unesco

Ethical Impact Assessment

A Tool of the Recommendation on the Ethics of Artificial Intelligence

SHS/REI/BIO/REC-AIETHICS-TOOL-EIA/2023

Published in 2023 by the United Nations Educational, Scientific and Cultural Organization
7, place de Fontenoy, 75352 Paris 07 SP, France

Cover photo: voronaman/Shutterstock.com; Artistdesign29/Shutterstock.com

Designed and printed by UNESCO

Printed in France
https://doi.org/10.54678/YTSA7796

unesco

Ethical Impact Assessment

A Tool of the Recommendation on the Ethics of Artificial Intelligence

Table of Contents

Foreword

The opportunities that artificial intelligence offers our economies, labour markets and lives are immense and continue to amaze us. Yet hardly a week goes by without news alerting us to real, indisputable risks that arise from the use, or misuse, of AI technology. This may be the result of AI developments being deployed prioritizing commercial or geopolitical interests rather than protecting and promoting human rights and human dignity.

Generative AI products, for example, were released to the market before full assurances were made about their safety and trustworthiness. They were quickly adopted by millions of people around the world but were nevertheless capable of delivering racist or biased outcomes. Questions pertaining to transparency or accountability were neither addressed in an ex-ante manner.

To avoid a backlash against these outcomes and to allow innovation to continue to flourish, AI should be developed in line with the common good in an ethical manner. This is the call that UNESCO has made since 2021, when its 193 Member States adopted the first global standard on the ethics of AI.

The Recommendation on the Ethics of AI provides a framework to ensure that AI developments align with the promotion and protection of human rights and human dignity, environmental sustainability, fairness, inclusion and gender equality. It underscores that these goals and principles should inform technological developments in an ex-ante manner. To support effective implementation, UNESCO developed two instruments, the Readiness Assessment Methodology and the Ethical Impact Assessment.

In this publication, we are pleased to unveil the Ethical Impact Assessment. This instrument has two goals: First, to assess whether specific algorithms are aligned with the values, principles and guidance set up by the Recommendation. And second, to ensure transparency by calling for information about AI systems and the way they were developed to be available to the public. This is not how it works today, even for basic information about AI safety and reliability.

Impact Assessment tools are gaining ground to assess the true impact of AI systems. In fact, impact assessments are mandated by the draft EU AI Act for high-risk systems, and they are proposed as part of the Council of Europe's discussion on a Convention for AI.

The UNESCO Recommendation is unique in that it considers the entire AI lifecycle. The Ethical Impact Assessment therefore includes ex-ante and ex-post requirements. At an early stage, it establishes the importance of ensuring quality and representativeness data, the diversity of the teams developing the products, the robustness and transparency of the algorithms, their auditability, and the possibility of inserting check points at different moments of the development process.

The EIA is proposed to procurers of AI systems, as this is one of the main channels in which algorithms make their way to highly sensitive public domains. But the questions and the structure of the document are designed so the tools can also be used more generally by developers of AI systems, in the public or private sectors, who wish to develop AI ethically and fully comply with international standards such as the Recommendation.

The document comprises two main parts that together strike a balance between procedure and substance. In the first part, related to scoping, the goal is to understand the basics of the system, as well as to lay out some preliminary questions, such as whether automation is the best solution for the case at hand. It also raises questions about the project team and whether plans are in place to engage different stakeholders. The second part is dedicated to implementing the principles in the UNESCO Recommendation. For each principle, questions will aim to assess:

a. Whether sufficient procedural safeguards have been put in place to ensure the system complies with the Recommendation; and
b. The (potential) positive outcomes and adverse impacts that may arise from the procurement and deployment of the system, specific to its context of use.

The EIA is part of a larger implementation plan for the Recommendation, and it complements another tool produced by UNESCO, the Readiness Assessment Methodology (RAM). The RAM helps governments assess how robust and agile their laws, policies and institutions are in addressing AI risks. It is a diagnostic tool and the first step for targeted capacity building to strengthen institutional and human capacities in government to deal with AI.

We hope that these two tools, and more broadly the work to implement the Recommendation on the Ethics of AI, will provide the basis for a trustworthy environment where AI technology can flourish safely and responsibly.

Gabriela Ramos
Assistant Director-General for Social and Human Sciences, UNESCO

The UNESCO Ethical Impact Assessment

What is this instrument for?

As stated in article 50 of UNESCO's Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021), the goal of this instrument is to "*identify and assess benefits, concerns and risks of AI systems, as well as appropriate risk prevention, mitigation, redressal and monitoring measures, among other assurance mechanisms. Such assessments should identify the impacts on human rights and fundamental freedoms, in particular but not limited to the rights of people in vulnerable and precarious situations, labour rights, the environment and ecosystems and ethical and social implications, and facilitate citizen participation in line with the values and principles set forth in this Recommendation*".

As such, this is a general methodology, intended as a set of criteria for how to conduct an ethical impact assessment (EIA). It is important to note that it is not a universal tool and therefore it will need to be adapted to the specific circumstances it is used in and to the regulatory regime in each country.

Who should use it?

This instrument is primarily designed to help government officials (individuals and teams) involved in the procurement of AI systems. The goal of the methodology is to equip procurement officers with the set of questions to ask in order to ensure that the AI systems they are purchasing are aligned with the ethical standards set out in the UNESCO Recommendation on the Ethics of AI. By implementing the methodology, procurement teams can acquire more information about the systems and ensure greater transparency. More generally, the EIA can also be used outside procurement procedures, to assess if any AI system is in accordance with the UNESCO Recommendation, for example it can be used by companies that are wishing to develop AI ethically, if they are in line with the standards of the Recommendation.

We recognize that the ability of procurers to fill out the different sections of the EIA, and to do so comprehensively, may be limited due to their position within the AI system lifecycle and breadth and depth of the knowledge they have access to. We also note that this ability varies between contexts and use cases. That said, **procurers should aim to fill out the assessment as comprehensively as possible, before assigning questions to and consulting with other parties where they lack requisite information.**

Since procurement teams are often not directly involved in the design and development of AI systems and may be unable to answer some questions, it is essential to hold open discussions involving the AI provider and other individuals further upstream in the AI lifecycle when completing the EIA. How this might be approached is left to the procurers' discretion.

For example, the EIA can also be used as part of the procurement process, making up parts of the questionnaire sent to bidders. The responses of bidders may therefore form a useful additional criterion in the tender process. Please refer to Fig. 1 for further examples and guidance on this.

Figure 1: Guidance on the division of responsibilities when filling in the EIA

There is no single way of filling out the ethical impact assessment. How it is completed will depend on the situation faced by the procuring team involving numerous factors, for example: whether the AI system being procured is off-the-shelf or custom-made, who owns the data used for training and prediction, and the breadth of the procurement exercise.

Therefore, UNESCO cannot offer a single set of instructions regarding who should fill out the various parts of the EIA. This will remain at the discretion of the procuring team, who can adapt it to their specific situation. The two examples below provide illustrations of how roles might be allocated depending on the situation:

Example 1:

Context: A team runs a procurement process to procure an AI system but do not own the underlying data it would rely on, nor do they know precisely what the system will look like.

The procurement team can likely fill out the following parts of the EIA in the early stages of the procurement process:

- The "project description" section (excluding the "dependencies" sub-section);
- The "proportionality screening and do no harm" section;
- Parts of the "project governance" section (excluding the "multi-stakeholder engagement" section).

Most other sections could either be sent out as part of the tendering process to potential suppliers to fill out (and could feed into the selection criteria where relevant) or completed at a later stage once a supplier is chosen. Additionally, the procuring team can use any of the categories to provide potential suppliers with extra information regarding the specifications they are looking for in terms of ethical standards of the AI system they intend to purchase (for example, requirements regarding stakeholder engagement).

Example 2:

Context: A team runs a procurement process to develop an AI system involving image recognition using its own database of images. Furthermore, the team started the procurement process following an initial consultation exercise with relevant stakeholders.

In this case, the procuring team can fill out everything that the procuring team filled out in Example 1, as well as:

- The "multi-stakeholder engagement" sub-section within the "project governance" section. If further stakeholder engagement is planned, this section can be updated later;
- Parts of the "Data quality and discriminatory bias prevention" sub-section within the "Fairness, non-discrimination, diversity" section;
- Most of the "privacy and data protection" section.

As in Example 1, other parts of the EIA can be sent out to suppliers to fill in during the tender process or once the project is developed.

With the above in mind, and for the sake of simplicity, please note that the term "project team" is henceforth used to refer broadly to both individuals involved in procurement roles and, where relevant, other individuals involved in the design and development of the AI system (e.g., project manager, administrators, technical team, etc.). The specificities of the division of work and assignment of EIA questions to answer are subject to the procurers' (and Member States') judgement.

It is worth noting that an aim of the EIA is to provide space for procurers to reflect on important questions and consider the ethical issues associated with AI systems. This should guide them to identify and address any gaps, including by implementing mitigative actions such as adjustments to the design of the AI system or incorporating relevant requirements into contracts with suppliers of AI systems as appropriate. The level of detail for answers in the EIA, the level of stakeholder engagement and the extent of mitigative actions should be proportionate to the scale and scope of the project, its urgency and expected impacts.

Another core goal of the EIA is to promote transparency in the governance of AI systems. It can also play an important role in raising awareness of ethical issues relating to AI. Therefore, governmental entities are encouraged to make the completed EIA publicly available.

Contents

The EIA is deeply rooted in ethics and its added value to the existing impact assessment space is that it incorporates ethics by design throughout the assessment. As the space of impact assessment tools for AI-based systems is already rather crowded, the EIA builds on the existing work of various reputable organizations instead of seeking to replace them: where applicable, some parts of the assessment borrow from existing tools or methods, referring to and crediting them in the process. In this way, the aim of producing this tool was not to reinvent the wheel, but rather to leverage pre-existing knowledge, fill existing gaps and create a cohesive methodology that emphasizes ethics throughout.

The EIA is composed of two main chapters.

1. **Scoping questions:** These questions assess the fundamentals of the AI project and whether you and your team are in a position to continue with the rest of the EIA. To do so, you should have established that the AI project you wish to procure is not prohibited by the Recommendation, that your approach is proportionate to your intended aims and that your plans to involve stakeholders in the project are in line with guidelines set out in UNESCO's Recommendation.
2. **Implementing the UNESCO principles:** This section assesses whether the design, development and deployment of the AI system you wish to procure will result in processes and outcomes consistent with the UNESCO principles for ethical AI. For each principle, questions will aim to assess:

    a. Whether sufficient procedural safeguards have been put in place to ensure this system is compliant with the Recommendation.
    b. The (potential) positive outcomes and adverse impacts that may arise from the procurement and deployment of the system, specific to its context of use. **For detailed guidance on how to fill in the tables in this sub-section, please refer to the Annex.**

Iterative Requirements – the AI lifecycle:

While the Recommendation does not aim to provide a single definition of AI, it approaches AI systems as systems that have the capacity to process data and information in a way that resembles intelligent behaviour, and typically include aspects of reasoning, learning, perception, prediction, planning or control. AI systems integrate models and algorithms that produce a capacity to learn and to perform cognitive tasks leading to outcomes such as prediction and decision-making in material and virtual environments (UNESCO, 2021).

According to the Recommendation, the AI lifecycle is understood to "range from research, design and development to deployment and use, including maintenance, operation, trade, financing, monitoring and evaluation, validation, end-of-use, disassembly and termination."

Since this tool is designed primarily for procurement purposes, the project team should begin to engage with it early in the AI lifecycle. It is important to ensure that the impact assessment is initiated during the "design" phase of an AI lifecycle in order to guarantee ethical practices are established from the outset.

The EIA is intended to be a living document that will be filled out progressively and iteratively at different stages including:

- During project research, design, development and pre-procurement (for example, to reflect on the scope of the project, its legitimate aims and whether AI is an appropriate solution);
- During the procurement process itself, when the EIA can help both in selecting a supplier and in formulating contractual obligations;
- Following project deployment, the EIA should be revisited at regular intervals, especially since the answers may change over time as the project evolves.

Scoping Questions

Prior to conducting an in-depth assessment of how an AI project aligns with the UNESCO Recommendation, it is important to establish strong ethical foundations for the procurement or development process. This introductory section helps project teams to do so through four core sections:

1. **Project description:** Here, the project team specifies the details of the system they plan to implement and the goal they hope to achieve. This helps to clarify the details of the system and the scope of the EIA. It also helps to ensure the plan is well suited to a real-world problem or challenge.
2. **Proportionality screening:** The UNESCO Recommendation contains some fundamental provisions which should be established before moving on to the next stage of impact assessment. Specifically, the system should not be used for social scoring or mass surveillance and the choice to use AI must be justified based upon whether the method is proportional to achieving the stated aim.
3. **Project governance:** To ensure an effective EIA, clear roles and responsibilities must be defined from the outset. Diversity of project team must be ensured as it will have an impact on the quality of the end product.
4. **Multi-stakeholder governance:** Plans must also be established to ensure diverse perspectives are included in the EIA, by incorporating the perspectives, views and experiences of those who will be impacted by the system.

1. Project Description

1.1. Description of system:

In this first stage of impact assessment, you will be asked to provide a description of the AI system being assessed and to specify the context in which it will operate. The full project team should collaborate in filling out this section, as certain details may only be known by the AI provider.

1.1.1 Please provide an initial description of the AI system you intend to design, develop or deploy:

1.1.2 Please describe the aim or objective of this system. If the aim is to address a specific problem, please specify the problem you are trying to solve. Please also specify how this system may fit within broader schemes of work:

1.1.3 Please describe the current status of your project, with reference to the project lifecycle:

1.1.4 Using the table below, please specify the following features of your AI system. Please note that this part is based on the OECD framework for classifying AI systems (OECD, 2022):

1.1.4.1. Who will the users who interact with your system be (include their level of competency)?	**Free text**
1.1.4.2. What degree of optionality will users have?	**Multiple choice, multiple selection possible:** Users cannot opt out of the AI system's output / Users can opt out of the AI system's output/ Users can challenge or correct the AI system's output / Users can reverse the AI system's output ex-post/ Other

1.1.4.3. What is the sector where this ai system will be applied?	**Multiple choice, multiple selection possible:** General purpose/ Agriculture, forestry and fishing/ Mining and quarrying/ Manufacturing/ Electricity, gas, steam and air conditioning supply / Water supply, sewerage, waste management and remediation activities / Construction / Wholesale and retail trade; repair of motor vehicles and motorcycles / Transportation and storage / Accommodation and food service activities / Information and communication / financial and insurance activities / Real estate activities / Professional, Scientific and technical activities / Administrative and support service activities / Public administration and defence; compulsory social security / Education / Human health and social work activities / Arts, entertainment and recreation / Other service activities / Activities of households as employers, undifferentiated goods and services-producing activities of households for own use / Activities of extraterritorial organisations and bodies/ Other
1.1.4.4. In what business function will this ai system be employed?	**Multiple choice, multiple selection possible:** General purpose/ Human resource management / Sales / ICT management and information security / Marketing and advertisement / Logistics / Citizen / Customer service / Procurement / Maintenance / Accounting / Monitoring and quality control / Production / Planning and budgeting / Research and development / Compliance and justice / Other
1.1.4.5. Impacts on critical functions and activities	**Multiple choice:** AI system deployed in a critical sector or infrastructure (e.g., energy, transport, water, health, digital infrastructure and finance)/ AI system performs or serves a critical function independent from its sector (e.g., conducting elections, maintaining supply chains, law enforcement, providing medical care, supporting the financial system) / Neither of the above, other
1.1.4.6. Please describe the breadth of deployment (i.E., Is the ai deployment a pilot, narrow, broad or widespread)	**Multiple choice, multiple selection possible:** Pilot / Narrow / Broad / Widespread / Other (with free text)

1.2. Dependencies

1.2.1. Is this project an expansion or adaptation of any existing project? If so, has a previous assessment been done? If so, what features of the system have changed since this initial assessment? (The Information Accountability Foundation, 2019)

1.2.2. Is the AI system, including the core model, developed for this specific aim or objective or is it built upon an off-the-shelf model (e.g., BERT, ChatGPT, etc.)?

1.2.3. Please list relevant dependencies of the system on other models not directly developed or data not directly used, but which may have an impact on the ethical impact assessment (e.g. specific machine learning packages or pre-trained models):

Dependency	Corresponding Risk(s)
E.g. pre-trained word embedding model	*E.g model reproduces biases in the training data which are not within control of this project*

Project teams are encouraged to factor the dependencies of the system when answering the rest of this methodology and can use this table as a primer for filling up the impact assessment tables within the 'Implementing the UNESCO Principles' section.

2. Proportionality Screening and Do No Harm

2.1. Why is this important?

When seeking to procure AI, it is imperative to consider the objectives of using AI and the specific system in question, as well as the proportionality of the technology in terms of whether the intended purpose warrants its usage, considering the risks, uncertainty and downsides of the technology (Commission Nationale de l'Informatique et des Libertés, 2022). Such reflection allows procurers to strike a balance between the means and the intended aim in an exercise to justify the *necessity* of using a particular method or system and demonstrating its *suitability,* ensuring that processes that are related to, or part of, the AI system do not exceed what is necessary to achieve legitimate aims (European Data Protection Supervisor, n.d.; UNESCO, 2021). Importantly, the Recommendation stresses that "*any possible limitations on human rights and fundamental freedoms must have a lawful basis, and be reasonable, necessary and proportionate, and consistent with States' obligations under international law.*" Additionally, while all the principles and values expounded in the Recommendation are important and desirable, in practice, they may sometimes come into conflict – this may arise for instance when the need for transparency and explainability may impact the ability to preserve privacy and data protection (Whittlestone et al., 2019). The principle of proportionality can therefore also play a crucial part in helping, when needed, to contextually reconcile tensions between different ethical principles and/or priorities, while still respecting human rights and fundamental freedoms (Karliuk, 2022).

2.2. Establishing Proportionality

Proportionality is a key principle in the Recommendation. The Recommendation states: *"The choice to use AI systems and which AI method to use should be justified in the following ways: (a) the AI method chosen should be appropriate and proportional to achieve a given legitimate aim; (b) the AI method chosen should not infringe upon the foundational values captured in this document, in particular, its use must not violate or abuse human rights; and (c) the AI method should be appropriate to the context and should be based on rigorous foundations."*

2.2.1. Has careful consideration been given to non-algorithmic options which may be used to achieve the same goal? If so, why is the option involving an AI system favoured?

2.2.2. Were different AI methods considered, including computationally simpler approaches? What was the rationale behind choosing this specific method?

2.2.3. Has the scope of this project been clearly defined? What limitations have been placed on the scope of this project to ensure it remains proportional to the stated objective?

2.2.4. [For ex-post analysis] How effective has the system been in achieving its stated aim?

2.3. Do No Harm

The Recommendation prohibits several uses of AI: *"In scenarios where decisions are understood to have an impact that is irreversible or difficult to reverse or may involve life and death decisions, final human determination should apply. In particular, AI systems should not be used for social scoring or mass surveillance purposes."*

2.3.1. Is the system intended to be used for social scoring? Could the system be adapted for social scoring by other actors? If so, have measures been put in place to safeguard against this?

2.3.2. Is the system intended to be used for mass surveillance? Could the system be adapted for mass surveillance by other actors? If so, have measures been put in place to safeguard against this?

2.3.3. Are the expected impacts irreversible or difficult to reverse or could they involve life and death decisions? (e.g., setting prison sentences or determining medical treatments)

2.3.4. Could the AI system and its application impact fundamental human rights (e.g., human dignity, freedom of expression, fair trial)?

3. Project Governance (establishing roles and responsibilities)

3.1. Why is this important?

It is crucial to ensure that actors are identified for transparency and to avoid any confusing diffusion of responsibility within the project team. As AI comes with inherent risk, it is important to determine who has responsibility over which aspect of the AI system.

Furthermore, project teams should be especially vigilant about poor representation of stakeholders including prospective users, and particularly those from marginalized communities. A lack of diversity within the project team means that certain perspectives may be missing, which may contribute to greater experienced harm for unrepresented or underrepresented communities. In contrast, more diversity within project teams may allow for the early identification of biases and mitigation of harms. Further, allowing users to provide feedback to contribute to model development is critical, as users are generally more diverse than developers and may notice these concerns earlier (Bommasani et al., 2021).

3.2. Roles and responsibilities

3.2.1. Who has ultimate decision-making authority within the project team responsible for this AI system?

3.2.2. Please describe who has responsibility for the major workstreams within this project, including any representatives of third-party or external organisations. Include a full description of roles and responsibilities within the team, and a map of different individuals and organisations involved.

3.2.3. Has consideration been given to the diversity of the AI project team, especially in terms of – but not limited to – gender, age, race, colour, descent, language, religion, national origin, ethnic origin, social origin, economic or social condition, disability, and sexual orientation, including how this reflects the complexity and diversity of expected user population, and how this could introduce biases?

3.2.4. To address these questions and consider what perspectives may be missing from your project team, conduct a positionality reflection as individuals or, ideally, as a team (see Leslie et al., 2022: 43-44). In particular, please refer to the positionality matrix developed by The Alan Turing Institute and provided below as a starting point for considering how your positionality could influence your ability to identify and understand affected stakeholders and the potential impacts of my project. The process of stakeholder engagement detailed below should also be used to help fill any perspectival gaps.

Figure 2: **Positionality Matrix (developed by The Alan Turing Institute)**

4. Multistakeholder Governance

4.1. Why is this important?

To inclusively assess the impacts of this AI project, it is necessary to consult a diverse range of stakeholders. The project team should therefore produce a stakeholder engagement plan during the early stages of designing their system. This stakeholder engagement plan will allow the project team to set out their engagement objectives, which should be reviewed at regular intervals to ensure that stakeholder engagement is not only done as a checklist exercise, but rather constitutes an integral and transformative aspect of the decision-making process.

Please use the following stakeholder engagement template, adopted from work by The Alan Turing Institute, to facilitate multistakeholder engagement.

You are encouraged to consult the following resources (GPAI, 2022) for more extensive information and guidance on how to conduct detailed stakeholder engagement, the various modes of stakeholder engagement, and their respective strengths and weaknesses.

Depending on the specific context of the project, including the division of labour and the point of time when the EIA is conducted, **project teams are encouraged to adapt the template accordingly**. For example, if stakeholders have already been consulted, the system is already operational or if system is at an advanced stage of lifecycle when the EIA is being conducted, teams should rephrase questions to be in the past tense, assessing how well they have adhered to their original stakeholder engagement plan or detailing the steps they are taking to fulfil the plan.

4.2. Stakeholder Engagement Plan:

QUESTIONS:	RESPONSES:
4.2.1. What stakeholder groups are most likely to be impacted by the deployment of the AI system? Here, the project team maps out impacted stakeholders and identifies salient stakeholders (considering how protected characteristics and contextual vulnerabilities may intersect with one another to make particular stakeholders more vulnerable to adverse impacts). *In answering this question, consider:* 4.2.1.1. Who has the greatest needs for this tool? 4.2.1.2. Who has the least power to influence the development of this tool?	
4.2.2. Based on the previous question and on the positionality reflection, which stakeholder groups will you involve or consult during the development, deployment and use of the AI system?	

QUESTIONS:	RESPONSES:
4.2.3. What objective do you have for engaging these stakeholders? (e.g., to ensure all adverse impacts have been identified, to increase the community's level of trust in the system being designed) *In answering this question consider:* 4.2.3.1. Why are you engaging stakeholders? 4.2.3.2. How much will stakeholder input influence project development and outcomes? 4.2.3.3. How will you ensure that stakeholder input is taken into account?	
4.2.4. What is your plan for engaging stakeholders? *In answering this question consider:* 4.2.4.1. How should the plans be tailored to the needs of specific stakeholders? (e.g., if children are involved, the project team will need to carefully obtain consent and present materials in accessible language) 4.2.4.2. What resources are available and what time constraints may limit participation? 4.2.4.3. Which modes of stakeholder engagement would be more appropriate (e.g., online or in-person)? 4.2.4.4. How will these stakeholders be engaged, and will this differ during the different stages of the AI lifecycle? 4.2.4.5. Which activities will help your team to identify potential impacts and ensure they are mitigated? 4.2.4.6. When are you operationalising this plan/ engaging different stakeholders? Please elaborate on the timeline of its implementation. 4.2.4.7. If the EIA is being filled in after deployment: have you/ your project team already operationalised the plan? Please elaborate on the progress of its implementation.	

Implementing the UNESCO Principles

5. Safety and Security

The Recommendation states that *"unwanted harms (safety risks), as well as vulnerabilities to attack (security risks) should be avoided and should be addressed, prevented and eliminated throughout the lifecycle of AI systems to ensure human, environmental and ecosystem safety and security. Safe and secure AI will be enabled by the development of sustainable, privacy-protective data access frameworks that foster better training and validation of AI models utilizing quality data"*.

5.1. Why is this important?

While having access to more data is generally perceived as an opportunity to enhance security and accuracy, it means that attackers can also learn more rapidly and use AI to constantly improve their attacks, combining speed with context (Gregory, 2021).

Safety and security in AI implies that, in the same manner that before a new car is allowed to drive on the roads it has to undergo safety tests, and before a new medication is sold to consumers it must satisfy strict safety standards, AI also has to comply with regulatory, technical and societal standards (Conn, 2017). The black-box character of many AI systems further underlines the need to comply with safety and security standards as these technologies are complex and often combine multiple systems.

AI systems typically handle enormous amounts of (sometimes sensitive) data. This poses significant threats to data safety and security. The risk of data breaches presents a major concern, given that sensitive personal data is often the target of cyber-attacks (Schuster et al., 2021). The main types of attacks include:

- Data poisoning: manipulating the behaviour of the model by changing the training data or its labels (Papernot et al., 2017).
- Input manipulation: this technique entails inputting malicious content to trick the system. This is a very relevant concern given the rise of large language models, as it includes situations of direct attacks where someone inserting a prompt into ChatGPT or Bing Chat to try to make it behave in a different way, and indirect attacks that rely on data being entered from elsewhere, i.e. by instructing the bot to read documents or websites that contains attacking code (Burgess, 2023)
- Membership inference: through this method, the attacker is able to determine whether data related to a specific individual was used in the training, and this information can allow the attacker to link more closely between pieces of information and the identity of the person, potentially deducing sensitive personal information (The OWASP Foundation, n.d.).

To mitigate these issues, AI systems need to be tested and revalidated periodically, especially when they are used for decision-making in sensitive contexts such as healthcare.

5.2. Procedural Assessment

5.2.1. What measures were put in place to ensure the safety and security of the AI system and protect it from system manipulation?

5.2.2. What measures were put in place to ensure the safety and security of the AI system's training data from data poisoning/corruption?

5.2.3. What measures were put in place to ensure the safety and security of the data processed by the AI system?

5.2.4. If the training data or data being processed by the AI system were poisoned or corrupted, or if your system was manipulated, how would you know?

5.2.5. Has the AI system been tested prior to use? Please elaborate.

5.2.5.1. Please provide details regarding the testing procedure that led to the selection of the specific model(s) used in your AI system.

5.2.6. If the AI system is already into use, has further testing and revalidation been conducted after the AI system entered into use?

5.2.6.1. How did you test the robustness of the AI system, and what were the results?

5.2.6.2. How often will the AI system be tested in the future and which components will be tested?

5.3. Identifying and Mitigating Impacts:

5.3.1 Positive Impacts

The questions above are not exhaustive. We want you to consider all outcomes that could arise from the design, development, deployment and use of your AI system, both positive *and* negative. For guidance on filling up the tables, please refer to the Annex.

What are the prospective positive impacts of the system on safety and security?	**Please assess the scale of the prospective positive outcome.**	**Please assess the scope of the prospective positive outcome.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the likelihood of the prospective positive outcome occurring.**
E.g. the AI system is used to secure user authentication (via facial recognition, CAPTCHA, fingerprint scanner, etc.)	***Significance level*** • *Very high* • *High* • *Medium* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • *Intergenerational*	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*

5.3.2 Negative Impacts

What are the prospective negative/ adverse impacts of the system on safety and security?	**Please assess the scale of the prospective negative impact.**	**Please assess the scope of the prospective negative impact.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the remediability of the prospective negative impact.**	**Please assess the likelihood of the prospective negative impact occurring.**	**To what extent do the procedural safeguards described above mitigate this impact?** **What additional mitigation and redressal strategies will you need to implement to combat this potential harm?** *Please also detail* *1) The extreme cases that would warrant immediate stopping of the AI system.* *2) The cases that would warrant investigation and redressals. For these cases, please indicate and justify the time it would take for redressal.* *3) How you will guarantee non-repetition of the potential harm.*
E.g. there is no way to detect data poisoning	***Gravity level:*** • *Catastrophic* • *Critical* • *Serious* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • Intergenerational	***Degree of remediability*** • *Very low* • *Low* • *Medium* • High	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*	

6. Fairness, Non-Discrimination, Diversity

The Recommendation states that AI actors should promote social justice and safeguard fairness and non-discrimination of any kind in compliance with international law. This implies an inclusive approach to ensuring that the benefits of AI technologies are available and accessible to all, taking into consideration the specific needs of different age groups, cultural systems and language groups, persons with disabilities, girls and women, and disadvantaged, marginalized and vulnerable people or people in vulnerable situations.

6.1. Why is this important?

Analyses of facial-analysis systems have found that individuals with darker skin, particularly women, were more likely to be misclassified on aggregate (Buolamwini and Gebru, 2018). Speech recognition systems may also be less accessible to users from certain ethnic groups, with strong accents, or not speaking their first language (Koenecke et al., 2020). Researchers have also highlighted many other ways in which biases in the analogue world have been transferred over to and even amplified by AI (Guo and Caliskan, 2021), making systems inaccessible and leading to discriminatory outcomes. This often stems from and is exacerbated by the fact that datasets, data sources and technology teams in the AI sector tend to lack diversity (Howard and Isbell, 2020). As such, to safeguard fairness and prevent AI from perpetuating discrimination, procurement teams should ensure that there are processes in place to test against biases, such as conducting intersectional algorithmic bias audits on datasets (Howard, 2021), and be clear about how fairness is being addressed in algorithms, while also making active efforts to promote diversity and inclusiveness.

6.2. Procedural Assessment

Throughout this section, when responding to questions about testing with particular groups, the project team should consider especially – but not only – race, colour, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, and disability. Please specify if testing was conducted on groups that combine several of these criteria i.e. if the system has been tested in terms of intersectionality.

6.2.1. Preventing discriminatory outcomes:

6.2.1.1. Has the algorithm been tested with different groups?

6.2.1.1.1. Was there a difference in terms of accuracy rate (or any other performance metric used)? Please describe any difference of this kind.

6.2.1.1.2. Was there a discriminatory effect for particular groups?

6.2.2. Data quality and preventing discriminatory bias:

6.2.2.1. Are processes in place to test data against biases?

6.2.2.1.1. Have you undertaken an analysis of the data to prevent societal and historical biases in data?

6.2.2.1.2. Is the data well-balanced and does it reflect the diversity of the targeted end-user population?

6.2.2.1.3. Are there any differences you can foresee between the data used for training and the data processed by the AI system which could result in the AI system producing discriminatory outcomes or performing differentially for different groups?

6.2.2.1.4. Have you developed a process to document how data quality issues can be resolved during the design process?

6.2.2.1.5. Did you put in place educational and awareness initiatives to help AI designers and developers gain awareness of the possible bias they can introduce through the design and development of the AI system?

6.2.2.1. Are processes in place to test data against biases?

6.2.3. Preventing discrimination in terms of accessibility

6.2.3.1. Does the design allow all people, especially marginalized groups, to access and interact with the AI system? Please specify any restrictions in terms of accessibility.

6.2.3.1.1. Did you assess whether the AI system is usable by those with disabilities (e.g., accessible to screen readers, including alt text for images, colour-blind friendly palettes, etc.)?

6.2.3.1.2. Did you assess whether the AI system is usable by those with a precarious economic situation?

6.2.3.2. How has the principle of fairness been approached from a technical perspective? For example, are you able to specify what the technical notion of fairness is that the AI system is calibrated for? (e.g., individual fairness, demographic parity, equal opportunity, etc.)[1]

6.2.3.3. To which segment of the population will the AI system be applied? Is the population affected particularly marginalised?

1. For a non exhaustive list, see (Abu Elyounes, 2019)

6.3. Identifying and Mitigating Impacts:

6.3.1. Positive Impacts

The questions above are not exhaustive. We want you to consider all outcomes that could arise from the design, development, deployment and use of your AI system, both positive *and* negative. For guidance on filling up the tables, please refer to the Annex.

What are the prospective positive impacts of the system on fairness, non-discrimination and diversity?	**Please assess the scale of the prospective positive outcome.**	**Please assess the scope of the prospective positive outcome.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the likelihood of the prospective positive outcome occurring.**
E.g. The AI system helps to identify biased and discriminatory diction used during interviews by hiring managers	***Significance level*** • *Very high* • *High* • *Medium* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • *Intergenerational*	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*

6.3.2. Negative Impacts

What are the prospective negative/ adverse impacts of the system on fairness, non-discrimination and diversity?	**Please assess the scale of the prospective negative impact.**	**Please assess the scope of the prospective negative impact.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the remediability of the prospective negative impact.**	**Please assess the likelihood of the prospective negative impact occurring.**	**To what extent do the procedural safeguards described above mitigate this impact?** **What additional mitigation and redressal strategies will you need to implement to combat this potential harm?** *Please also detail* *1) The extreme cases that would warrant immediate stopping of the AI system.* *2) The cases that would warrant investigation and redressals. For these cases, please indicate and justify the time it would take for redressal.*
E.g. AI system helps to spread false information, or representational harms or abuse that may threaten users' psychological well-being (e.g., misgendering of persons by machine translation systems, abusive language in dialogue systems)	***Gravity level:*** • *Catastrophic* • *Critical* • *Serious* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • *Intergenerational*	***Degree of remediability*** • *Very low* • *Low* • *Medium* • *High*	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*	

7. Sustainability

The Recommendation states that the continuous assessment of the human, social, cultural, economic and environmental impacts of AI technologies should be carried out with full cognizance of the implications of AI technologies for sustainability. Sustainability here is understood as a set of constantly evolving goals across a range of dimensions, such as currently identified in the Sustainable Development Goals (SDGs) of the UN. In general, environmental and ecosystem flourishing should be protected and promoted through the lifecycle of AI systems.

7.1. Why is this important?

AI can be used to improve climate modelling, both for emissions predictions as well as to support mitigation such as through solar forecasting (Stein, 2020). While it is important that developers minimise the adverse impacts of algorithms on the environment, it is also imperative that AI use cases themselves are sustainable, given that it is possible for algorithms to run on low-carbon energy while at the same time supporting environmentally inimical use cases or encouraging emissions and resource-intensive activities (Kaack et al., 2022). For instance, AI-based recommendation systems may encourage the over-consumption of resources, while AI-powered autonomous vehicles may discourage individuals from choosing greener public transportation options.

The significant environmental impacts of many AI systems are well-documented, particularly those relating to the resources and infrastructure needed to keep models and algorithms running. Research has shown that the carbon footprint of training a large language model is equivalent to approximately 300,000 kg of carbon dioxide emissions (Strubell et al., 2019). The design, development and use of machine learning models also typically requires large amounts of electricity and water resources. It is important to be aware of such impacts and take actions to mitigate them.

The adverse environmental impacts of AI systems depend on several factors, and it is important that procurement teams are cognizant of these so they can make the right decisions. For example, the energy grid used by the system has significant effects on emissions, as different regions are powered by different combinations of renewable and non-renewable energy. As such, the carbon impacts of model training can be partly mitigated by selecting energy grids with minimal carbon emissions (Henderson et al., 2020; Lacoste et al., 2019; Patterson et al., 2021). The cloud provider is another factor; since providers differ in terms of commitment to sustainability, the project team should be careful in selecting the cloud provider for the AI system.

7.2. Procedural Assessment

7.2.1. Has an environmental impact assessment of your AI system ever been conducted?

7.2.1.1. Did you consult the environmental laws and policies that apply in your country/region during this process?

7.2.2. Are you using accountability metrics for responsible innovation (SDGs, ESGs) to project how your AI system can increase environmental flourishing (long-term sustainability) versus just avoiding immediate, likely regional and short-term harms?

7.2.3. Your AI system may harm the environment and ecosystems in different ways throughout its lifecycle. Some of these questions may apply more directly to embedded AI systems (application of AI algorithms and models at the device level). Please answer the following questions:

7.2.3.1. Research/design/development phase

7.2.3.1.1. Have you estimated the environmental impact of raw material extraction, processing and transportation involved in manufacturing the hardware of your AI system? If so, please describe how your methodology.

7.2.3.1.2. Have you measured your system's power consumption and if so, how?[2]

7.2.3.2. Use phase

7.2.3.2.1. Depending on what your AI system will be used for, it may encourage emissions- or resource-intensive activities. For instance, AI-based recommendation systems may boost consumption of resources. AI-powered autonomous vehicles may discourage individuals from choosing greener public transportation options. Is there a specific consideration of the environmental impacts of the use cases that your AI system is facilitating?

7.2.3.3. End-of-use/disassembly/termination phase

7.2.3.3.1. Once your system is decommissioned, how will you handle the process of dismantling, recycling and/or disposing of obsolete IT hardware?[3]

2. Project teams may wish to consult online tools such as Green Algorithms, ML CO2 and/or integrated tools such as Experimental Impact Tracker.

3. Project teams are encouraged to consult Circular Tech's Guide to the Circular Economy of Digital Devices (Navarro et al., 2021) to better understand the concept of circularity and reflect on how to maintain a focus on sustainability in the later to end stage(s) of the AI project lifecycle.

7.3. Identifying and Mitigating Impacts:

7.3.1. Positive Impacts

The questions above are not exhaustive. We want you to consider all outcomes that could arise from the design, development, deployment and use of your AI system, both positive *and* negative. For guidance on filling up the tables, please refer to the Annex.

What are the prospective positive impacts of the system on the environment and ecosystem flourishing?	**Please assess the scale of the prospective positive outcome.**	**Please assess the scope of the prospective positive outcome.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the likelihood of the prospective positive outcome occurring.**
E.g.: • *Protection, monitoring and management of natural resources* • *Prediction, prevention, control and mitigation of climate-related problems* • *Efficient and sustainable food ecosystem* • *Acceleration of access to and mass adoption of sustainable energy* • *Mainstreaming of sustainable infrastructure, business models and finance for sustainable development* • *Detection of pollutants or prediction of levels of pollutions* • *Other (please indicate, for example, identification of endangered animals or poachers)*	***Significance level*** • *Very high* • *High* • *Medium* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • *Intergenerational*	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*

7.3.2. Negative Impacts

What are the prospective negative/ adverse impacts of the system on the environment and ecosystem flourishing?	**Please assess the scale of the prospective negative impact.**	**Please assess the scope of the prospective negative impact.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the remediability of the prospective negative impact.**	**Please assess the likelihood of the prospective negative impact occurring.**	**To what extent do the procedural safeguards described above mitigate this impact?** **What additional mitigation and redressal strategies will you need to implement to combat this potential harm?** *Please also detail* *1) The extreme cases that would warrant immediate stopping of the AI system.* *2) The cases that would warrant investigation and redressals. For these cases, please indicate and justify the time it would take for redressal.*
E.g. destruction of natural habitats or increased production of fossil fuels	***Gravity level:*** • *Catastrophic* • *Critical* • *Serious* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • *Intergenerational*	***Degree of remediability*** • *Very low* • *Low* • *Medium* • *High*	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*	

8. Privacy and Data Protection

The Recommendation states that adequate data protection frameworks and governance mechanisms should be established in a multi-stakeholder approach at the national or international level, protected by judicial systems, and ensured throughout the lifecycle of AI systems.

8.1. Why is this important?

AI models are often trained by amassing vast amounts of data. When AI is combined with big data and the internet of things, such a practice raises concerns about users' privacy, and misuse of data. The impact of big data on AI is often characterized by three Vs, volume (the amount of data used for training), variety (the component that enables new and unanticipated inferences); and velocity (the component that facilitates analysis and sharing in real time) (Kerry, 2020). Our constant connectivity and interaction with connected devices mean that if privacy is not properly guarded, a full mosaic of our movements, personality, habits and taste can be created and taken advantage of (Idziniak, 2023). Training an efficient AI model for sensitive contexts (e.g., healthcare) requires significant amounts of privacy-sensitive data, and when AI is being used, there is always the risk of inferring sensitive data from so-called non-sensitive data or anonymized data. This happens because of the ability to link between several non-sensitive attributes and deduce sensitive information (van Bekkum and Zuiderveen Borgesius, 2023).

From a legal perspective, contrary to other principles, the domain of privacy and data protection is relatively more regulated around the world, although the strength and breadth of such protection vary. On the international level, Article 12 of the Universal Declaration of Human Rights anchor the protection of the right in a binding document (United Nations General Assembly, 1948). Similar protection can also be found in Article 8 of the European Convention on Human Rights (Council of Europe, 1950).

Data protection is an expression of the right to privacy, and it operationalizes it (Andrasko et al., 2021). In terms of its status, while privacy is an internationally recognized human right, data protection is not, despite the fact that it is also a heavily regulated subject: see, for example, the European General Data Protection Regulation that inspired several other countries around the world (Keane, 2021). Data protection laws often include requirements such as specification about the type of data being collected, the length of time the data is being stored, and consent specification (Zenonos, 2022).

Besides the legal requirements, from a technical perspective there are different methods under the umbrella of privacy by design, or privacy-preserving machine learning, that attempt to minimize as much as possible the risks to users' privacy, the possibility of identifying an individual, or leakage of personal information (Kourtellis, 2021). One such method is known as differential privacy, which adds noise to existing data in order to distance the individual from his or her identifiable information in case data is leaked (Wood et al., 2018).

8.2. Procedural Assessment

8.2.1. Data Protection

8.2.1.1. What types of personal data does the AI system have access to?

8.2.1.2. Are the data and input collected by humans, automated sensors or both?

8.2.1.3. Are the data and input from experts provided, observed, synthetic or derived?

8.2.1.4. If the data is coming from external entities, are there written agreements detailing the conditions for data sharing?

8.2.1.5. Is the data being stored at a level of security commensurate to its sensitivity?

- 8.2.1.5.1. If so, how and where?
- 8.2.1.5.2. For how long will data be retained?
- 8.2.1.5.3. Will the data be securely deleted when it is no longer required?

8.2.1.6. Is the data minimization principle being applied? In other words, is there an ex-ante assessment of the relevance and necessity of including each one of the data types in the system?

8.2.1.7. If the data is personal:

- 8.2.1.7.1. Are different types of personal data being subjected to different processing standards (especially sensitive types of data)?
- 8.2.1.7.2. Is the data anonymised or pseudonymised?
- 8.2.1.7.3. Does the system actively link between different databases?
- 8.2.1.7.4. Do people actively consent for the processing of their data by the AI system?

8.2.2. Privacy

8.2.2.1. Has a privacy impact assessment been conducted with regard to the AI system?

8.2.2.2. Has the quality of the training data been evaluated in terms of fairness and non-discrimination?

8.2.2.3. Is privacy by design being applied in the system? Please elaborate how.

8.2.2.4. Can users request the deletion of their data and stop the processing by the AI system?

8.2.2.5. If the data is accessible to third parties, are there provisions to protect against ill-intentioned actions, where relevant?

8.3. Identifying and Mitigating Impacts:

8.3.1. Positive Impacts

The questions above are not exhaustive. We want you to consider all outcomes that could arise from the design, development, deployment and use of your AI system, both positive *and* negative. For guidance on filling up the tables, please refer to the Annex.

What are the prospective positive outcomes of the system on the privacy of individuals and groups?	**Please assess the scale of the prospective positive outcome.**	**Please assess the scope of the prospective positive outcome.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the likelihood of the prospective positive outcome occurring.**
E.g. the AI system is used to identify malware and counter attacks against individuals	***Significance level*** • *Very high* • *High* • *Medium* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • *Intergenerational*	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*

8.3.2. Negative Impacts

What are the prospective negative/adverse impacts of the system on the privacy of individuals and groups?	**Please assess the scale of the prospective negative impact.**	**Please assess the scope of the prospective negative impact.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the remediability of the prospective negative impact.**	**Please assess the likelihood of the prospective negative impact occurring.**	**To what extent do the procedural safeguards described above mitigate this impact?** **What additional mitigation and redressal strategies will you need to implement to combat this potential harm?** *Please also detail* *1) The extreme cases that would warrant immediate stopping of the AI system.* *2) The cases that would warrant investigation and redressals. For these cases, please indicate and justify the time it would take for redressal.*
E.g. the AI system will be able to predict sensitive information about users from non-sensitive input (e.g., keyboard typing patterns may be used to predict emotional states)	***Gravity level:*** • *Catastrophic* • *Critical* • *Serious* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • *Intergenerational*	***Degree of remediability*** • *Very low* • *Low* • *Medium* • *High*	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*	

9. Human Oversight and Determination

The Recommendation states that Member States should ensure that it is always possible to attribute ethical and legal responsibility for any stage of the lifecycle of AI systems. It emphasizes that AI systems can never replace ultimate human responsibility and accountability. Human oversight refers not only to individual human oversight, but also to inclusive public oversight, as appropriate.

9.1. Why is this important?

Human oversight is crucial to supporting and respecting human autonomy (European Commission High-Level Expert Group on Artificial Intelligence, 2019). It helps to address process-based concerns, including helping to reduce the dehumanising effects of automation/ algorithmic decision-making and ensuring transparency and explainability, as well as outcome-based concerns, by assigning discretionary power and reducing discriminatory decisions (Koulu, 2020). In machine learning systems, human-in-the-loop and human-on-the-loop mechanisms are often utilized to provide a level of model oversight, for instance providing insights or making decisions for edge or outlier classification or prediction cases, and validating models.

This can be accomplished by ensuring that it is always possible to attribute responsibility for the system's decisions and outcomes to physical persons or legal entities, and that there are procedures in place for appointed staff to override the system if the need arises. However, human oversight needs to be included in a meaningful and effective way to avoid producing or amplifying negative impacts such as introducing bias into the system (Green and Chen, 2019; Skeem et al., 2019). Thus, it is also important that oversight mechanisms are accompanied by other procedural measures, including those stipulated in other sections of this EIA.

9.2. Procedural Assessment:

9.2.1. Does the model evolve and / or acquire abilities from interacting with data in the field?

9.2.2. Is the AI system (a) replacing an existing computer system; (b) replacing human beings; (c) adding new functionality or supplementing existing functionality?

9.2.3. If the AI system took over a task that was previously conducted by humans, how was the knowledge transfer preserved? How involved were the humans who were conducting the task previously in the development and training of the AI system?

9.2.4. Does the AI system have the authority to make a decision that would impact people?

9.2.4.1. If yes, is the decision subjected to meaningful human oversight before it takes effect?

9.2.5. Is it always possible to attribute ethical and legal responsibility for any stage of the lifecycle of the AI system to physical persons or to existing legal entities?

9.2.5.1. Who bears such responsibility in your project team?

9.2.6. Are there mechanisms in place for a human entity to override decisions made by the AI system?

9.2.6.1. If so, which individuals are given the authority to do so?

9.2.6.2. Please reflect on possible biases that may result from such authority (it may be helpful to refer to the positionality reflection from the Project Governance section of the EIA).

9.2.7. Is there a risk of over-reliance on AI systems such that human autonomy is adversely affected or compromised?

9.3. Identifying and Mitigating Impacts:

9.3.1. Positive Impacts

The questions above are not exhaustive. We want you to consider all outcomes that could arise from the design, development, deployment and use of your AI system, both positive and negative. For guidance on filling up the tables, please refer to the Annex.

What are the prospective positive impacts of the system on human oversight?	**Please assess the scale of the prospective positive outcome.**	**Please assess the scope of the prospective positive outcome.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the likelihood of the prospective positive outcome occurring.**
E.g. A strong emphasis on human oversight as part of the AI system, with clear disclaimers about the nature of human oversight and the parties involved, to the effect of an increase in the collective public understanding of AI systems and contribute cumulatively to making clear human oversight over AI systems a norm.	***Significance level*** • *Very high* • *High* • *Medium* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • *Intergenerational*	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*

9.3.2. Negative Impacts

What are the prospective negative/adverse impacts of the system on human oversight or the lack thereof?	**Please assess the scale of the prospective negative impact.**	**Please assess the scope of the prospective negative impact.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the remediability of the prospective negative impact.**	**Please assess the likelihood of the prospective negative impact occurring.**	**To what extent do the procedural safeguards described above mitigate this impact?** **What additional mitigation and redressal strategies will you need to implement to combat this potential harm?** *Please also detail* *1) The extreme cases that would warrant immediate stopping of the AI system.* *2) The cases that would warrant investigation and redressals. For these cases, please indicate and justify the time it would take for redressal.*
E.g., the AI system can take final healthcare decisions for patients, leading to reduced human autonomy	***Gravity level:*** • *Catastrophic* • *Critical* • *Serious* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • *Intergenerational*	***Degree of remediability*** • *Very low* • *Low* • *Medium* • *High*	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*	

10. Transparency and Explainability; Accountability and Responsibility

The Recommendation highlights that transparent and explainable AI systems which are accountable and have in place mechanisms to attribute responsibility are imperative to the respect, protection and promotion of human rights, fundamental freedoms and ethical principles. This includes the development of appropriate oversight, impact assessment, audit and due diligence mechanisms, including whistle-blowers' protection, to ensure accountability.

10.1. Why is this important?

It is important to ensure that all AI systems, including machine learning or robotic systems, regardless of the extent to which physical persons are in the loop, are subject to precise regulation and transparency and accountability requirements. Such mechanisms should come in the form of system- and context-dependent regulation able to facilitate explanations of the logic and reasons behind outcomes and decisions made by AI systems, thereby ensuring such information is easily accessible (Wachter et al., 2017). Explanations are crucial not only for preventing errors in outcomes and processes, but also for developing trust from end users (Doshi-Velez and Kim, 2017). Care should be taken to avoid opaque mechanisms and systems, including avoiding the use of black-box decision-making algorithms and design choices that imply the agency of the AI system (Long and Magerko, 2020).

As such, appropriate safeguards and mechanisms should be put in place in four key areas: system awareness, robust audits, algorithm explainability and transparency, and procedures for impact mitigation (e.g. appeals and complaints).

Procedural Assessment

10.2.1. System Awareness:

10.2.1.1. Are users made fully aware when they are interacting with an AI system, as opposed to a human being?

10.2.1.2. Are individuals (directly or indirectly) impacted by the AI system made fully aware of when a decision (that impacted them) was informed by or made on the basis of an AI system or AI algorithms?

10.2.1.2.1. Are they made aware of the extent to which they are impacted, including the rationale, benefits and limitations of the decision(s)?

10.2.1.3. Have appropriate explanations been put in place to help users and other impacted individuals understand the decision-making process or how the system works when required?

10.2.1.4. Have appropriate explanations been put in place to help the government bodies in charge of regulation understand the decision-making process or how the system works when required?

10.2.1.5. Has the decision to adopt the AI system been documented and communicated online?

10.2.1.6. Can the AI system make any decisions which the physical persons or legal entities in charge of the system lack expertise or competence to critique, modify or override?

10.2.2. Audits:

10.2.2.1. What technical and institutional designs have been put in place to ensure the accountability, auditability and traceability of (the working of) AI systems?

10.2.2.2. Is there a designated board, committee or person(s), or similar bodies designated to review issues of accountability and responsibility, and other ethical issues?

10.2.2.3. Is there any auditing process for the system?

10.2.2.3.1. Who oversees this audit process?

10.2.2.3.2. Does this involve internal, external, or third-party auditors?

10.2.2.3.3. Have you done the relevant checks to ensure there are no potential or existing conflicts of interest regarding the auditors?

10.2.2.3.4. Does the auditing process cover the entire project lifecycle? If not, which phases does the process cover?

10.2.2.3.5. When and how often is this audit conducted?

10.2.2.4. Is there an audit trail that keeps a record of all the decisions taken by the AI system?

10.2.2.5. Are all key decision-making checkpoints identifiable within the audit trail?

10.2.2.6. How is liability attributed?

10.2.3. Algorithmic Explainability:

10.2.3.1. Is the algorithm, including its inner-working logic, open to the public or any oversight authority? Is the code of the AI system in an open-source format?

10.2.3.2. Can public authorities request a copy of the code?

10.2.3.3. Are the datasets used for training the system known and traceable?

10.2.4. Mitigative Assessment:

10.2.4.1. Is there a protocol regarding liability allocation in case of malfeasance caused by the algorithm?

10.2.4.2. Is there a designated project team member or public sector institution who can review complaints, inform impacted individuals of the explanation(s), and correct the decision if needed?

10.2.4.2.1. Who is this person/ institution?

10.2.4.2.2. What is the timeframe for the review of complaints?

10.2.4.2.3. What is the timeframe for the correction of the decision?

10.2.4.3. Can individuals impacted by the AI system submit claims, complaints or requests for an explanation of how a decision was made to this project team member?

10.2.4.3.1. If yes, how are they made aware that it is possible to request an explanation?

10.2.4.4. Can individuals appeal a decision made by an AI system?

10.2.4.4.1. Are details on how to do so provided to them?

10.2.4.5. Are there mechanisms in place to monitor:

10.2.4.5.1. The designated project team members and public sector institutions who are in contact with members of the public?

10.2.4.5.2. The project team members who are overseeing and can override decisions made by the AI system?

10.2.4.6. Are there mechanisms in place to revoke access of individuals to the system (including the capacity to override decisions)?

10.2.4.6.1. How quickly can access be revoked?

10.2.4.7. Is there a procedure in place to investigate claims raised about the system by the general public, researchers or the media?

10.2.4.7.1. If yes, please elaborate on the procedure(s) and deadlines for the investigation

10.2.4.8. What provisions for whistle-blower protection have been made?

10.3. Identifying and Mitigating Impacts:

10.3.1. Positive Impacts

The questions above are not exhaustive. We want you to consider all outcomes that could arise from the design, development, deployment and use of your AI system, both positive *and* negative. For guidance on filling up the tables, please refer to the Annex.

What are the prospective positive impacts of the system on transparency and explainability, accountability and responsibility?	**Please assess the scale of the prospective positive outcome.**	**Please assess the scope of the prospective positive outcome.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the likelihood of the prospective positive outcome occurring.**
E.g. AI is used as part of a chatbox to allow users to provide feedback, lodge complaints and raise claims about (another) AI system.	***Significance level*** • *Very high* • *High* • *Medium* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • *Intergenerational*	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*

10.3.2. Negative Impacts

What are the prospective negative/adverse impacts relating to transparency, explainability, accountability and responsibility, or a lack thereof?	**Please assess the scale of the prospective negative impact.**	**Please assess the scope of the prospective negative impact.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the remediability of the prospective negative impact.**	**Please assess the likelihood of the prospective negative impact occurring.**	**To what extent do the procedural safeguards described above mitigate this impact?** **What additional mitigation and redressal strategies will you need to implement to combat this potential harm?** *Please also detail* *1) The extreme cases that would warrant immediate stopping of the AI system.* *2) The cases that would warrant investigation and redressals. For these cases, please indicate and justify the time it would take for redressal.*
E.g. A lack of manpower to handle appeals and complaints can lead to [impacted parties] failing to receive the attention, justification and redressal measures that they require.	***Gravity level:*** • *Catastrophic* • *Critical* • *Serious* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • Intergenerational	***Degree of remediability*** • *Very low* • *Low* • *Medium* • *High*	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*	

11. Awareness and Literacy

The Recommendation notes that for the ethical development and deployment of AI, public awareness and understanding of AI technologies and the value of data should be cultivated and grounded by their impact on human rights, fundamental freedoms and the environment and ecosystems. This can be promoted through open and accessible education, civic engagement, digital skills and AI ethics training, media and information literacy and training led jointly by governments, intergovernmental organizations, civil society, academia, the media, community leaders and the private sector.

11.1. Why is this important?

AI products and tools are increasingly being integrated and utilised in user-facing technology across the world (Long and Magerko, 2020). It is therefore crucial that everyone, including end users, civil society, policymakers and students, not just developers and procurers, becomes educated about AI and, eventually, AI literate (Firth-Butterfield et al., 2022). Widespread AI literacy will ensure that users are able to serve as critical consumers: using AI effectively as a tool, evaluating AI systems and holding developers/ procurers accountable (Ibid.). It can also foster innovation, support and promote transparency and accountability, quell unfounded fears among users and prevent misunderstandings.

Fostering AI awareness and literacy is closely linked to ensuring transparency, explainability and accountability.

11.2. Procedural Assessment:

11.2.1. Has there been a public announcement regarding the intention to design this AI system?

11.2.2. Can information be found online regarding the system, its capabilities, its purpose and functionality? If not, is there a plan to publish this information at a particular stage of the project lifecycle?

11.2.2.1. Is the language used to present the system appropriate for the general public?

11.2.3. Will the system be used by the public or only internally?

11.2.3.1. If the system will only be used internally, what is the level of competency of those who will interact with it?

11.2.3.2. If the system will be used by the public, can people report their experience interacting with the system and concerns related to its impacts? Is the process for doing so simple, accessible and clearly advertised?

11.2.4. Have any schemes been put in place to help educate users and impacted groups about this system and the reason behind its deployment? For example, an educational media campaign or workshops involving community leaders.

11.3. Identifying and mitigating impacts:

11.3.1 Positive Impacts

The questions above are not exhaustive. We want you to consider all outcomes that could arise from the design, development, deployment and use of your AI system, both positive *and* negative. For guidance on filling up the tables, please refer to the Annex.

What are the prospective positive outcomes of the system on AI awareness and literacy? **How, if at all, could the deployment of this system increase awareness surrounding AI? Are there any other ways in which this system could increase awareness and literacy?**	**Please assess the scale of the prospective positive outcome.**	**Please assess the scope of the prospective positive outcome.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the likelihood of the prospective positive outcome occurring.**
E.g. The AI system is used to support learning as part of an AI literacy educational module *E.g. the deployment of the AI system is accompanied by the publication of an online document/video explaining the aim and characteristics of the AI system in layman's terms*	***Significance level*** • Very high • High • Medium • Moderate/ minor	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • *Intergenerational*	***Likelihood (of occurrence)*** • Low • Medium • High • Very high

11.3.2. Negative Impacts

What are the prospective negative/ adverse impacts of the system on AI awareness and literacy? **How, if at all, could this system decrease awareness surrounding AI? Are there any other ways in which this system could have an adverse impact on awareness and literacy?**	**Please assess the scale of the prospective negative impact.**	**Please assess the scope of the prospective negative impact.** *This should include:* • *Descriptions of the impacted persons/ groups/ entities. This may include any living organisms (including humans).* • *The timescale of the impact*	**Please assess the remediability of the prospective negative impact.**	**Please assess the likelihood of the prospective negative impact occurring.**	**To what extent do the procedural safeguards described above mitigate this impact?** **What additional mitigation and redressal strategies will you need to implement to combat this potential harm?** *Please also detail* *1) The extreme cases that would warrant immediate stopping of the AI system.* *2) The cases that would warrant investigation and redressals. For these cases, please indicate and justify the time it would take for redressal.*
E.g. The AI system runs the risk of perpetuating the common misconception that AI systems have agency because it uses language that implies intentionality and sentience, such as: "I think that this choice is something that you would enjoy".	***Gravity level:*** • *Catastrophic* • *Critical* • *Serious* • *Moderate/ minor*	***Extent of Impact*** *Impacted Parties* • *Primary* • *Secondary* • *Unexpected/ unintended* *Timescale* • *Short-term* • *Medium-term* • *Long-term* • *Intergenerational*	***Degree of remediability*** • *Very low* • *Low* • *Medium* • *High*	***Likelihood (of occurrence)*** • *Low* • *Medium* • *High* • *Very high*	

Annex

Guidance on Filling in the Sub-Sections 'Identifying and Mitigating Impacts'

The sub-sections on impact assessment aim to identify and classify both potential (or existing, for ex-post assessments) positive and adverse consequences, thereby guiding the procurement (and implicitly, the development and innovation) process as well as, where applicable, the identification and implementation of corresponding redress and mitigation measures.

Under each section of the second chapter on Implementing the UNESCO Principles, project teams should identify any potential positive outcomes and adverse impacts that may arise from the procurement, deployment and use of the AI system in question, specific to a given UNESCO principle and the AI system's context of use. Importantly, this exercise requires engaging with a range of potentially affected individuals, representatives and communities through multistakeholder consultations proportionate to the scale and scope of the project, its urgency and expected impacts. While the structure of this impact assessment is such that each impact and its corresponding assessment is recorded under disparate principles, we recognise that impacts are often complex and applicable to an interplay of principles and values. Project teams are therefore encouraged to remain cognizant of and actively acknowledge this when filling in the impact tables.

Additionally, it must be acknowledged that while this impact assessment aims to be comprehensive, it is not exhaustive. This is especially relevant for ex-ante impact assessments at earlier points in the project lifecycle, which procurers are encouraged to engage in, reflecting upon potential future impacts. As with other parts of the EIA, the impact assessment segments are meant to be engaged with iteratively, and should be adjusted, contextualized and expanded in accordance with the stage of the AI lifecycle.

The scoring framework for the impacts is adapted from Human Rights Impact Assessment (HRIA) literature, methodologies and criteria, particularly the UNHR's Guiding Principles on Business and Human Rights (United Nations Human Rights Office of the High Commissioner, 2011), which has long served as an internationally recognized guide to assessing and addressing the human rights impacts of businesses. Given that existing HRIA models are often highly granular and oversized, these frameworks have been tailored and contextualized for AI applications for the purpose of this impact assessment exercise (Mantelero, 2022). Additionally, while the Recommendation and this tool are rooted in respect for human rights and dignity, unlike traditional human rights frameworks, this impact assessment is intended to be conducted with reference not just to affected physical persons or groups/communities, but also the physical environment and other living organisms, in accordance with the Recommendation's chapter on the environment and ecosystems. Less tangible and wider, more cumulative or aggregate impacts, such as effects on the general literacy or perception of AI systems or effects on diversity and general perceptions of discriminated groups, should also be accounted for, but should as far as possible be explicitly linked to impacted groups.

For each individual impact, there are two categories to consider: **severity** (or **significance level** for positive outcomes) and **likelihood**. Under severity, there exist three sub-categories: **scale**, **scope** and **remediability**. When scoring each category or sub-category, project teams are encouraged to elaborate extensively on their score, with accompanying justifications, details and nuances.

Severity/ Significance Level

Based on Article 14 of the UNGP, the assessment of the severity of a given **negative** impact is based on the consideration of 3 factors/ sub-categories: *scale, scope* and *remediability*.

For **positive** impacts, the assessment of significance is based on 2 factors/ sub-categories: *scale* and *scope*.

We recognize that severity is a relative and not an absolute concept, and it differs based on the specific situation and perspectives of stakeholders, as well as the arbitrary judgement of project teams. This set of instructions thus aims to support you in filling out the respective tables so that this may constitute a productive and illuminating exercise.

Ideally, the respective assessments of scale, scope and remediability should recognise the nuances in how different impacted parties are affected in different ways and to varying degrees (e.g. vulnerable, marginalized or minority communities might suffer an impact to a greater extent than other populations). Thus, different risks may be faced by different groups – this should be specified in the impact assessment segment as far as possible. This consideration is relevant for situations where the impacts or potential impacts of the AI system are particularly broad, involving, for instance, entire populations.

Scale

Scale here refers to the gravity or seriousness of a given impact.

For **positive impacts**, the scale of a given impact can be assessed along a continuum of four Significance level: Low, Medium, High and Very High (transformative). Project teams should use the following table to guide assessments of scale:

SIGNIFICANCE LEVEL	DESCRIPTION
Very high	Extremely transformative and significant positive outcomes, possibly systemic change, will benefit all or some of the affected parties, in ways that lastingly or permanently heighten the quality of life; affirm and secure respect for human rights; improve and/or enhance the welfare of entire groups or communities; improve democratic society; support the legal order; transform all or parts of education, critical thinking, awareness and/or literacy, digital or otherwise; encourage community; contribute to sustainability and/or environmental flourishing; support or enhance infrastructure.
High	Very significant positive outcomes will benefit some or all of the affected parties, in ways that lead to notable and lasting enhancement of the quality of life; support of democratic society; support of the legal order; support of parts of education, critical thinking, awareness and/or literacy, digital or otherwise; encouragement of community; contribution to sustainability and/or environmental flourishing; support of infrastructure.
Medium	Significant positive outcomes will benefit some or all of the affected parties, in ways that lead to temporary enhancement of the quality of life; support of democratic society; support of the legal order; support of parts of education, critical thinking, awareness and/or literacy, digital or otherwise; encouragement of community; contribution to sustainability and/or environmental flourishing; support of infrastructure.
Moderate/ minor	Moderate or minor positive outcomes will benefit all or some of the affected parties, infrastructure and/or the biosphere and natural environment.

For **negative impacts**, the scale of a given impact can be assessed along a continuum of four Gravity level: Moderate/ Minor, Serious, Critical and Catastrophic. Project teams should use the following table to guide assessments of scale:

GRAVITY LEVEL	DESCRIPTION
Catastrophic	Catastrophic harms and negative impacts are accrued to all or some of the affected parties, in ways that lead to the deprivation of the right to life; irreversible injury to physical, psychological, or moral integrity; deprivation of the welfare of entire groups or communities; catastrophic harm to democratic society, the rule of law, or to the preconditions of democratic ways of life and just legal order; deprivation of individual freedom and of the right to liberty and security; harm to the biosphere and/or infrastructure.
Critical	Critical harms and negative impacts are accrued to all or some of the affected parties, in ways that lead to the significant and enduring degradation of human dignity, autonomy, physical, psychological, or moral integrity, or the integrity of communal life, democratic society, or just legal order
Serious	Serious harms and negative impacts are accrued to all or some of the affected parties, in ways that lead to the temporary degradation of human dignity, autonomy, physical, psychological, or moral integrity, or the integrity of communal life, democratic society, or just legal order or that harm to the information and communication environment
Moderate/ minor	Moderate or minor harms and negative impacts are accrued to all or some of the affected parties, in ways that do not lead to any significant, enduring, or temporary degradation of human dignity, autonomy, physical, psychological, or moral integrity, or the integrity of communal life, democratic society, or just legal order

(Table adapted from The Alan Turing Institute)

Scope

Scope here refers to the extent of impact: how many people are or could be affected, and/or how widespread the adverse impact is/could be, as well as the timescale of the impact. In the impact assessment tables, project teams are asked to describe the extent of impact through listing impacted parties and noting the timescale of the impact.

Impacted Parties

Project teams are to first list out all impacted/potentially impacted persons, organisms, groups and/or environments, and the extent to which each of these impacted parties are affected by the positive or negative impact.

These can be categorized under primary, secondary and unexpected/unintended impacted parties. To ensure a comprehensive assessment, this exercise should be undertaken in consultation with a variety of stakeholders, including experts, developers, impacted/potentially impacted parties, etc.

	DESCRIPTION	EXAMPLES
Primary	Persons directly involved in the development, deployment and use of the AI system. This includes developers, procurers, as well as immediate end users or objectives of the AI system. Persons, organisms, environments or things that the AI systems provide direct input into and are involved in its operation.	*End users of a particular algorithm: the students for an AI system used for learning a module, patients for an AI system used to improve efficiency of healthcare diagnoses, etc.*
Secondary	Persons, organisms, environments or things that are or may be impacted indirectly or proximately as a function of the AI system's impact on primary impacted parties.	*The family of primary impacted parties, the environment of primary impacted parties, those dependent on impacted parties (e.g. consumers of crops from farmers that utilise an AI system for more efficient harvesting), etc.*
Unexpected/ unintended	Persons, organisms, environments or things that may be impacted or could be impacted unexpectedly. Such impacted parties are often affected due to impacts that are not foreseen. While these are hard to gauge and identify during ex-ante assessment, you are encouraged to use this category as a brainstorming exercise for extreme and/or less likely scenarios to bolster preparedness.	*The climate/ environment, quality of life, communities, the general population (who may for instance be impacted by cumulative impacts – see 'Timescale' below) etc.*

It is also important to note whether any of the impacted parties possess characteristics that could make them more susceptible or vulnerable to higher, more prolonged or more intense levels of impact (Leslie et al., 2022).

Timescale

You should next consider and elaborate upon the timescale of the given impact. For negative impacts, identifying the timescale of the impact or potential impact can help project teams gauge the extent of action required for mitigation.

Impacts can be categorised as **short-term, medium-term, long-term** or **intergenerational impacts.**

Additionally, project teams are also encouraged to identify and elaborate on whether there are **cumulative** or **aggregate** impacts of the system (Götzmann et al., 2020). This refers to impacts that are "incremental, combined, and successive across space and time" (Franks et al., 2011). Examples of this may include (but are not limited to) impacts on the wider digital literacy of specific communities or populations, or contributing to underlying prejudices and biases.

Remediability (applicable only to negative impacts)

Remediability here refers to the capacity for reparability and/or restoration: whether and the ease with which impacted persons and/or objects can be restored to a situation equivalent to their situation immediately prior to the impact.

As highlighted under Article 24 of the UNGP, it is important that project teams prioritize preventing and mitigating those impacts or potential impacts that are most severe or where a delayed response would preclude remediation. It is thus crucial that project teams are able to identify and establish the remediability of impacts or potential impacts.

Project teams may refer to the following table for guidance on how to classify the remediability of a potential negative impact.

DEGREE OF REMEDIABILITY	EFFORT
Very Low	Suffered harm may be irreversible and may not be overcome (e.g., long-term psychological or physical ailments, death, etc.)
Low	Suffered harm can be overcome albeit with serious difficulties and enduring effects (e.g., economic loss, property damage, worsening health, loss of social trust, deterioration of confidence in the legal order, etc.)
Medium	Suffered harm can be overcome despite some difficulties (e.g., extra costs, fear, lack of understanding, stress, minor physical ailments, etc.)
High	Suffered harm can be overcome without any problem (e.g., time spent amending information, annoyances, irritations, etc.)

(Table adapted from The Alan Turing Institute)

Likelihood

The likelihood of a given impact refers to the probability that a given positive/negative impact is going to occur.

In assessing the likelihood of an impact, project teams should consider several factors, among others:

1. End-user interests, motivations and incentives
2. Developer interests, motivations and incentives
3. National policies and laws
4. End-user AI awareness and literacy

Likelihood can be scored across four categories: *low*, *medium*, *high* and *very high*.

LIKELIHOOD OF IMPACT	DESCRIPTION
Very High	The likelihood of the impact occurring is very high. It is highly probably that it will occur.
High	The likelihood of the impact occurring is high. It is probable that it will occur.
Medium	The likelihood of the impact occurring is moderate. It is possible that it may occur.
Low	The likelihood of the impact occurring is low. It is improbable that it may occur.

Bibliography

Abu Elyounes D (2019) Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness. *Journal of Law, Technology and Policy*. Elsevier BV. DOI: 10.2139/SSRN.3478296.

Andraško J, Mesarčík M and Hamuľák O (2021) The regulatory intersections between artificial intelligence, data protection and cyber security: challenges and opportunities for the EU legal framework. *AI and Society* 36(2). Springer Science and Business Media Deutschland GmbH: 623–636. DOI: 10.1007/S00146-020-01125-5/METRICS.

Bommasani R, Hudson DA, Adeli E, et al. (2021) On the Opportunities and Risks of Foundation Models. *Arrive*. Available at: https://arxiv.org/abs/2108.07258v3 (accessed 13 June 2023).

Buolamwini J and Gebru T (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR. Available at: https://proceedings.mlr.press/v81/buolamwini18a.html (accessed 17 April 2023).

Commission Nationale de l'Informatique et des Libertés (2022) Asking the right questions before using an artificial intelligence system. Available at: https://www.cnil.fr/en/asking-right-questions-using-artificial-intelligence-system (accessed 22 March 2023).

Council of Europe (1950) European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14. Available at: https://www.refworld.org/docid/3ae6b3b04.html (accessed 13 June 2023).

Doshi-Velez F and Kim B (2017) Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608*. Available at: https://arxiv.org/abs/1702.08608v2 (accessed 17 April 2023).

European Commission High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed 13 June 2023).

European Data Protection Supervisor (n.d.) Necessity & Proportionality. Available at: https://edps.europa.eu/data-protection/our-work/subjects/necessity-proportionality_en (accessed 22 March 2023).

Firth-Butterfield K, Topic L, Anthony A, et al. (2022) Without universal AI literacy, AI will fail us. Available at: https://www.weforum.org/agenda/2022/03/without-universal-ai-literacy-ai-will-fail-us/ (accessed 17 April 2023).

Franks D, Brereton D and Moran C (2011) Cumulative Social Impacts. In: Vanclay F and Esteves AM (eds) *New Directions in Social Impact Assessment: Conceptual and Methodological Advances*. Cheltenham: Edward Elgar.

Götzmann N, Bansal T, Wrzoncki E, et al. (2020) Human rights impact assessment guidance and toolbox. Available at: https://www.humanrights.dk/tools/human-rights-impact-assessment-guidance-toolbox (accessed 13 June 2023).

GPAI (2022) *Data Justice in Practice: A Guide for Policymakers*.

Green B and Chen Y (2019) Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc: 90–99. DOI: 10.1145/3287560.3287563.

Guo W and Caliskan A (2021) Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. AIES 2021 - *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, Inc: 122–133. DOI: 10.1145/3461702.3462536.

Henderson P, Hu J, Romoff J, et al. (2020) Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *Journal of Machine Learning Research* 21. Microtome Publishing: 1–44. Available at: https://arxiv.org/abs/2002.05651v2 (accessed 17 April 2023).

Howard A (2021) Real Talk: Intersectionality and AI. Available at:https://sloanreview.mit.edu/article/real-talk-intersectionality-and-ai/ (accessed 17 April 2023).

Howard A and Isbell C (2020) Diversity in AI: The Invisible Men and Women. Available at: https://sloanreview.mit.edu/article/diversity-in-ai-the-invisible-men-and-women/ (accessed 17 April 2023).

Idziniak S (2023) AI and data privacy: protecting information in a new era. Available at: https://technologymagazine.com/articles/ai-and-data-privacy-protecting-information-in-a-new-era (accessed 13 June 2023).

Kaack LH, Donti PL, Strubell E, et al. (2022) Aligning artificial intelligence with climate change mitigation. *Nature Climate Change 2022 12:6* 12(6). Nature Publishing Group: 518–527. DOI: 10.1038/s41558-022-01377-7.

Keane J (2021) From California to Brazil: GDPR has created recipe for the world. Available at: https://www.cnbc.com/2021/04/08/from-california-to-brazil-gdpr-has-created-recipe-for-the-world.html (accessed 14 June 2023).

Kerry C (2020) Protecting privacy in an AI-driven world. Available at: https://www.brookings.edu/research/protecting-privacy-in-an-ai-driven-world/ (accessed 13 June 2023).

Koenecke A, Nam A, Lake E, et al. (2020) Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America* 117(14). National Academy of Sciences: 7684–7689. DOI: 10.1073/PNAS.1915768117/SUPPL_FILE/PNAS.1915768117.SAPP.PDF.

Koulu R (2020) Proceduralizing control and discretion: Human oversight in artificial intelligence policy. *Maastricht Journal of European and Comparative Law* 27(6). SAGE Publications Ltd: 720–735. DOI: 10.1177/1023263X20978649/SUPPL_FILE/16.11.2020_KOULU_RIIKKA_FINAL_SUB.DOCX.

Kourtellis N (2021) Building Machine Learning Models with Privacy by Design in Mind. Available at: https://www.concordia-h2020.eu/blog-post/building-machine-learning-models-with-privacy-by-design-in-mind/ (accessed 13 June 2023).

Lacoste A, Luccioni A, Schmidt V, et al. (2019) Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700*. Available at: https://arxiv.org/abs/1910.09700v2 (accessed 17 April 2023).

Leslie D, Burr C, Aitken M, et al. (2022) Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal. Zenodo. DOI: 10.5281/ZENODO.5981676.

Long D and Magerko B (2020) What is AI Literacy? Competencies and Design Considerations. CHI '20: Proceedings of the 2020 *CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery: 1–16. DOI: 10.1145/3313831.3376727.

Mantelero A (2022) Human Rights Impact Assessment and AI. In: *Beyond Data: Human Rights, Ethical, and Social Impact Assessment in AI.* T.M.C. Asser Press.

Navarro L, Roura M, Franquesa D, et al. (2021) A guide to the circular economy of digital devices. Available at: https://circulartech.apc.org/books/a-guide-to-the-circular-economy-of-digital-devices (accessed 13 June 2023).

OECD (2022) The OECD Framework for the Classification of AI systems. Available at: www.oecd.ai/classification (accessed 13 June 2023).

Patterson D, Gonzalez J, Le Q, et al. (2021) Carbon Emissions and Large Neural Network Training. Available at: https://arxiv.org/abs/2104.10350v3 (accessed 17 April 2023).

Schuster R, Song C, Tromer E, et al. (2021) You Autocomplete Me: Poisoning Vulnerabilities in Neural Code Completion You Autocomplete Me: Poisoning Vulnerabilities in Neural Code Completion. In: *Proceedings of the 30th USENIX Security Symposium*, 2021. Available at: https://www.usenix.org/conference/usenixsecurity21/presentation/schuster (accessed 7 May 2023).

Skeem JL, Scurich N and Monahan J (2019) Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants. *Virginia Public Law and Legal Theory Research Paper*. Available at: https://papers.ssrn.com/abstract=3316266 (accessed 13 June 2023).

Stein AL (2020) Artificial Intelligence and Climate Change. *Yale Journal on Regulation* 37. Springer-Verlag. DOI: 10.1007/SPRINGERREFERENCE_22452.

Strubell E, Ganesh A and McCallum A (2019) Energy and Policy Considerations for Deep Learning in NLP. *ArXiv Preprint* 53(1): 3645–3650. Available at: https://arxiv.org/abs/1906.02243v1 (accessed 17 April 2023).

The Information Accountability Foundation (2019) Ethical Data Impact Assessments and Oversight Models. Available at: https://www.immd.gov.hk/pdf/PCAReport.pdf (accessed 13 June 2023).

UNESCO (2021) Recommendation on the Ethics of Artificial Intelligence. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000380455 (accessed 17 April 2023).

United Nations General Assembly (1948) Universal Declaration of Human Rights. United Nations. Available at: https://www.un.org/en/about-us/universal-declaration-of-human-rights (accessed 13 June 2023).

United Nations Human Rights Office of the High Commissioner (2011) *Guiding principles on business and human rights – Implementing the United Nations "Protect, Respect and Remedy" Framework.*

van Bekkum M and Zuiderveen Borgesius F (2023) Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? *Computer Law & Security Review* 48. Elsevier Advanced Technology: 105770. DOI: 10.1016/J.CLSR.2022.105770.

Wachter S, Mittelstadt B and Russell C (2017) Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 32(2). Elsevier BV. DOI: 10.2139/SSRN.3063289.

Whittlestone J, Alexandrova A, Nyrup R, et al. (2019) The role and limits of principles in AI ethics: Towards a focus on tensions. *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, Inc: 195–200. DOI: 10.1145/3306618.3314289.

Wood A, Altman M, Bembenek A, et al. (2018) Differential Privacy: A Primer for a Non-Technical Audience. *Vanderbilt Journal of Entertainment & Technology Law*. Available at: http://privacytools.seas.harvard.edu (accessed 13 June 2023).

Zenonos A (2022) Artificial Intelligence and Data Protection. Available at: https://towardsdatascience.com/artificial-intelligence-and-data-protection-62b333180a27 (accessed 13 June 2023).