause, contribute, directly linked)***

Assess the enterprise's involvement with the actual or potential adverse impacts identified. Specifically assess whether the organisation or relevant business relationship caused (or would cause) the adverse impact; or contributed (or would contribute) to the adverse impact; or whether the adverse impact is (or would be) directly linked to its operations, products or services by a business relationship. An enterprise's relationship to adverse impact is not static. It may change, for example as situations evolve and depending upon the degree to which due diligence and steps taken to address identified risks and adverse impacts decrease the risk of the impacts occurring.

**Box 2.4. Understanding involvement with the risk**

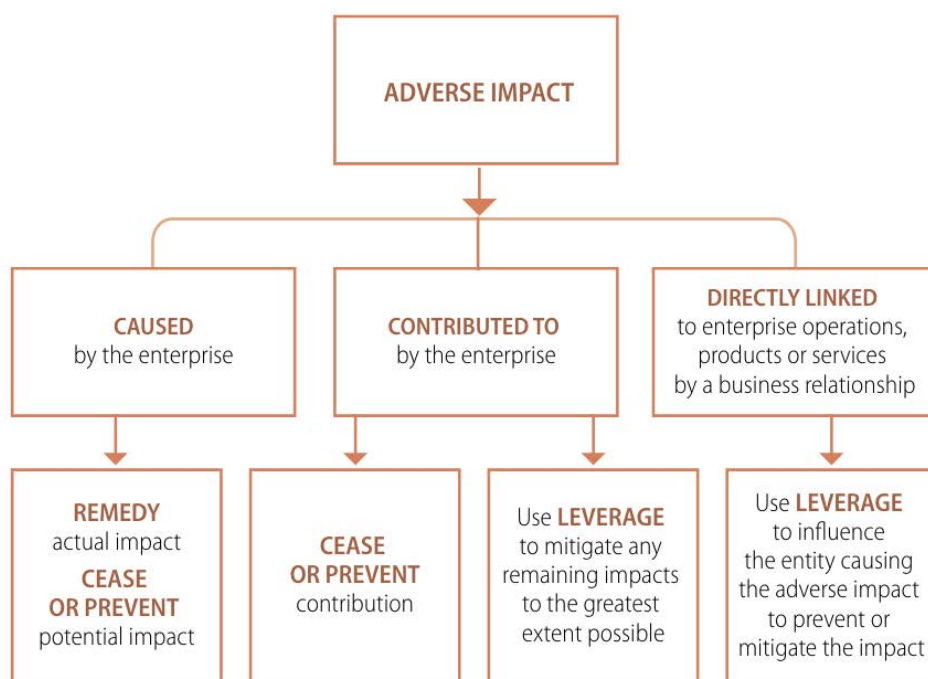
While all enterprises are expected to conduct due diligence, the level of due diligence will vary depending on the involvement with the actual or potential adverse impact. Enterprises causing adverse impacts are expected to cease or prevent potential impacts and remediate harm caused by actual impacts.

Enterprises contributing to adverse impacts are expected to cease or prevent their contribution to the potential impacts and remediate their contribution to the harm and use, and where necessary, build their leverage with business relationships to prevent or mitigate additional risk. An enterprise contributes to an impact if its activities, in combination with the activities of other entities cause the impact, or if the activities of the enterprise cause, facilitate or incentivise another entity to cause an adverse impact. Contribution must be substantial, meaning that it does not include minor or trivial contributions.

Enterprises directly linked to adverse impacts are expected to use, and where necessary build, leverage to prompt the business relationship(s) to prevent or mitigate adverse impacts or risks. “Linkage” is defined by the relationship between the adverse impact and the enterprise’s products, services or operations through another entity (i.e., business relationship). Contribution or direct linkage to an adverse impact once an AI system is deployed, sold or re-sold can often be associated with unmitigated risks in product design or high-risk end users.

See Box 2.5 for scenarios illustrating the involvement framework and the OECD RBC Guidance for a detailed explanation of these concepts.

Figure 2.1. Due diligence expectations based on involvement with the adverse impact



Source: OECD (2018<sup>[3]</sup>), *OECD Due Diligence Guidance for Responsible Business Conduct*, <http://mneguidelines.oecd.org/OECD-Due-Diligence-Guidance-for-Responsible-Business-Conduct.pdf>.

## Box 2.5. Scenarios illustrating the involvement framework

### Causing the impact

Company V developed and deploys a generative AI model capable of creating text, audio and video outputs based on prompts from users. The model is trained using data scraped from publicly available sources. The model is also trained on private data of individual users gathered by Company V when users subscribe to use the model and when users enter personal information in the prompts. Company V does not set up effective guardrails to inform users of how the model is trained on their data. It also fails to prevent private data from leaking through outputs generated by the model.

Company V is causing the adverse impacts on privacy rights.

### Contributing to the impact

Company X developed and operates an AI powered surveillance system that is designed to monitor and analyse worker productivity including through gathering and analysing data through cameras, emails, programs used on company laptops, and internal worker communication tools. The AI surveillance system has the capability to flag when workers exhibit certain types of behaviour such as frequent negative statements towards the company, fatigue, aggression towards colleagues, misuse of company data, etc. Company X is selling this AI surveillance system to other companies. Company A purchases the AI system from Company X and uses it to monitor workers in its distribution warehouse. Company A is reported to have been fined for unlawful union busting activities and monitoring workers without their informed consent. The information gathered and analysis about employee sentiment made by the AI system are then used to inform decisions to terminate the contracts of employees that are believed to be organising a union.

Company X is contributing to the adverse impacts in this scenario because it substantially facilitated Company A in causing the impact through designing an AI system whose intended use was likely to result in adverse impacts and not including sufficient safeguards in the design of the product. Furthermore, it provided that system to a customer whose track record enhanced the likelihood of the adverse impacts occurring.

In a separate scenario involving the same AI system, an Academic Institution is provided access to the AI system to use it in a study on worker productivity. The Academic Institution, however, also uses the AI system to monitor the behaviour of its own employees without their informed consent. In this scenario, the reported intended use of the AI system by the Academic Institution was not likely to result in adverse impacts and the customer was not foreseeably high risk, and as such Company X did not substantially facilitate the adverse impacts.

### Directly linked to the impact

Company Y is in the process of developing an AI system used to detect a form of skin cancer. The AI system is trained using a dataset that contains images of individuals infected by the skin cancer. The dataset was purchased from a service provider, Company B. Company B develops its datasets by having its workers review and label thousands of images, including graphic images. Many such workers have reported adverse impacts on their mental health as a result of this work. Company B is based in a jurisdiction with little protection for workers. Company B has no systems in place to engage with its workers, nor does it provide any form of mental support services to workers.

Here, Company Y is not causing or contributing to the impact on Company B's workers. However, Company Y is directly linked to the adverse impacts caused by Company B because Company Y has a business relationship with Company B.

As noted above, an enterprise's relationship to a risk is not static. If Company Y continues to purchase services from Company B and fails to take any steps to mitigate the risk, then its direct linkage to the risk may evolve into a contribution to the impact over time.

### Step 2.4 – Prioritise the most significant (i.e., most salient) risks

Drawing from the information obtained on actual and potential adverse impacts, prioritise the most significant (i.e., most salient) risks and adverse impacts for action, based on severity and likelihood. Prioritisation will be relevant where it is not possible to address all potential and actual adverse impacts immediately. Once the most significant adverse impacts are identified and dealt with, the enterprise should move on to address less significant foreseeable impacts.

Where the risk of adverse impacts is most significant will be specific to the enterprise. Thus, it is important for enterprises to demonstrate a credible prioritisation process.

Engaging with relevant stakeholders, including workers, workers' representatives and trade unions, on how to prioritise and publicly communicating the rationale behind how prioritisation decisions are made can be useful for establishing trust in the enterprise's due diligence approach. In some cases, prioritisation may also be informed by domestic legal obligations.

**Table 2.3. Factors to consider when prioritising risk**

Scale	Scope	Irremediability (or irreversibility)	Likelihood / Probability / Foreseeability
The gravity of the impact	The reach of the impact	To what extent the impact can be remediated (i.e., any limits on the ability to restore the individuals or environment affected to a situation equivalent to their situation before the adverse impact.	Estimation of the likelihood of an impact occurring
<i>Examples of impacts with a significant scale</i>	<i>Examples of impacts with a significant scope</i>	<i>Examples of impacts with an irremediable character</i>	<i>Examples of likely, probable or foreseeable impacts</i>
<p>Example: An AI system is being used to determine the length of sentences in criminal cases.</p> <p>Example: An AI system is used to generate explicit images of an individual for the purposes of sexual harassment and blackmail.</p> <p>Example: An AI system is used to gather information on and monitor the movements and daily habits of a target group to facilitate actions against them.</p>	<p>Example: Data centres used to power AI systems are consuming excessive amounts of water in communities where local water basins are at risk.</p> <p>Example: Biased recommendations generated by an AI system used in government welfare services results in the cancellation of financial aid for thousands of families.</p>	<p>Example: An AI chatbot recommends to an individual user that they commit self-harm or harm against others.</p>	<p>Example: Numerous reports of AI systems with similar functionality being misused by bad actors.</p> <p>Example: 20% of users of an AI chatbot report that the chatbot shares violent content during their conversations.</p>

## Step 3 - Cease, prevent and mitigate adverse impacts

**Table 2.4. Step 3: Roadmap of related provisions in existing frameworks**

ASEAN Guide	Section C.2: Determining the level of human involvement in AI-augmented decision-making; C.3. Operations Management; and Annex A: 3: Determining the level of human involvement in AI-augmented decision-making
Australia Guidance for AI Adoption (Implementation Practices)	Implementation Practices 1, 2, 3, 4, 5
Canada CoC	Safety Measures 2 & 3; Fairness and Equity; Transparency; Human Oversight and Monitoring; Validity and Robustness
CoE HUDERIA	Impact Mitigation Plan (IMP) and Access to Remedies
EU AI ACT	Recital 115, Art. 9(2a), Art. 9(2)(d), Art. 9(4)-(5), Art. 50 (transparency obligations), Art. 55(1)(b)
EU DSA	Art. 35: Mitigation of risks
EU CSDDD	Arts. 10 and 11: Prevent and (where not possible or immediately possible) mitigate potential adverse impacts; and bring actual adverse impacts to an end and minimise their extent
Hiroshima Process CoC	Principles 1, 2, 6-7, and 11
IEEE 7000	10. Ethical Risk-Based Design Process
ISO 31000 & ISO/IEC 23894	6.5: Risk treatment
ISO/IEC 42001	6.1.3 AI risk treatment; 8.3 AI risk treatment; Annex A A.5 and sub-controls A.5.2–A.5.5 for risk management; A.7 (Data for AI systems) and sub-controls A.7.2–A.7.6 to cover data quality and management
Japan AI Guidelines for Business	Part 2C. Common Guiding Principles, Part 3, 4, 5; Appendix 3, 4, 5.
Korea AI Basic Act	Arts. 31, 32, and 34
Singapore AI Verify Testing Framework	Safety 4.1.1 – 4.6.1 Security 5.1.1 – 5.7.1 Robustness 6.1.1 – 6.5.3
UNGPs	Operational Principle 19
UK DSIT AI Assurance Framework	4.2: AI assurance mechanisms; 5.2: AI assurance spectrum; 5.3: Assuring data, models, systems, and governance in practice; 6.1: Steps to build AI assurance
US NIST AI RMF	Map 1, Manage 1-4

### **Step 3.1 – Addressing risks that the enterprise causes or contributes to**

Cease activities that are causing or contributing to adverse impacts based on the enterprise’s assessment of its involvement with the impact.

Develop and implement plans to prevent and mitigate potential (future) adverse impacts.

#### **Box 2.6. Tailoring risk management to the enterprise’s circumstances**

The nature and extent of due diligence can be affected by factors such as the context of an enterprise’s operations and should be proportionate to the resources of the enterprise, its involvement with an adverse impact and the severity of adverse impacts. Large enterprises with expansive operations and many products or services may need more formalised and extensive systems than smaller enterprises with a limited range of products or services to effectively identify and manage risks.

There are practical limits on how enterprises are expected to respond to risks and impacts. Risk mitigation efforts should be proportionate, taking into account:

- context of the enterprise's operations (e.g., the enterprise's role in the AI value chain, leverage and the enterprise's technical capabilities)
- size of the enterprise
- involvement with an adverse impact (i.e., causation, contribution or direct linkage, see Box 2.5)
- severity of adverse impacts.

Each enterprise should do its part and one enterprise's responsibility should not be shifted to others. Likewise, RBC due diligence is not a standard of perfection, but a standard of improvement. Enterprises are not expected to *immediately* resolve all risks and impacts they are involved with. Rather, enterprises should aim to progressively improve their systems and processes to avoid and address adverse impacts.

### **Practical examples for implementation (for all enterprises in the AI value chain, Groups 1-3)**

1. Assign responsibility to relevant senior staff for ensuring that activities that cause or contribute to adverse impacts cease, and for preventing activities that may cause or contribute to adverse impacts in the future. Depending on the context of the risk, this might include staff from different business units who have the means to take the necessary action to address risks (e.g., research and product development, procurement, customer relationship management, sales, or legal).
2. In the case of complex activities that may be difficult to cease due to operational, contractual or legal issues (e.g., provision of government services, long-term contracts, reliance on a business relationship), create a roadmap for how to cease the activities causing or contributing to adverse impacts. Enterprises may benefit from publicly explaining the complexity of the situation and efforts made to progressively stop activities over time.
3. Consult and engage with impacted and potentially impacted stakeholders and their representatives to devise appropriate actions and implement the plan (see Chapter 1, Meaningful stakeholder engagement).
4. Draw from the findings of the risk assessment to update and strengthen management systems to better track information and flag risks before adverse impacts occur.
5. Update the enterprise's policies with active engagement of stakeholders to provide guidance on how to avoid and address the adverse impacts in the future and ensure their implementation.
6. When deciding which mitigation action to take, consider weighing this action against other possible risks that might occur as well as the benefits of deploying or using the AI system, where relevant.

#### **Box 2.7. Using AI to support RBC due diligence**

Beyond AI value chains, AI has potential to support the implementation of RBC due diligence. For example:

- helping analyse vast amounts of supplier data to identify potential adverse impacts across complex global value chains that could be difficult or costly to monitor manually
- scanning news, reports, and social media to provide early warning of emerging issues related to an enterprise's operations or business relationships and can also provide targeted information on compliance across multiple jurisdictions, helping MNEs navigate evolving frameworks in different markets
- helping verify product origins and conditions of production, enhancing transparency and enabling more effective remediation of identified problems.

## **Practical examples for implementation (for enterprises in the AI system lifecycle, Group 2)**

Actions to avoid and mitigate risk that the AI system causes or contributes to can be broadly categorised in the four below points, however appropriate actions will be based on the specific AI system and use-case and a range of additional actions may be considered:

- **Responsible sourcing and use of data to train of AI models**, including during the data collection and processing stage of the AI system lifecycle (e.g., actions to assess and improve the quality of the data and performance of the AI system and prevention or mitigation of risks of sourcing data gathered or annotated in ways that cause adverse impacts)
- **Transparency, explainability and traceability** especially after deployment (e.g., actions to keep stakeholders informed of the AI system functionality, capabilities and risks).
- **Security, including physical and cybersecurity and robustness** throughout the AI system lifecycle (e.g., actions to ensure the AI system is resilient against attack and demonstrates reliability, repeatability, reproducibility, and predictability).
- **Responsible deployment**, including responsible operation and monitoring and, if appropriate, retirement from production (e.g., actions to assess whether the model is safe to deploy and implementing appropriate guardrails)

### *Responsible sourcing and training*

1. Conduct data quality reviews to identify and address issues such as incorrect labels and representativeness.
2. Implement privacy preserving and responsible data governance approaches to collecting data and training AI systems such as data cleaning, on-device processing, and federated learning. Monitor pre-trained models used for development as part of regular AI system monitoring and maintenance, including through data quality reviews.
3. If the enterprise is not confident it can train a safe model at the scale it initially had planned, they could consider incremental scaling (i.e., training a smaller or otherwise weaker model).
4. Apply state-of-the-art alignment and safety techniques such as inverse reinforcement learning (Centre for the Governance of AI, 2023<sup>[19]</sup>).
5. Take steps to prevent or mitigate risks linked to data collection and processing.
6. Risks related to data quality and sourcing might also be linked to data enrichment services.

### *Transparency, explainability and traceability*

7. Seek to enable transparency, explainability and traceability, in relation to sourcing data from subcontractors, datasets, processes, relevant decisions made during system development, including on human review of significant decisions as well as appeal processes (see Box 2.8).

### Box 2.8. Enabling transparency, explainability and traceability throughout the AI system lifecycle

**Transparency** in this context refers to disclosure to ensure people are aware that AI is being used in a prediction, recommendation, decision, or in an interaction (e.g., a chatbot). Transparency also means enabling people to understand how an AI system is developed, trained, operated, and deployed in the application domain, so that, for example, users and consumers can make more informed choices. Depending on the context and unless required by law, transparency need not extend to the disclosure of the source or other proprietary code or datasets, all of which might be too technically complex to be useful for understanding an outcome.

**Explainability** means enabling stakeholders to understand how an outcome of an AI system is determined. This entails providing easy-to-understand information that can enable those adversely affected to challenge the outcome, notably – to the extent practicable – the factors and logic that led to an outcome.

Explainability can be achieved in different ways depending on the context (such as, the significance of the outcomes). For example, for some types of AI systems, requiring explainability may negatively affect the accuracy and performance of the system (as it may require reducing the solution variables to a set small enough that humans can understand, which could be suboptimal in complex, high-dimensional problems), or privacy and security. It may also increase complexity and costs, potentially putting AI actors that are SMEs at a disproportionate disadvantage.

Therefore, when AI actors provide an explanation of an outcome, they may consider providing – in clear and simple terms, and as appropriate to the context – the main factors in a decision, the determinant factors, the data, logic or algorithm behind the specific outcome, or explaining why similar-looking circumstances generated a different outcome. This should be done in a way that allows individuals to understand and challenge the outcome while respecting personal data protection obligations, if relevant.

From a development perspective, explainability is considered crucial for learning from system failures. Without it, valuable insights from mistakes cannot be gained. The ability to understand decision-making processes that led to failures is seen as essential for improving AI systems and building trust.

**Traceability** in AI describes an effort to maintain a record of the provenance of data, processes, code, and other elements in the development of an AI system. Traceability often captures granular information about an element or component of an AI system, such as the input data or model. It is essential to enable the auditing of a system. To enable transparency and traceability, enterprises could document the following information, where relevant and feasible to their specific activity throughout the AI system lifecycle:

- use and risk information identified in Step 2.1
- data sources, data collection processes and data processing information
- complete code, including necessary libraries
- information on how the code should be executed to guarantee reproducibility of outputs, including detailed documentation of the parameters and computing requirements
- information on how the outputs of the model are used and whether outputs or decisions are generated by AI systems (e.g., disclosure of AI generated images, audio or text)

- information about the monitoring strategy, including performance metrics, thresholds, expected model behaviour, and mitigation actions; information about the deficiencies, limitations, and biases of the model, as well as if and how they are communicated to the relevant stakeholders.

Disclosures should be tailored to their audiences, which may require the creation of multiple means of disclosure at varying levels of detail (e.g., AI nutrition labels, data sheets, model cards, system cards, technical reports, etc).

Sources: OECD (2024<sup>[20]</sup>), AI, data governance and privacy: Synergies and areas of international co-operation, <https://doi.org/10.1787/2476b1a4-en>; US Department of Commerce. National Telecommunications and Information Administration (2024<sup>[21]</sup>), AI Accountability Policy Report, <https://www.ntia.gov/sites/default/files/publications/ntia-ai-report-final.pdf>; European Union (2024<sup>[14]</sup>), Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, [https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L\\_202401689](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689).

8. Seek to enable ways to generate and provide interpretations and explanations of an AI system's output by including the below information in model explanations, if relevant:
  - a. the type and source of model input data
  - b. the high-level data transformation process
  - c. the decision-making criteria and rationale
  - d. a disclosure about using AI.
9. Enterprises should implement mechanisms to provide clear, accessible, and meaningful explanations of automated decision-making processes, especially when such decisions may significantly affect individuals. These explanations should include the logic, main parameters, and potential outcomes of the algorithmic process, tailored to the average user's understanding.
10. Develop and deploy reliable content authentication and provenance mechanisms, where technically feasible (see Box 2.9).

### Box 2.9. Content authentication and provenance mechanisms

Provenance generally refers to the basic, trustworthy facts about the origins of a piece of digital content (image, video, audio recording, document). It may include information such as who created it and how, when, and where it was created or edited. Detecting the provenance of digital content is currently extremely difficult at internet scale and speed because manipulation software is increasingly more sophisticated, metadata can also easily be manipulated and provides no proof of its origins. Technical solutions and best practices are currently under development through collaborative industry and multi-stakeholder efforts such as the Coalition for Content Provenance and Authenticity (C2PA) Guidance for Implementers and Guidance for AI and ML, and the Partnership on AI (PAI) Responsible Practices for Synthetic Media.

Generally speaking, existing best practice on this issue includes three main elements:

- transparency on the capabilities, functionality, limitations, and the potential risks of technology that produce synthetic media
- integrating direct or indirect disclosure methods (e.g., content labels watermarks)
- Investing in the research and development of detection methods and durability of cryptographic disclosure.

Sources: OECD (2022<sup>[13]</sup>), *OECD Framework for the Classification of AI systems*, <https://doi.org/10.1787/cb6d9eca-en>; Coalition for Content Provenance and Authenticity (C2PA) (n.d.<sup>[22]</sup>), *Guiding Principles for C2PA Designs and Specifications*, <https://c2pa.org/principles/>; Partnership on AI (n.d.<sup>[23]</sup>), *Responsible Practices for Synthetic Media*, [https://syntheticmedia.partnershiponai.org/#read\\_the\\_framework](https://syntheticmedia.partnershiponai.org/#read_the_framework).

11. Consider contributing to the advancement and standardisation of AI measurement science to fully understand the long-term benefits and risks of AI systems.
12. Consider developing a guide for external stakeholders that provides resources to help stakeholders better understand the AI system. This could take the form of responsible AI documentation, offering stakeholders a single repository for information on intended use cases and limitations, responsible AI design choices, and best practices for deployment and performance optimisation. Such a guide could address issues related to the OECD AI Principles. This would be a constantly evolving document as the AI system and related risks are better understood.
13. Make the information in disclosures sufficiently clear and understandable to enable deployers and users as appropriate and relevant to interpret the model/system's output and to enable users to use it appropriately, and that disclosures should be supported and informed by robust documentation processes.

### *Security, safety and robustness*

1. Develop approaches to support robustness, security and safety throughout the AI system lifecycle, for example:
  - a. red teaming the different phases of the AI system lifecycle, where commensurate with the risk
  - b. monitoring the AI system behaviour including mechanisms for capturing and evaluating input from users and other relevant AI organisations, appeal and override, decommissioning, incident response, recovery, and change management.
  - c. establishing mechanisms to quickly respond to AI system failures, such as mechanisms to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.
  - d. establishing a robust insider threat detection programme.
  - e. securing model weights.
  - f. monitoring performance changes against agreed metrics during the AI system lifecycle.
  - g. establishing cybersecurity protections.
  - h. monitoring AI system outcomes for data and model drift.

### *Responsible deployment*

1. Engage with stakeholders for input on system requirements and design decisions (e.g., “the system shall respect the privacy of its users”) pre-deployment (see Box 2.10).

#### **Box 2.10. Pre-deployment response plan**

A clear and transparent response plan is an essential element for preventing risks when deploying an AI system. This plan could clearly define the following aspects:

- Risk scenarios that warrant deployment corrections, as developed in the threat modelling process, and triggers to identify deviations from expected behaviour.
- The composition of the response team, comprising representatives from IT, cybersecurity, AI development, legal, communications, relevant business units, and external domain experts. Due to the variety of potential risk scenarios, incident response may require expertise beyond what AI developers can handle alone, and require inputs from multiple parties.

- The roles and responsibilities of different teams and individuals involved in the incident response process. To act swiftly, everyone on the team should know their responsibilities and the decisions that are theirs to make.
- The extent to which decisions are automated versus left to human operators.
- The extent to which authority is shared.
- The extent to which deployment correction protocols are binding.

For more detail, see Institute for AI Policy and Strategy (2023<sup>[24]</sup>), ISO (2023<sup>[25]</sup>).

2. Consider measures to gradually deploy the AI system as evidence about risks emerges. Research has identified a ‘gradient system of access’ when deploying generative AI models, ranging from fully closed and gradual/staged release at one end of the gradient to downloadable and fully open at the other end (Solaiman, 2023<sup>[26]</sup>). Each level of access comes with risks and trade-offs that should be taken into account when considering how to prevent and mitigate risks linked to the AI system (see Box 2.11).

### Box 2.11. Preventing or mitigating risks when deploying AI systems

Research has identified specific technical tools that enterprises can invest in and non-technical actions that enterprises can take to address risks when deploying generative AI systems. The relevance of these will vary across systems and deployment scenarios, so they are provided as exemplary potential tools and actions for enterprises to consider.

Technical tools include:

- Rate limiting – Constricting the number of outputs a user can generate
- Safety and content filters – Filters developed to trigger blank response when given a potentially unsafe input
- Detection models – Technical and human detection of AI generated content
- Hardcoding responses – Predetermined safe outputs triggered for a given input which can be hardcoded into a model interface

Non-technical actions include:

- Internal risk policies and codes of conduct (see Step 1).
- Establishing legal guardrails through licenses where enforcement for violating terms of use can be pursued by the license owner. This is possible through more restricted types of AI systems, but potentially difficult for fully downloadable or open systems.
- Investments in human rights and labour rights training for non-experts involved in the design and deployment of the AI system.
- Investments in risk foresight exercises.

Source: OECD (2023<sup>[17]</sup>), Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI, <https://doi.org/10.1787/2448f04b-en>.

3. Consider monitoring and controlling model or system usage (e.g., by collecting know-your-customer (KYC) information and restricting access to the system or some capabilities of the system). This can be done through a risk-based gating and escalation approach (see (BSR, 2022<sup>[27]</sup>)).

4. Develop adequate assessments and monitoring measures internally and support external researchers who have the resources and understanding to support post-deployment assessment.
5. Establish and integrate feedback processes for end users and relevant stakeholders to report problems and appeal system outcomes.
6. Update and fine-tune the AI system post-deployment based on on-going monitoring and post-deployment assessment. Methods for fine-tuning include reinforcement learning with human feedback or making adjustments to datasets.
7. Where significant adverse impacts are imminent or severe harms are actually occurring – it is important to cease development and deployment of the AI system in a responsible manner until risks can be sufficiently managed.
8. Assess whether legal requirements on AI development and use may cause impacts (see Box 2.12).

### Box 2.12. Deployment in contexts where laws are inconsistent with international standards on RBC

In contexts where domestic legal requirements may contradict international standards on RBC, enterprises should clearly and widely communicate commitment to respect internationally recognised human rights.

As a preventative measure, this commitment can be clearly communicated and negotiated upfront, prior to deployment. Where the legal context has changed, encourage governments to comply with their human rights obligations, particularly where there are direct links with enterprise's operations. Avoid contributing to the unjust criminalisation of human rights defenders or the use of AI systems to repress peaceful protest. Consider not entering or withdrawing from contexts where human rights cannot be respected.

9. Examples of actions that could be used by enterprises to mitigate risks of misuse of AI systems include the following:
  - a. user-based restrictions (e.g., consider declining to work with specific users or groups where risk of misuse is significant)
  - b. access frequency limits (e.g., limiting the number of outputs a system can produce per hour)
  - c. capability or feature restrictions (e.g., filtering outputs or reducing a system's context window)
  - d. use case restrictions (e.g., prohibitions of certain applications of the system in certain contexts)
  - e. temporary or permanent suspension of the functioning of the AI system (see Box 2.13).

### Box 2.13. Temporary or permanent suspension of the functioning of the AI system

Suspension protocols can include a process for authorising re-deployment, or for alternative recovery plans. It is important that this recovery process go through extensive testing and validation, ideally involving external stakeholders. There should be a high bar for re-deploying a model that is demonstrably capable of producing adverse impacts. Where fixes are not possible or sufficiently robust, alternative plans to re-deployment should be pursued (e.g., decommissioning the model, and/or coordinating with other actors in government or industry to manage industry-wide responses). In extreme cases, recovery may not be possible.

Source: OECD (2025<sup>[28]</sup>), Towards a common reporting framework for AI incidents, <https://doi.org/10.1787/f326d4ac-en>.

10. Take steps to ensure that business relationships deploying the enterprise's AI systems are also meaningfully engaging with stakeholders before deployment, particularly in the workplace. See Step 3.2 – Addressing risks directly linked to the enterprise throughout the AI value chain.

### ***Practical examples for implementation (for users AI systems, Group 3)***

1. In the context of using AI systems for operational decision-making, enterprise may need to engage with workers, workers' representatives and trade unions to mitigate risks. Engagement with workers can include:
  - a. disclosure of information about the AI system, including on its design and intended use
  - b. developing governance mechanisms that could provide workers with access and rights over the collection and analysis of data that concerns them
  - c. developing or maintaining grievance mechanisms for workers affected by AI
  - d. developing re-skilling and AI literacy programmes.
2. When using AI systems in products and services, enterprises could conduct independent tests where technically feasible (e.g., red teaming or other types of testing exercises) to verify the quality of the outputs and test vulnerabilities. Enterprises should also consider disclosing when outputs are generated or informed by AI systems and to what extent (see Box 2.8).

### ***Step 3.2 – Addressing risks directly linked to the enterprise throughout the AI value chain***

Based on the risk prioritisation, develop and implement plans to prevent or mitigate actual or potential adverse impacts directly linked to the enterprise by business relationships (e.g., temporary suspension of the relationship, continuation of the relationship throughout the course of risk mitigation effort, or disengagement).

Enterprises throughout the development and use of AI might be directly linked to adverse impacts caused by (1) other AI actors in the system lifecycle; or (2) business relationships outside of the AI system lifecycle, such as suppliers of AI inputs and users of the AI system.

Appropriate responses to risks associated with business relationships may at times include:

- **continuation of the relationship** throughout the course of risk mitigation efforts
- **temporary suspension of the relationship** while pursuing ongoing risk mitigation
- **disengagement with the business relationship** either after failed attempts at mitigation, or where the enterprise deems mitigation not feasible, or because of the severity of the adverse impact. A decision and subsequent plan to disengage should take into account potential social, environmental and economic adverse impacts and should include meaningful stakeholder engagement. These plans should detail the actions the organisation will take, as well as its expectations of its suppliers, buyers and other business relationships (see Box 2.15). Disengagement actions should be in line with applicable laws, including competition laws.

### ***Practical examples for implementation (for all enterprises in the AI value chain, Groups 1-3)***

1. Assign responsibility for developing, implementing and monitoring plans to prevent or mitigate actual or potential adverse impacts directly linked to the enterprise by business relationships.
2. Support or collaborate with the relevant business relationship(s) in developing fit-for-purpose plans for them to prevent or mitigate adverse impacts identified within reasonable and clearly defined

timelines, using qualitative and quantitative indicators for defining and measuring improvement (sometimes referred to as “corrective action plans”).

3. Use leverage, to the extent possible and in line with competition law obligations, to prompt the business relationship(s) to prevent or mitigate adverse impacts or risks. Once a product or service has been sold or re-sold, consider ways to exercise leverage through restricting the provision of essential services that the AI system relies on to run (e.g., customer support, updates, running servers, etc.). Using leverage may include:
  - a. engagement with the business relationship to urge them to prevent and/or mitigate impacts through direct communications with staff responsible for addressing risks at the operational, senior management and/or board level to express views on RBC issues
  - b. building expectations around RBC and due diligence specifically into commercial contracts
  - c. linking business incentives – such as the commitment to long-term contracts and future orders – with performance on RBC
  - d. engagement with regulators and policymakers on RBC issues for them to effect change in the wrongful practices of the entity causing the harm
  - e. communicating – publicly or privately – the possibility of disengagement if expectations around RBC are not respected.
4. If the enterprise does not have sufficient leverage to encourage a business relationship to prevent or mitigate an adverse impact, consider ways to build additional leverage with the business relationship in line with competition law, including for example through outreach from senior management and through support and incentives as relevant.
5. To the extent possible, in line with competition law, cooperate with other enterprises or stakeholders to build and exert leverage to encourage the prevention and mitigation of adverse impacts, for example through collaborative approaches in industry associations, or through engagement with governments.
6. To prevent potential (future) adverse impacts and address actual impacts, seek to build leverage into new and existing business relationships (e.g., through policies or codes of conduct, contracts, or written agreements) (see Box 2.14).

### Box 2.14. Special considerations for enterprises engaging with ‘control points’

In the RBC due diligence context, control points are understood to be key points of visibility, leverage or transformation in the development or trade of a product. Assuming it is in line with the risk-based approach and in compliance with domestic laws, including competition laws, focusing due diligence on control points helps to make overall due diligence efforts more efficient.

Generally, control point characteristics include (1) relatively few actors in the value chain that have direct or indirect business relationships with many other enterprises; and (2) they are (or soon will be) the subject of regulations and audits. In the context of the development and use of AI systems, some large enterprises may meet many ‘control point’ characteristics (e.g., semiconductor manufacturers, foundation models, and very large online platforms).

Conducting due diligence on control points to determine whether they are in turn conducting due diligence provides some comfort that risks have been identified, prevented and mitigated, without having to conduct detailed due diligence on every other relevant enterprise. In addition, control points are usually already subject to audits, public reporting requirements or to some form of regulatory control, indicating that focusing due diligence on this point could help avoid duplicative efforts.\*

Identification and engagement with control points can be carried out by including requirements in contracts with supplier and business relationships that control points be identified and meet due diligence expectations; using confidential information-sharing systems on suppliers, and/or through industry wide schemes.

Note: \* Examples include risk management regulations in the EU such as the Digital Services Act, the Corporate Sustainability Due Diligence Directive, and the AI Act.

1. Include conditions and expectations on RBC issues in supplier, sales partner and/or user contracts or other forms of written agreements (e.g., the development of “responsible use guides” or “acceptable use policies” for users).
2. Encourage business relationships causing or contributing to adverse impacts to consult and engage with impacted or potentially impacted stakeholders or their representatives in developing and implementing corrective action plans.
3. Support relevant business relationships in the prevention or mitigation of adverse impacts or risks (e.g., through training or strengthening of their management systems, striving for continuous improvement through measurable, time-bound targets).
4. Encourage relevant authorities in the country where the adverse impact is occurring to act, (e.g., through inspections, enforcement and application of existing laws and regulations).
5. Engage with other enterprises and stakeholders to cease adverse impacts and/or prevent them from recurring or to prevent risks from materialising (e.g., through participating in industry initiatives and engagement with governments).
6. Where a business relationship’s due diligence information is not publicly available, seek to engage with business relationships to increase transparency or to demonstrate due diligence through confidential bilateral or multilateral arrangements (e.g., disclosure to trusted industry or multi-stakeholder initiatives or non-disclosure agreements).
7. As a last resort, consider disengaging from the business relationship (see Box 2.15).

### Box 2.15. Understanding disengagement from business relationships in the context of risks

The dynamics between business relationships in the development and use of AI systems are constantly evolving and more research and consultation is necessary to understand the full implications of disengagement from business relationships in this context. In some cases, disengagement might not be possible. AI is being increasingly integrated into personal and business tools and is becoming a fundamental part of doing business in many sectors of the economy. Likewise, there are only a small handful of enterprises at some stages of the development of AI systems, such as developers of general-purpose AI and semiconductors manufacturers, that are essential sources of supply and difficult or impossible to disengage from.

Generally speaking, according to the MNE Guidelines, disengagement from a business relationship may be considered as a last resort, either after failed attempts at mitigation, or where the enterprise deems mitigation not feasible, or because of the severity of the adverse impact. Where it is possible for enterprises to continue the relationship and demonstrate a realistic prospect of, or actual improvement over time, such an approach will often be preferable to disengagement.

Thus, among the factors that will enter into the determination of the appropriate action in such situations are the enterprise's leverage over the entity concerned and the severity of the impact. How crucial the relationship is to the enterprise, and the potential social, environmental and economic adverse impacts related to the decision to disengage are also relevant factors to consider.

In cases where disengagement is not possible, it is recommended that enterprises report the situation internally, continue to monitor the business relationship, for example, through maintaining a knowledge database, and revisit their decision to continue the business relationship where circumstances change or as part of the enterprise's long-term strategy to systemically respond to all adverse impacts.

It may also be in the enterprise's interest to publicly explain the decision not to end the business relationship, how this decision aligns with their policies and priorities, what actions are being taken to attempt to apply leverage to mitigate the adverse impacts, and how the business relationship will continue to be monitored in the future.

### Box 2.16. Practical examples of due diligence for investors and financial institutions investing in the development of AI systems

Investors and financial institutions play a key role in the development of AI systems and AI firms are attracting substantial financing. For example, the global annual value of AI venture capital (VC) has risen dramatically, from about USD 6.4 billion in 2012 to USD 147 billion in 2024, accounting for 56% of the value of all VC investment by Q3 2025 (OECD, n.d.<sup>[29]</sup>). Some investors, in particular early-stage investors, are thus often in a position to exert leverage as they are supporting investees to shape and define a project's outlook.\* Another area where investors and financial institutions have strong agenda setting ability is the provision of guidance on how investees and clients can themselves invest in AI systems. Investors and financial institutions can play a role in encouraging investees to identify and address actual or potential adverse impacts. Practical implementation examples are to:

- Include risks of adverse impacts in portfolio risk assessments or investment analyses.
- Engage with stakeholders to support risk identification related to investee companies.
- Engage in bilateral dialogues with investees to raise concerns flagged by the risk assessment (including from engagement with stakeholders), learn more about investee company due diligence practices and request additional information/action from the investee company.
- Request investees provide a clear and concise rationale for adopting AI systems.
- Participate in peer-to-peer coordination efforts to build a stronger, more unified investor approach with regards to responsible AI.
- Establish or engage in initiatives to encourage or develop best practice/improved market standards among companies related to responsible AI.
- Sign public pledges with like-minded investors addressing RBC concerns.
- Where necessary, file shareholder resolutions addressing risks of adverse impacts.
- Where other methods of improving due diligence efforts with investee companies are unsuccessful, consider voting against board members, or divesting if compatible with mandates.
- Make public announcements when divesting in a company or excluding a company from an investment portfolio due to its failure to conduct due diligence.

Note: \* Since 2022, investors under the banner of the Collective Impact Coalition for Ethical AI have led outreach to 44 of the 150 companies assessed in the 2021 World Benchmarking Alliance Digital Inclusion Benchmark, focusing on those that did not have a public set of principles to steer their development and use of AI. These engagements have helped push 14 additional companies to announce their AI principles in the 2023 DIB, bringing the total of those with disclosed principles up to 47 out of 150 (31%) (World Benchmarking Alliance, 2023<sup>[30]</sup>).

Source: The OECD Centre for RBC has worked to operationalise RBC due diligence for different financial transactions and actors through developing fit-for-purpose guidance for institutional investors, corporate lending and securities underwriting, and project and asset finance transactions. OECD (2022<sup>[31]</sup>), Responsible business conduct due diligence for project and asset finance transactions, <https://doi.org/10.1787/952805e9-en>.

## Step 4 - Track implementation and results of due diligence activities

Track the implementation and effectiveness of the enterprise's due diligence activities, i.e., its measures to identify, prevent, mitigate and, where appropriate, support remediation of adverse impacts.

**Table 2.5. Step 4: Roadmap of related provisions in existing frameworks**

ASEAN Guide	Section C.3 and Annex A:3
Australia Guidance for AI Adoption (Implementation Practices)	Implementation Practices 1, 2, 3, 4, 5
Canada CoC	Human Oversight and Monitoring
CoE HUDERIA	Iterative requirements
EU AI ACT	Recital 114, Art. 9(5)-(8): Testing, Art. 55(1)(c): Incident reporting for general purpose AI models, Art. 60: Testing of High-Risk AI Systems in Real World Conditions Outside AI Regulatory Sandboxes, Art. 72: Post market monitoring
EU DSA	Art. 37: Independent audit
EU CSDDD	Art. 14: Establish and maintain a notification mechanism and complaints procedure: Art. 15: Monitor the effectiveness of due diligence policy and measures
	Principles 4
ISO 31000 & ISO/IEC 23894	6.6: Monitor and review
ISO/IEC 42001	8.1: Operational planning and control; 9: Performance evaluation; 10: Improvement ; Annex A.6.2.6 and A.6.2.8 (monitoring and logging); Annex A.8.4 (incident communication)
Korea AI Basic Act	Arts. 32 and 34
Singapore AI Verify Testing Framework	Safety 4.1.1 – 4.6.1; Security 5.1.1 – 5.7.1; Robustness 6.1.1 – 6.5.3
UNGPs	Operational Principle 20
UK DSIT AI Assurance Framework	5.7: Compliance audit; 6.1.3: Review internal governance and risk management
US NIST AI RMF	Measure 1, 4

### ***Practical examples for implementation (for all enterprises in the AI value chain, Groups 1-3)***

1. Identify adverse impacts or risks that may have been overlooked in past due diligence processes and include these in the future.
2. Assess whether previously undetected risks exist or previously assessed risks are no longer acceptable.
3. Assess effectiveness of stakeholder engagement efforts (e.g., looking at whether engagement is timely, accessible, and appropriate and safe for stakeholders).
4. Include feedback of lessons learned into the enterprise's due diligence in order to improve the process and outcomes in the future.
5. Monitor and track AI system performance or assurance criteria qualitatively or quantitatively for conditions similar to deployment setting(s), for example, by:
  - a. Documenting test sets, metrics, and details about the tools used during Test and Evaluation, Verification and Validation (TEVV).
  - b. Identifying and documenting measurable performance improvements or declines based on consultations with relevant AI actors and other stakeholders, including affected communities, and field data about context relevant risks and trustworthiness characteristics.
  - c. Documenting and sharing information about incidents with stakeholders, including affected communities, governments, workers, workers' representatives and trade unions, civil society, and academia. This should be done with a view to advancing safety, security and trustworthiness of advanced AI systems.

- d. Documenting and sharing monitoring results regarding AI system trustworthiness in deployment context(s) and across the AI system lifecycle with domain experts and relevant AI actors and stakeholders to validate whether the system is performing consistently as intended.
6. Monitor and track implementation and effectiveness of the organisation's own internal commitments, activities and goals on due diligence (e.g., by carrying out periodic internal or third-party reviews or audits of the outcomes achieved and communicating results at relevant levels within the organisation).
7. Carry out periodic assessments of business relationships, to verify that risk mitigation measures are being pursued or to validate that adverse impacts have actually been prevented or mitigated.

## Step 5 - Communicate actions to address impacts

Communicate externally relevant information on due diligence policies, processes, activities conducted to identify and address actual or potential adverse impacts, including the findings and outcomes of those activities. Communication could take a variety of forms depending on the target audience (e.g., stakeholder consultations and public communication through the enterprise's annual, sustainability or corporate responsibility reports or other appropriate forms of disclosure required by legislation or voluntary initiatives).

**Table 2.6. Step 5: Roadmap of related provisions in existing frameworks**

ASEAN Guide	Section C.4. Stakeholder interaction and communication and Annex A:5
Australia Guidance for AI Adoption (Implementation Practices)	Implementation Practices 1, 2, 3, 4
Canada CoC	Transparency
CoE HUDERIA	Stakeholder Engagement Process (SEP)
EU AI ACT	Art. 13: Transparency and provision of information to deployers, Art. 53(1)(a) and (d), Art. 55(1)(c)
EU DSA	Art. 42: Transparency reporting obligations
EU CSDDD	Art. 16: Publicly communicate on due diligence
Hiroshima Process CoC	Principles 4 and 5
IEEE 7000	11: Transparency management process
ISO 31000 & ISO/IEC 23894	6.7: Recording and reporting
ISO/IEC 42001	7.4 Communication; Annex A.6.2.7, A.8.2, A.8.4, A.8.5.
Japan AI Guidelines for Business	Part 2C. Common Guiding Principles 6, 7; Appendix 3, 4, 5, B. Descriptions of "Common guiding principles" in Part 2. 6, 7
Korea AI Basic Act	Arts. 28 and 31
Singapore AI Verify Testing Framework	Transparency 1.1.1 – 1.5.1
UNGPs	Operational Principle 21
UK DSIT AI Assurance Framework	4.1.3: Communicate
US NIST AI RMF	Manage 4

### ***Practical examples for implementation (for all enterprises in the AI value chain, Groups 1-3)***

1. Publicly communicate all relevant information on due diligence processes, with due regard for commercial confidentiality, competition law, and other competitive or security concerns.<sup>12</sup> Include information on the following:

- a. RBC policies, per Step 1, including information on commitments to and implementation of relevant voluntary initiatives.
  - b. Processes for tracking, responding to, and recovering from significant incidents and errors.
  - c. Significant adverse impacts or risks identified, prioritised and assessed. For human rights impacts or other notable risks that the enterprise causes or contributes to, communicate with impacted or potentially impacted stakeholders in a timely, culturally sensitive and accessible manner, all information that is relevant to them, in particular when relevant concerns are raised by them or on their behalf.
  - d. Risk prioritisation criteria and processes
  - e. Actions taken or planned to prevent or mitigate risks, including where possible estimated timelines and benchmarks for improvement and their outcomes, such as details of the evaluations (including red-teaming) conducted for risks of adverse impacts
  - f. Measures to track implementation and results.
  - g. Provision of or co-operation in any remediation.
  - h. AI system capabilities, limitations and domains of appropriate and inappropriate use
  - i. Meaningful information for all new significant releases of AI systems that may be widely used
  - j. How relevant stakeholders are engaged in the design and implementation of these due diligence processes
  - k. Where relevant (e.g., for Group 2), information on incidents and attempts by AI actors to circumvent safeguards, in particular incidents related to general-purpose AI systems.
2. Disclose the above information in a way that is user friendly, regular, timely, reliable, clear, complete, accurate and with sufficient detail (see MNE Guidelines Chapter III: Disclosure for more detail).
  3. Ensure information is presented appropriately for different target audiences and may take special steps to make information available to vulnerable stakeholders (e.g., workers).

## Step 6 - Provide for or co-operate in remediation when appropriate

When an enterprise has caused or contributed to actual adverse impacts, seek to restore the affected person or persons to the situation they would be in had the adverse impact not occurred (where possible) and enable remediation that is proportionate to the significance and scale of the adverse impact.

**Table 2.7. Step 6: Roadmap of related provisions in existing frameworks**

ASEAN Guide	Section C.4: 4. Stakeholder interaction and communication
Australia Guidance for AI Adoption (Implementation Practices)	Implementation Practices 1 and 2
EU DSA	Art. 14: Terms and conditions
EU CSDDD	Art. 12: Provide remediation for actual adverse impacts
Singapore AI Verify Testing Framework	Transparency 1.4.1 – 1.5.1
UNGPs	Operational Principles 22, 29, 30, 31
UK DSIT AI Assurance Framework	5.3: Assuring data, models, systems, and governance in practice

### **Practical examples for implementation (for all enterprises in the AI value chain, Groups 1-3)**

1. When appropriate (i.e., in instances where the adverse impact is caused or contributed to by the enterprise), provide for or cooperate with legitimate remediation mechanisms through which impacted stakeholders can raise complaints and seek to have them addressed with the enterprise (see Box 2.17).
2. For adverse impacts directly linked to business relationships, enterprises are expected to apply leverage on the business relationship, in line with competition law, to provide for or cooperate with remediation mechanisms.

#### **Box 2.17. Potential options for remedying adverse impacts**

The ultimate goal of remediation is to restore the affected person or persons to the situation they would be in had the adverse impact not occurred. Effective remedy is context dependant and a number of potential options exist. Multiple options for remedy may be used depending on the scale and context of the harm inflicted. These options broadly include:

- Restitution: Restoring an affected person or group of people to the position they would have been had the harm not occurred.
- Compensation: Financial compensation to compensate for harm that can be economically assessed, which could include compensation for physical or mental harm, lost earnings, or for costs of expert help, such as medical costs.
- Rehabilitation: Providing medical and psychological care as well as legal and social services.
- Satisfaction: Acknowledgement of facts, decisions to restore the dignity of affected people and groups, acknowledgements that human rights were not respected, apologies, legal sanctions against the person responsible for the harm such as fines and prison sentences, commemorations.
- Guarantees of non-repetition: Measures that contribute to future prevention, including termination of the system, injunctions and changes to corporate policies, procedures and strategies.

Mechanisms that can support implementation of remediation take a variety forms, including:

- through direct action from the enterprise (in consultation with stakeholders), without recourse to a dispute resolution mechanism
- algorithmic audits by independent, multidisciplinary panels
- judicial mechanisms: Domestic and regional courts
- state-based non-judicial mechanisms: Mechanisms connected with the State which may have the potential to deliver remedies in some shape or form, such as National Contact Points for RBC, ombudspersons, inspectorates, and national human rights institutions; and
- non-State-based grievance mechanisms: Remediation mechanisms that are developed and administered by private entities such as companies or, in some cases, industry associations or multi-stakeholder groups (see UNGP 31 for more information on what makes a remediation mechanism effective (UN OHCHR, 2021<sup>[32]</sup>)).

Source: UN OHCHR B-Tech Project (2021<sup>[33]</sup>), Access to remedy and the technology sector: basic concepts and principles, <https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/access-to-remedy-concepts-and-principles.pdf>.

# References

- BSR (2022), *Sales partners and human rights due diligence in the technology sector: Best practices brief*, [https://www.bsr.org/reports/Sales Partner - Best Practice Brief.pdf](https://www.bsr.org/reports/Sales_Partner_-_Best_Practice_Brief.pdf). [27 ]
- Centre for the Governance of AI (2023), *Towards best practices in AGI safety and governance: A survey of expert opinion*, <https://arxiv.org/pdf/2305.07153>. [19 ]
- Coalition for Content Provenance and Authenticity (C2PA) (n.d.), *Guiding Principles for C2PA Designs and Specifications*, <https://c2pa.org/principles/>. [22 ]
- European Center for Not-for-Profit Law (2023), *Framework for Meaningful Engagement*, <https://oecd.ai/en/catalogue/tools/framework-for-meaningful-engagement-of-external-stakeholders-in-ai-development>. [38 ]
- European Union (2024), *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 an*, [https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L\\_202401689](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689). [14 ]
- G7 (2023), *Hiroshima Process International Code of Conduct for Advanced AI Systems*, <https://www.soumu.go.jp/hiroshimaaiprocess/en/documents.html>. [15 ]
- ILO (2023), *Tripartite Declaration of Principles concerning Multinational Enterprises and Social Policy*, <https://www.ilo.org/publications/tripartite-declaration-principles-concerning-multinational-enterprises-and-3>. [5 ]
- Institute for AI Policy and Strategy (IAPS) (2023), *Deployment corrections: An incident response framework for frontier AI*, [https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/651c397fc04af033499df9f8/1696348544356/Deployment+corrections\\_+an+incident+response+framework+for+frontier+AI+models.pdf](https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/651c397fc04af033499df9f8/1696348544356/Deployment+corrections_+an+incident+response+framework+for+frontier+AI+models.pdf). [24 ]
- ISO (2023), *ISO/IEC 27035-1:2023: Information security incident management*, <https://www.iso.org/standard/78973.html>. [25 ]
- ISO (2018), *ISO 31000:2018 Risk management — Guidelines*, <https://www.iso.org/standard/65694.html>. [36 ]
- NIST (2024), *NIST AI 600-1 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, <https://doi.org/10.6028/NIST.AI.600-1>. [41 ]
- OECD (2025), *RBC Spotlight: Due diligence in electronics and vehicle manufacturing*, [39 ]

- [https://www.oecd.org/en/publications/responsible-business-conduct-spotlights\\_03a75bf9-en/due-diligence-in-electronics-and-vehicle-manufacturing\\_993bb959-en.html](https://www.oecd.org/en/publications/responsible-business-conduct-spotlights_03a75bf9-en/due-diligence-in-electronics-and-vehicle-manufacturing_993bb959-en.html).
- OECD (2025), "Towards a common reporting framework for AI incidents", *OECD Artificial Intelligence Papers*, No. 34, OECD Publishing, Paris, <https://doi.org/10.1787/f326d4ac-en>. [28 ]
- OECD (2024), "AI, data governance and privacy: Synergies and areas of international co-operation", *OECD Artificial Intelligence Papers*, No. 22, OECD Publishing, Paris, <https://doi.org/10.1787/2476b1a4-en>. [20 ]
- OECD (2024), "Defining AI incidents and related terms", *OECD Artificial Intelligence Papers*, No. 16, OECD Publishing, Paris, <https://doi.org/10.1787/d1a8d965-en>. [34 ]
- OECD (2024), *Explanatory memorandum on the updated OECD definition of an AI system*, [https://www.oecd-ilibrary.org/science-and-technology/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system\\_623da898-en](https://www.oecd-ilibrary.org/science-and-technology/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en). [2]
- OECD (2024), "Explanatory memorandum on the updated OECD definition of an AI system", *OECD Artificial Intelligence Papers*, No. 8, OECD Publishing, Paris, <https://doi.org/10.1787/623da898-en>. [6]
- OECD (2024), *Recommendation of the Council on Artificial Intelligence*, <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>. [9]
- OECD (2023), "Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI", *OECD Digital Economy Papers*, No. 349, OECD Publishing, Paris, <https://doi.org/10.1787/2448f04b-en>. [17 ]
- OECD (2023), *Common Guideposts to Promote Interoperability in AI Risk Management*, <https://www.oecd-ilibrary.org/docserver/ba602d18-en.pdf?expires=1700737987&id=id&accname=ocid84004878&checksum=74F08ECC1794AF825ADFFE292B62FD50>. [10 ]
- OECD (2023), *OECD Guidelines for Multinational Enterprises on Responsible Business Conduct*, OECD Publishing, Paris, <https://doi.org/10.1787/81f92357-en>. [1]
- OECD (2022), *Measuring environmental impacts of AI compute and applications*, <https://www.oecd-ilibrary.org/docserver/7babf571-en.pdf?expires=1716997196&id=id&accname=ocid84004878&checksum=C13188D88E3A07543BCC1AAAD636B592>. [7]
- OECD (2022), "OECD Framework for the Classification of AI systems", *OECD Digital Economy Papers*, No. 323, OECD Publishing, Paris, <https://doi.org/10.1787/cb6d9eca-en>. [13 ]
- OECD (2022), *Report on the Implementation of the Recommendation of the OECD Council on Due Diligence Guidance for Responsible Supply Chains of Minerals from Conflict-Affected and High-Risk Areas*, <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0386>. [37 ]
- OECD (2022), *Responsible Business Conduct Due Diligence for Project and Asset Finance Transactions*, <https://doi.org/10.1787/952805e9-en>. [31 ]
- OECD (2018), *OECD Due Diligence Guidance for Responsible Business Conduct*, OECD Publishing, Paris, <https://doi.org/10.1787/15f5f4b3-en>. [3]

- OECD (2016), *OECD Due Diligence Guidance for Responsible Supply Chains of Minerals from Conflict-Affected and High-Risk Areas: Third Edition*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264252479-en>. [40]
- OECD (2015), *Competition Law and Responsible Business Conduct*, <https://mneguidelines.oecd.org/global-forum/2015GFRBC-Competition-Law-RBC.pdf>. [11]
- OECD (2015), *Competition Law and Responsible Business Conduct*, <https://mneguidelines.oecd.org/global-forum/2015GFRBC-Competition-Law-RBC.pdf>. [16]
- OECD (2013), *Guidelines governing the protection of privacy and transborder flows of personal data*, <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0188>. [18]
- OECD (n.d.), *Catalogue of tools and metrics for trustworthy AI*, <https://oecd.ai/en/catalogue/tools>. [12]
- OECD (n.d.), *Worldwide investments in AI*, <https://oecd.ai/en/data?selectedArea=investments-in-ai-and-data> (accessed on 1 October 2025). [29]
- Partnership on AI (2021), *Responsible Sourcing of Data Enrichment Services*, <http://partnershiponai.org/wp-content/uploads/2021/08/PAI-Responsible-Sourcing-of-Data-Enrichment-Services.pdf>. [8]
- Partnership on AI (n.d.), *Responsible Practices for Synthetic Media*, [https://syntheticmedia.partnershiponai.org/#read\\_the\\_framework](https://syntheticmedia.partnershiponai.org/#read_the_framework). [23]
- Solaiman, I. (2023), *The Gradient of Generative AI Release: Methods and Considerations*, <https://arxiv.org/pdf/2302.04844>. [26]
- UN OHCHR (2021), *OHCHR Accountability and Remedy Project: Meeting the UNGPs' Effectiveness Criteria*, <https://www.ohchr.org/sites/default/files/2022-01/arp-note-meeting-effectiveness-criteria.pdf>. [32]
- UN OHCHR B-Tech Project (2021), *Access to remedy and the technology sector: Basic concepts and principles*, <https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/access-to-remedy-concepts-and-principles.pdf>. [33]
- United Nations Office of the High Commissioner on Human Rights (2012), *United Nations Guiding Principles on Business and Human Rights*, <https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights>. [4]
- US Department of Commerce. National Telecommunications and Information Administration (2024), *AI Accountability Policy Report*, <https://www.ntia.gov/sites/default/files/publications/ntia-ai-report-final.pdf>. [21]
- US National Institute of Standards and Technology (2023), *AI Risk Management Framework 1.0*, <https://doi.org/10.6028/NIST.AI.100-1>. [35]
- World Benchmarking Alliance (2023), *Digital Inclusion Collective Impact Coalition 2023 Progress Report*, <https://www.worldbenchmarkingalliance.org/impact/digital-inclusion-collective-impact-coalition-progress-report/>. [30]

# Glossary

## AI system

Term as used in OECD standards and reports referred to in this guidance	The definition for an AI system is from the OECD Recommendation on AI that was updated in 2023 (OECD, 2024 <sup>[6]</sup> ). The Recommendation defines an AI system as a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment. The reasoning behind this definition is explained in detail in a memorandum published by the OECD (2024 <sup>[2]</sup> ).
Term as used by other risk management frameworks	The EU AIA, US NIST RMF, ASEAN Guide, and Council of Europe Convention all either directly adopt, refer to, or only slightly deviate from the definition used in the OECD Recommendation on AI.
Differences in terminology and application for this guidance	For the purposes of this guidance, an AI system can be understood according to the definition used in the OECD Recommendation on AI.

## Assess / Measure / Evaluate

Term as used in OECD standards and reports referred to in this guidance	The Interoperability Framework describes this step in the process as discovering risks, analysing the mechanisms by which those risks may occur, and evaluating their likelihood of occurring as well as their severity. The OECD DDG uses the same term in much the same way, but also recommending that businesses assess their linkage or contribution to impacts caused by business relationships. Included in this step is also the risk prioritisation process, where businesses prioritise the most significant risks and impacts for action, based on severity and likelihood.
Term as used by other risk management frameworks	Other risk management frameworks such as the NIST RMF, ISO 31000, IEEE 7000-21, and ISO/IEC Guide 51 refer to this step in the process as either measuring or evaluating. Under the NIST RMF, measuring risks includes using metrics for trustworthy characteristics and social impact to track risks. The UNGPs are aligned with the MNE Guidelines and OECD DDG (see Principles 17 and 24).
Differences in terminology and application for this guidance	These instruments use the terms in similar ways but with slightly different scopes. For example, the assessment could be in relation to trustworthiness objectives, to involvement with the impact (e.g., cause, contribute or directly linked), or to risks for the company. For the purposes of this guidance, the term 'ASSESS' can include all of these aspects.

\* The Interoperability Framework is derived from the paper on Advancing Accountability in AI which sets out common high-level steps for AI risk management that appear across multiple frameworks (OECD, 2023<sup>[17]</sup>).

## Define / Identify / Map / Scope

Term as used in OECD standards and reports referred to in this guidance	The Interoperability Framework* recommends to 'Define' the scope and context of an AI system and the criteria for evaluating risk e.g., at the governance level, process level, and/or technical level. The equivalent term used in the OECD DDG is 'Identify' and is essentially a broad scoping exercise to identify all areas of the business, across its operations, products and services, and also its relationships where risks are most likely to be present and significant.
Terms as used by other key risk management frameworks	Other risk management frameworks such as the NIST RMF and ISO 31000 refer to this process as 'Mapping' and 'Scoping', respectively. Similar to the Interoperability Framework, the NIST RMF offers specific recommendations on what needs to be mapped, e.g., the purpose and uses of the AI system, risks relating to components of the AI system, including its software and the data it uses, as well as its impacts on individuals, groups and society.

	The UNGPs are aligned with the MNE Guidelines and OECD DDG (see Principle 15).
Differences in terminology and application for this guidance	Different instruments use the terms 'define', 'identify', 'map' and 'scope' to generally refer to the same broad set of actions. The 'MAP' function of the NIST RMF seems to be the most specific and applicable to AI systems. Likewise, all the risk management frameworks describe the 'Define' (or 'Identify' / 'Map' / 'Scope') as a distinct step in the risk management process that is essential as a foundation to assess risks and impacts. For the purposes of this guidance, the term 'DEFINE' is considered to include all of these aspects.

## Due diligence

Term as used in OECD standards and reports referred to in this guidance	Due diligence is understood in the MNE Guidelines as the process through which enterprises can identify, prevent, mitigate and account for how they address their actual and potential adverse impacts as an integral part of business decision-making and risk management systems. The MNE Guidelines outline the following measures as part of a due diligence process: 1. embedding RBC into policies and management systems; 2. identifying and assessing actual and potential adverse impacts associated with the enterprise's operations, products or services; 3. ceasing, preventing and mitigating adverse impacts; 4. tracking implementation and results; 5. communicating how impacts are addressed; and 6. providing for or co-operating in remediation when appropriate.
Term as used by other risk management frameworks	The UN Guiding Principles on Business and Human Rights (UNGPs), developed alongside the 2011 update of the MNE Guidelines and are aligned with the latter on the term due diligence. The UNGPs define due diligence as a process that includes "assessing actual and potential human rights impacts, integrating and acting upon the findings, tracking responses, and communicating how impacts are addressed" (OECD, 2023 <sup>[17]</sup> ). Likewise, the UNGPs also extend the due diligence expectation to impacts that enterprises cause, contribute to or are directly linked to through their business relationships. Other risk frameworks usually refer to or adapt the ISO 31000 definition of "risk management" to refer to processes similar or related to the due diligence expectations of the MNE Guidelines. ISO 31000 provides that "[r]isk management refers to coordinated activities to direct and control an organisation with regard to risk."
Differences in terminology and application for this guidance	Generally, the concepts related to due diligence are aligned across frameworks. The key difference between the terminology is the explicit reference to due diligence pertaining to enterprises' business relationships in the MNE Guidelines and UNGPs, rather than focusing solely on their own operations, products and services. For the purpose of this guidance, due diligence can be understood as the more expansive term used in the MNE Guidelines / OECD DDG and UNGPs, which includes management of risks in an organisation's own operations, products, and services, as well as in an organisation's business relationships.

## Monitor and review / Track

Term as used in OECD standards and reports referred to in this guidance	'Monitoring and reviewing' are understood as a continuous activity to check risks and the steps taken to treat them. The equivalent term used in the OECD DDG is "Track". Specifically, "tracking" implementation and effectiveness and outcomes of due diligence activities (i.e., measures to identify, prevent, mitigate and remedy impacts), including with business relationships. Tracking is conducted on a periodic basis and can come in the form of independent, third-party audits.
Term as used in other key risk management frameworks	Other frameworks refer to monitoring and review in similar ways. In some cases, it is a distinct step of the due diligence process (e.g., in ISO 31000) and in other cases, it is part of a broader step focused on internal governance (e.g., in the NIST Risk Management Framework (NIST RMF)). The UNGPs are aligned with the MNE Guidelines and OECD DDG (see Principles 17 and 20).
Differences in terminology and application for this guidance	Generally, the concepts are equivalent across frameworks.

## Risks / Incidents / Hazards

Term as used in OECD standards and reports referred to in this guidance	OECD RBC instruments describe adverse impacts or potential adverse impacts (i.e., risks) in the context of topics covered in the chapters of the MNE Guidelines: human rights, including workers and industrial relations, environment, bribery and corruption, disclosure, and consumer interests. The MNE Guidelines refer to risk as the likelihood of adverse impacts on people, the environment and society that enterprises cause, contribute to, or to which they are directly linked. In other words, it is an outward-facing approach to risk. The likelihood of adverse impacts increases in situations where an enterprise's behaviour or the circumstances associated with their supply chains or business relationships are not consistent with the recommendations in the MNE Guidelines. A risk of adverse impacts may exist when there is the potential for behaviour that is inconsistent with the recommendations in the MNE Guidelines because it involves impacts that may occur in the future.
---	--

	<p>In a separate, but related workstream on monitoring AI incidents, the OECD Network of Experts is working towards developing a definition of an AI incidents and hazards (OECD, 2024<sup>[34]</sup>), which are defined as follows:</p> <p>An AI incident is an event, circumstance or series of events where the development, use or malfunction of one or more AI systems directly or indirectly leads to any of the following harms: (a) injury or harm to the health of a person or groups of people; (b) disruption of the management and operation of critical infrastructure; (c) failure to respect human rights or a breach of obligations under the applicable law intended to protect labour and intellectual property rights; (d) harm to property, communities or the environment.</p> <p>An AI hazard is an event, circumstance or series of events where the development, use or malfunction of one or more AI systems could plausibly lead to an AI incident, i.e., any of the following harms: (a) injury or harm to the health of a person or groups of people; (b) disruption of the management and operation of critical infrastructure; (c) failure to respect human rights or a breach of obligations under the applicable law intended to protect labour and intellectual property rights; (d) harm to property, communities or the environment.</p>
Terms as used by other risk management frameworks	<p>The EU AIA refers to risks as the combination of the probability of an occurrence of harm and the severity of that harm (European Union, 2024<sup>[14]</sup>), with “harm” being understood as harm to public interests and fundamental rights that are protected by European Union law. Such harm might be material or immaterial, including physical, psychological, societal or economic harm.</p> <p>The US NIST RMF adopts the same approach as ISO 31000 in framing risk as both positive and negative (US National Institute of Standards and Technology, 2023<sup>[35]</sup>) (ISO, 2018<sup>[35]</sup>). It refers to risk as the composite measure of an event’s probability of occurring and the magnitude or degree of the consequences of the corresponding event. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats.</p>
Differences in terminology and application for this guidance	<p>The definition of risk in the MNE Guidelines is broader in scope in that it includes impacts that enterprises “can cause, contribute to or to which they are directly linked.” Essentially, this expands relevant risks and impacts beyond own operations and into the realm of impacts and risks associated with business relationships in the value chain. The scope of specific risks covered also varies across frameworks.</p> <p>For the purpose of this guidance, risk can be understood using the definition from the MNE Guidelines as it includes AI Incidents and hazards, but is more expansive for the purposes of value chain due diligence. Where relevant and more specific, the guidance refers to incidents and hazards.</p>

## Stakeholders

Term as used in OECD standards and reports referred to in this guidance	<p>The OECD Recommendation on AI defines stakeholders as all organisations and individuals involved in, or affected by, AI systems, directly or indirectly.</p> <p>The OECD MNE Guidelines describes relevant stakeholders as persons or groups, or their legitimate representatives, who have rights or interests related to the matters covered by the MNE Guidelines that are or could be affected by adverse impacts associated with the enterprise’s operations, products or services.</p>
Term as used by other risk management frameworks	<p>Other frameworks use the term stakeholders to generally refer to organisations and individuals. In different contexts, “relevant stakeholders” usually refers to users of the AI system, civil society, workers’ representatives, SMEs, and other enterprises.</p>
Differences in terminology and application for this guidance	<p>For the purposes of this guidance, stakeholders should be understood in the broadest sense as persons, groups or organisations involved in or affected by AI systems and the enterprises involved in their development and use.</p>

## Treat / Cease, prevent, mitigate and remedy / Manage

Term as used in OECD standards and reports referred to in this guidance	<p>The Interoperability Framework defines ‘risk treatment’ as using techniques to prevent, mitigate or cease risks, based on their likelihood and impact.</p> <p>The OECD due diligence recommendations take a similar approach to risk treatment, but further specifies the types of action to be taken, based on a company’s responsibility in causing the risk (e.g., whether they caused, contributed or were directly linked to the risk). Companies are expected to cease, prevent, and/or mitigate identified risks and impacts.</p> <p>In circumstances where a company is contributing to or directly linked to an impact through a business relationship, it should seek, to the extent possible, to use its leverage, individually or in collaboration with others, to effect change.</p> <p>When a company has caused or contributed to an adverse impact, the company is expected to provide for, or co-operate in, remediation (see further discussion on the term “REMEDY” below).</p>
Term as used by other risk management frameworks	<p>The NIST RMF refers to risk treatment under its MANAGE function as “plans to respond to, recover from, and communicate about incidents or events.” Although it uses different terminology, the set of actions under this</p>

	function is generally consistent with OECD due diligence recommendations. Expected actions include prioritising and allocating resources to risk management, engaging with impacted stakeholders, and continuously monitoring and documenting risk management efforts. The NIST RMF also notes that risk response options can include mitigating, transferring, avoiding, or accepting.
Differences in terminology and application for this guidance	For the purposes of this guidance, 'Treat' can be understood to mean ceasing, preventing or mitigating risk that companies cause, contribute to, or are directly linked to through their business relationships.

# Notes

<sup>1</sup> The Recommendation on AI identifies five complementary values-based principles relevant to all stakeholders and five recommendations to policymakers, referred together in this guidance as the “AI Principles”.

<sup>2</sup> According to the MNE Guidelines Ch. I, paragraph 4, “A precise definition of multinational enterprises is not required for the purposes of the Guidelines. While the Guidelines allow for a broad approach in identifying which entities may be considered multinational enterprises for the purposes of the Guidelines, the international nature of an enterprise’s structure or activities and its commercial form, purpose, or activities are main factors to consider in this regard.”

<sup>3</sup> An AI system lifecycle typically involves several phases that include to: plan and design; collect and process data; build model(s) and/or adapt existing model(s) to specific tasks; test, evaluate, verify and validate; make available for use/deploy; operate and monitor; and retire/decommission. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase. The AI Principles refer to enterprises in the AI system lifecycle as “AI actors” (OECD, 2024<sup>[2]</sup>).

<sup>4</sup> The grouping based on activities was developed in the OECD report on *Advancing accountability in AI* (OECD, 2023<sup>[17]</sup>) and further detailed in a follow up draft report *Draft mapping and consolidation of relevant actors, issues, and terminology for Responsible Business Conduct in AI* [DSTI/CDEP/AIGO(2023)12], which was discussed at the November 2022 AIGO and WPRBC meetings.

<sup>5</sup> For more information on addressing risks associated with raw materials, see the OECD Due Diligence Guidance for Responsible Supply Chains for Minerals (OECD, 2016<sup>[39]</sup>); for more on addressing risks related to electronics and vehicle manufacturing see the RBC Spotlight on Due Diligence in Electronics and Vehicle Manufacturing (OECD, 2025<sup>[38]</sup>).

<sup>6</sup> The term ‘deploy’ is used differently in the AI Principles than in the EU AI Act. Under the AI Principles, deployment can be understood as making the AI system available for use. Under the EU AI Act, “deployer means a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity” (European Union, 2024<sup>[14]</sup>). The EU AI Act definition of deployer is more akin to what this guidance describes as Group 3: Users of the AI system.

<sup>7</sup> NCPs have the mandate of furthering the effectiveness of the MNE Guidelines by undertaking promotional activities, handling enquiries and contributing to the resolution of issues that arise relating to the implementation of the MNE Guidelines in specific instances. Any individual or organisation can bring a specific instance (case) against an enterprise to the NCP where the enterprise is operating or based

regarding the enterprise's operations anywhere in the world. NCPs facilitate access to consensual and non-adversarial procedures, such as conciliation or mediation, to assist the parties in dealing with the issues. NCPs are required to issue final statements upon concluding the specific instance processes. NCPs can also make recommendations based on the circumstances of the specific instance. Organisations can refer to the Responsible Business Conduct OECD Guidelines for Multinational Enterprises (OECD, 2025<sup>[38]</sup>) for information on the NCP process, specific NCPs or cases.

<sup>8</sup> The MNE Guidelines also note that “Enterprises should pay special attention to any particular adverse impacts on individuals, for example human rights defenders, who may be at heightened risk due to marginalisation, vulnerability or other circumstances, individually or as members of certain groups or populations, including Indigenous Peoples. OECD due diligence guidance, including the OECD Due Diligence Guidance on Responsible Business Conduct, the OECD Due Diligence Guidance on Meaningful Stakeholder Engagement in the Extractive Sector, and the OECD-FAO Guidance for Responsible Agricultural Supply Chains provides further practical guidance in this regard, including in relation to Free, Prior and Informed Consent (FPIC). United Nations instruments have elaborated on the rights of Indigenous Peoples (UN Declaration on the Rights of Indigenous Peoples).” (Commentary 45).

<sup>9</sup> To be meaningful, stakeholder engagement should be two-way, conducted in good faith and responsive to stakeholders' views. Stakeholders should be provided with truthful and complete information and should be given timely opportunity to provide input prior to major decisions being made that may affect them (see (OECD, 2018<sup>[3]</sup>)).

<sup>10</sup> For more information on how to meaningfully engage stakeholders when designing AI powered products and services, see the ECNL Framework for meaningful engagement of external stakeholders in AI development (European Center for Not-for-Profit Law, 2023<sup>[37]</sup>).

<sup>11</sup> Examples include the OECD AI Incidents Monitor (OECD, 2025<sup>[28]</sup>) and the OECD National Contact Point Specific Instance Database (OECD, 2022<sup>[36]</sup>).

<sup>12</sup> Disclosure should not place unreasonable administrative or cost burdens on enterprises. Nor should enterprises be expected to disclose information that may endanger their competitive position unless disclosure is necessary to fully inform an investor's decisions and to avoid misleading investors.

# OECD Due Diligence Guidance for Responsible AI

This report provides practical guidance to enterprises for implementing OECD standards on responsible business conduct (RBC) and the OECD AI Principles when developing and using artificial intelligence (AI). It aims to support innovation, investment and growth of enterprises in the AI value chain by helping enterprises proactively address adverse impacts. The report promotes policy coherence, and where possible interoperability, between the OECD and other national or international AI risk management frameworks.



PRINT ISBN 978-92-64-31703-1  
PDF ISBN 978-92-64-31822-9



9 789264 317031