



4 PEOPLE
INSTITUTE



THE AI4PEOPLE PLAYBOOK:
IMPLEMENTING A PROPORTIONAL
APPROACH TO ETHICAL
AI REQUIREMENTS

AI4PEOPLE

AI4People Institute brings together leading AI companies, academics, civil society groups and governments to consider the risks of AI and advocate for responsible and beneficial AI development and deployment. Close contact with leading policy makers is one of the key factors in the success of its action which is at the origin of the regulatory process in Europe.

AI4People focuses on various aspects related to AI, including ethics, policy, governance, privacy, inclusivity, and the responsible use of AI across different domains. Following the publication of the report “Towards an Ethics by Design Approach for AI”, the AI4People Institute has launched an initiative to develop the AI4People’s AI Ethics Playbook, intended to facilitate the implementation of the Ethics by Design Approach for AI.



The aim of the Playbook is to promote a balanced and pragmatic approach to core ethical principles, one that goes “beyond regulatory compliance”. By offering actionable recommendations, highlighting existing resources, and showcasing real-world use cases and activities, the guidance emphasizes that adopting an ethical perspective is the most effective way to prevent costly mistakes. Such an approach serves as a comprehensive safeguard against risks that may otherwise go unnoticed until it is too late.

I would like to thank all the members of the Working Group for making this playbook possible, and in particular for the brilliant and crucial coordination role of Virginia Ghiara, Senior Research Manager at Fujitsu Research of Europe. Thanks also to the members of the Scientific Committee for their contributions.

I dedicate this latest effort of the AI4People Institute to former French President Valéry Giscard d’Estaing, who passed away exactly five years ago today and who strongly supported the creation of AI4People in 2018.

Bruxelles, 2 December 2025

Michelangelo Baracchi Bonvicini
President, AI4People Institute



AUTHORS

- **Chazerand, Patrice** | Director, Public Affairs, AI4People Institute
- **Ghiara, Virginia** | Senior Research Manager, Fujitsu Research of Europe (Rapporteur)
- **Horzyk, Amanda** | Doctoral Researcher in Designing Responsible Natural Language Processing, University of Edinburgh, Futures Institute
- **Kennedy, Rónán** | Associate Professor at University of Galway and Lero
- **Leclère, Camilah** | AI4People Project Officer, master's student at Université Panthéon Sorbonne
- **Liu, Xinpeng** | Doctoral researcher in generative AI and copyright at University of Galway School of Law
- **Massucci, Flavia** | AI Compliance Specialist, Intesa Sanpaolo
- **Schafer, Burkhard** | Professor of Computational Legal Theory, University of Edinburgh
- **Slosser, Jacob Livingston** | Director of Research, Sapien Institute; iCourts Global Fellow, University of Copenhagen Faculty of Law
- **Tschap, Nicole** | European Affairs Policy Manager, Fujitsu



SCIENTIFIC COMMITTEE

- **Burkhard Schafer (Chairman)** | Professor of Computational Legal Theory, University of Edinburgh
- **Sergei Bobrovskyi** | Data Scientist, AI Platforms team, Airbus
- **Patrice Chazerand** | Director of Public Affairs, AI4People Institute
- **Donald Combs** | Vice President and Dean, School of Health Professions, Eastern Virginia Medical School
- **Mario D'Almo** | Head of the Digital Strategy and AI Service, Banca Intesa San Paolo
- **Fausto Pedro Garcia Marquez** | Full Professor, Castilla-La Mancha University, Spain (UCLM)
- **Virginia Ghiara** | Principal Researcher AI Ethics, Fujitsu Research of Europe
- **Hiroki Habuka** | Research Professor, Graduate School of Law, Kyoto University
- **Hiroya Inakoshi** | Project Director, Fujitsu



SCIENTIFIC COMMITTEE

- **Rónán Kennedy** | Senior Lecturer, University of Galway
- **Jacob Livingston Slosser** | Assistant Professor of Law and Cognition, University of Copenhagen
- **Robert Madelin** | Former Director General, DG Connect, European Commission
- **Cory Robinson** | Associate Professor in Communication Design & Information Systems, Linköping University
- **Jay Stoltzenberg** | Head of AI Services, Airbus



TABLE OF CONTENTS

1 INTRODUCTION	8
2 ETHICS BEYOND COMPLIANCE	12
3 A PROPORTIONAL APPROACH TO ETHICAL REQUIREMENTS	14
3.1 AI Ethics Impact Assessment & Risk management systems	14
3.2 Human oversight	18
3.3 Fairness	24
3.4 Accuracy, robustness and cybersecurity	29
3.5 Quality management system	35
3.6 The ethical development and use of Generative AI	38
3.7 Explainability	45
3.8 Environmental Impact	51
4 THE EU'S APPROACH TO AI GOVERNANCE	56
5 RECOMMENDATIONS TO POLICYMAKERS	61
ANNEX 1: AI PLAYBOOKS	65
ANNEX 2: HUMAN OVERSIGHT	67
ANNEX 3: GUIDANCE ON THE LEGAL IMPLICATIONS OF GENERATIVE AI	72
ANNEX 4: WHISTLEBLOWING	76
ANNEX 5: ENVIRONMENTAL IMPACT INDICATORS	78
ANNEX 6: RESOURCE LIST	80
ANNEX 7: POLICY LANDSCAPE	82
GAME 1: OVERSIGHT IN ACTION	84
GAME 2: WHAT TYPE OF BIAS?	86
GAME 3: THE ETHICAL RISKS OF ACCURACY, ROBUSTNESS & CYBERSECURITY	88
GAME 4: EXPLAINABILITY CROSSWORD	89
GAME 5: POLICY TRACKER CHALLENGES	90



INTRODUCTION

The AI4People Playbook

The AI4People Playbook breaks down AI ethics into actionable steps, and it includes instructions and practical material to support capacity-building activities, empowering you to put principles into practice effectively. While many resources support responsible AI, from the AI Office's GPAI Code of Practice to various global standards and simplified guides, this playbook stands out for its clear, straightforward message: **AI ethics goes beyond compliance, and can be applied proportionally, regardless of how your AI system is classified.**

This playbook is intended as a beta version, and a living document. It will continue to evolve with the growing capabilities of AI, input from users, stakeholders, and interested parties. Your feedback is essential: please share your thoughts with us at info@ai4people.org.

Intended Audience

The AI4People Playbook is designed for a broad range of audiences involved in developing or interacting with AI systems. Whether you are new to AI ethics or an experienced professional looking for practical guidance, this playbook intends to provide valuable information, hands-on activities, and a curated list of resources to support you.

Some sections are tailored to specific groups to better address their specific roles and needs. To help you navigate the content, we have identified such sections with clear icons for:



AI DEVELOPERS



GENERAL AI USERS



LAWYERS USING AI



JOURNALISTS USING AI



POLICYMAKERS



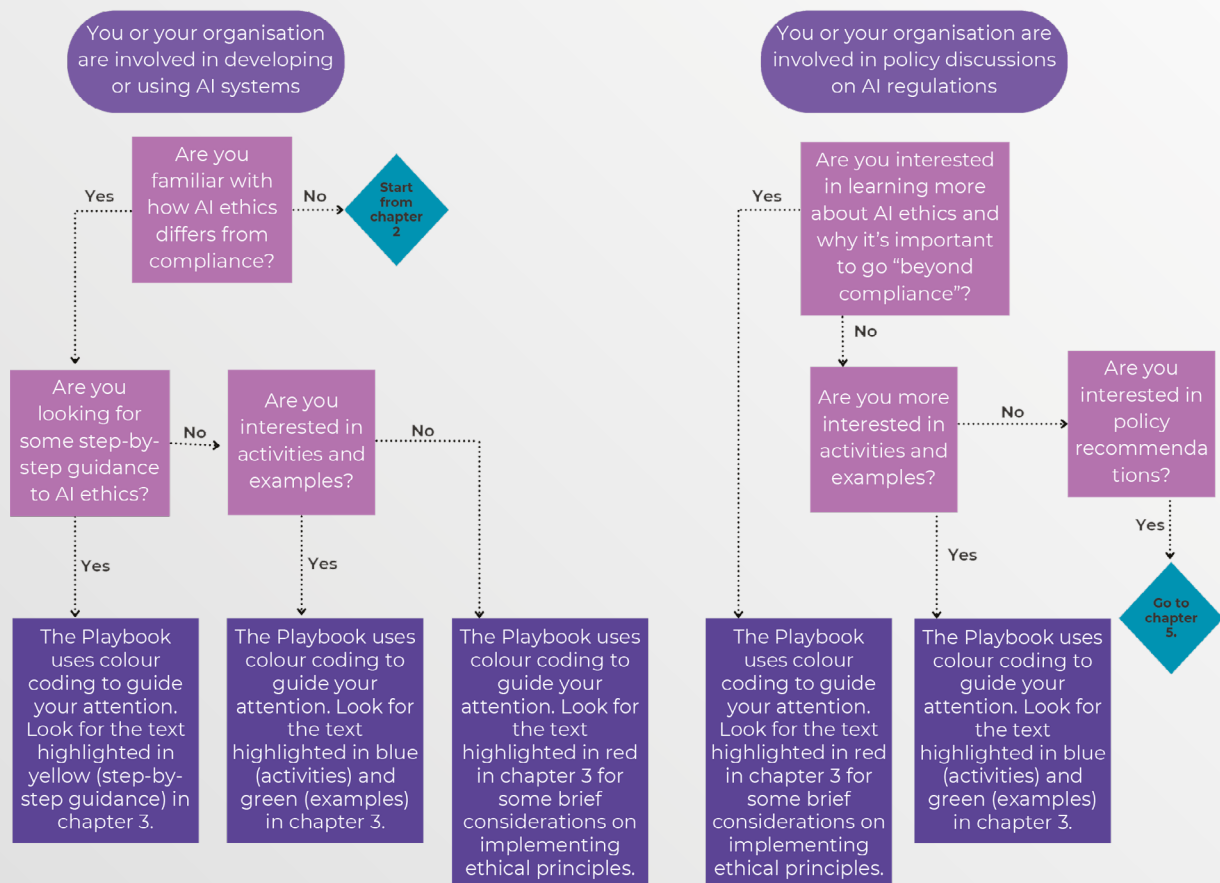
STUDENTS



ORGANISATIONS



Moreover, most sections follow the same structure for consistency, and the content is colour-coded to make navigation easier. Use the icons and colour code to quickly find the sections most relevant to you and get the most out of the Playbook, but don't stop there: take time to explore other parts too. You might be surprised by what new insights you will discover!



How the AI4People Playbook interacts with other resources

There are many AI playbooks and similar resources offering practical guidance (see [Annex 1](#)). The **AI4People Playbook** stands out in three ways:

- **Focus on ethics, not just compliance:** it moves beyond regulations and legal checklists to encourage deeper ethical reasoning.
- **Inclusive audience:** it is designed for AI developers, policymakers, and general users alike, rather than targeting only specific sectors or professionals.
- **Practical engagement:** it enables readers to *practice* ethics through activities and



examples, rather than simply reading about principles.

How AI tools were used to support the drafting process of the AI4People Playbook

We recognise that the use of AI in academic and policy writing is an evolving practice. Accordingly, we are committed to following established ethical standards and guidelines, such as those from the Committee on Publication Ethics (COPE) and the American Psychological Association (APA), and the *Guidelines for ethical use and acknowledgement of large language models in academic writing* published by Porsdam et al. (2024)¹. We encourage readers of this playbook to adopt the same commitment to openness and integrity.

AI tools were used to *support* and *enhance* the drafting process of the AI4People Playbook, not to replace it: the Playbook is the product of a collaborative effort by a dedicated working group, which actively drafted each page and met regularly to refine the content and format. Moreover, the Playbook has undergone three iterations, each reviewed by the AI4People Scientific Committee.

Generative AI tools were employed to support the drafting process in the following ways:

- Enhance grammar and sentence structure, and to identify typographical errors.
- Assist in formatting and structuring the document.
- Provide stylistic suggestions to improve clarity and coherence.

The members of the AI4People Working Group Playbook are fully responsible for the content of the Playbook.

About AI4People Institute

The AI4People Institute brings together leading AI companies, academic researchers, civil society organisations, and government bodies to tackle the societal risks of AI and to champion its ethical and beneficial development. AI4People works across key domains critical to trustworthy AI: Ethics & Governance, Policy & Regulation, Inclusion & Fairness, Responsible AI Deployment.

¹ Porsdam Mann, S., Vazirani, A. A., Aboy, M., Earp, B. D., Minssen, T., Cohen, I. G., & Savulescu, J. (2024). Guidelines for ethical use and acknowledgement of large language models in academic writing. *Nature Machine Intelligence*, 6(11), 1272-1274. <https://doi.org/10.1038/s42256-024-00922-7>



Over the past eight years, AI4People has published four milestone reports that have shaped Europe's AI policy agenda, marking a clear evolution from high-level principles to hands-on, actionable guidance.

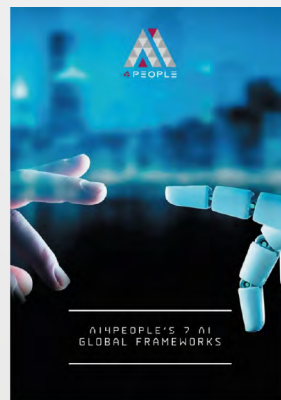
The first report laid the theoretical groundwork: it identified the societal risks and opportunities of AI, proposed five core ethical principles, and outlined 20 concrete recommendations to guide responsible development. The second report shifted focus from theory to application, examining how these principles could be implemented through real-world governance mechanisms. The third report focused on the deployment of ethical AI across seven key sectors, translating principles into domain-specific strategies. The most recent White Paper marked a turning point: it introduced a practical Ethics by Design framework to support organisations of all sizes in proactively integrating ethics throughout the entire AI lifecycle.



AI4PEOPLE'S ETHICAL FRAMEWORK FOR A GOOD AI SOCIETY: OPPORTUNITIES, RISKS, PRINCIPLES, AND RECOMMENDATIONS



ON GOOD AI GOVERNANCE: 14 PRIORITY ACTIONS, A S.M.A.R.T. MODEL OF GOVERNANCE, AND A REGULATORY TOOLBOX



AI4PEOPLE'S 7 AI GLOBAL FRAMEWORKS



AI4PEOPLE'S REPORT TOWARDS AN ETHICS BY DESIGN APPROACH FOR AI





ETHICS BEYOND COMPLIANCE

3.1. Four reasons to move beyond compliance and adopt an Ethics by Design approach

If your organisation develops or uses AI, or you are actively involved in the development and deployment of new AI systems, the importance of integrating ethical principles throughout the entire lifecycle of an AI system has never been more evident. Beyond meeting legal obligations, embracing an Ethics by Design approach can yield long-term benefits, fostering trust, enhancing reputations, and driving innovation. The following points outline why prioritising ethical considerations in AI systems is not only essential for regulatory compliance but also key for ensuring your organisation's success and positive societal impact.

1. Embracing an Ethics by Design approach is an investment in the organisation's long-term viability and reputation

- **Organisations can adhere to key ethical principles, including accountability, transparency, and respect for human rights, adopting an Ethics by Design approach in a proportionate manner**, tailoring their strategies based on the specific use cases associated with their AI systems.
- With recent data showing that, globally, people have become less trusting and more worried about AI², this approach can help **your organisation to enhance its reputation and align its operations with societal values**.
- Reflecting on guiding ethical principles can **clarify your company's mission and aspirations**, fostering a deeper alignment between its AI strategies and its overarching goals.

2 Gillespie, N., Lockey, S., Ward, T., Macdade, A., & Hassed, G. (2025). *Trust, attitudes and use of artificial intelligence: A global study 2025*. The University of Melbourne and KPMG. https://mbs.edu/-/media/PDF/Research/Trust_in_AI_Report.pdf?rev=0ee82285b2b0439bba524dbddc58214a



2. Embracing an Ethics by Design approach helps when the compliance requirements are not clear

- **An ethical approach that goes beyond mere compliance helps you to mitigate the risks associated with risk level misclassification** and to avoid significant reputational damage, financial burdens from rising operational costs, service disruptions, and potential legal liabilities.
- It also cultivates a stronger relationship with regulators, stakeholders, and the public, and positions your company as a leader in trustworthy development and use of AI systems.

3. Embracing an Ethics by Design approach fosters the development of better products

- By incorporating ethical principles and requirements since the design phase, **developers can build AI systems that are not only more reliable and trustworthy but also more aligned with societal values.**
- Using a proportional Ethics by Design approach, therefore, you can improve the overall quality of the final AI products.

4. Embracing an Ethics by Design approach promotes interoperability and easier compliance with the growing framework of laws

- As AI technology continues to evolve and spread across industries and nations, each country is adopting its own regulatory frameworks, often shaped by its unique legal, cultural, and economic context. These **regulatory divergences pose important interoperability challenges.**
- While these regulatory frameworks may differ, they are underpinned by some shared ethical principles. Comprehensive reviews³ of AI ethics discussions indicate that, although values can vary between regions, core principles such as fairness, transparency, and privacy are widely recognised. Embracing a proportional Ethics by Design approach, therefore, accelerates compliance and promotes interoperability.

3 Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399. <https://doi.org/10.1038/s42256-019-0088-2> Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), 99-120. <https://doi.org/10.1007/s11023-020-09517-8>





A PROPORTIONAL APPROACH TO ETHICAL REQUIREMENTS

What does it mean to apply a proportional approach to ethical requirements?

A proportional approach to ethical requirements means tailoring the depth of ethical design, scrutiny, and governance to match the potential ethical impact or complexity of an AI system.

3.1 AI Ethics Impact Assessment & Risk management systems

Why are Ethics Impact Assessment & Risk Management Systems important?

AI systems, regardless of their application domain and the related potential consequences can negatively impact individuals' rights, privacy, and society at large. Completing an Ethics Impact Assessments (EIA) to understand what impact an AI system might have, and defining measures to help reduce the likelihood or impact of any possible risks through an AI Risk Management System (AIRMS) can ensure that AI development and deployment remain ethical, responsible, and aligned with ethical principles, independent of how the system is classified under regulatory frameworks such as the EU AI Act.

Key considerations before conducting an ethics impact assessment and developing an ai risk management system:

- An EIA involves a systematic evaluation of potential ethical risks (including risks to fundamental rights), An AIRMS provides the framework and mechanisms to ensure the ongoing monitoring and management of the risks identified in the EIA. It follows that EIA and AIRMS are deeply interconnected and should not be considered separate approaches.
- EIA and AIRMS should always encompass the entire AI system's lifecycle. Partial assessments, for example not including data sources or overlooking long-term implications, might result in an incomplete understanding of the impact of the AI system on fundamental rights.



- One effective approach to complete the EIA is to develop an Ethical Principles Inventory outlining the ethical principles that apply both the organisation and its AI initiatives. This inventory should incorporate relevant ethical principles and values reflected in internal codes and standards. For practical guidance, refer to our previous report, specifically, [Guidance 7: “How to Compile an Ethical Principles Inventory?”](#) and [Guidance 8: “List of Main Ethics Principles.”](#)

Steps to embracing a proportional approach:

Step 1 Risk identification and evaluation	• Create a comprehensive list of ethical principles, fundamental rights, ethical considerations, challenges, and potential conflicts that must be taken into account when evaluating the impact of the AI system. Guidance 8 of our previous report offers a list of “minimum recommended” ethical principles that should be considered as part of an ethics impact assessment. • Analyse and assess whether the AI system could potentially infringe upon ethical principles or fundamental rights, and create a list of associated ethical risks that might be relevant to your particular use case. • Some European countries published guidance and templates that might guide a fundamental right impact assessment, for instance the Danish Institute for Human Rights’ Guidance on Human Rights Impact Assessment of Digital Activities, and the Fundamental Rights and Algorithm Impact Assessment developed by the Dutch Ministry of the Interior and Kingdom Relations. The project ALIGNER, funded by the European Union, released a publicly available Fundamental Rights Impact Assessment template.
Step 2 Risk management measures & residual risk assessment	• Consider the ethical risks identified in Step 1 and identify management measures to help reduce the likelihood or impact of the identified risks. • Some examples of control measures include monitoring, preventing, mitigating, or compensating for the ethical impacts and effects of the system.



Step 3 Providing Documentation for EIA Completion	• Share with the deployer(s) of the AI system information regarding how the system will be used, who might be affected, the potential risks, the human oversight measures and the defined actions to take if negative impact occurs.
Step 4 Monitoring and Reporting	• Establish a clear process for monitoring and evaluating the performance of the AI system, as well as the progress of mitigation measures. • This represents a critical step in ensuring that AI systems operate ethically and responsibly.

Further Guidance:

- The [ISO/IEEE 24748-7000](#) standard offers practical guidance on how to comprehensively evaluate the applicability of ethical principles to specific AI use cases.
- It is important to acknowledge the ongoing efforts to establish a more standardised approach to AI Risk Management. The [NIST AI RMF](#) and the [ISO/IEC 23894](#) standard provide valuable resources for organisations seeking to implement robust and comprehensive RMS for their high-risk AI systems.
- Watch this space: the AI Office will develop and publish a questionnaire template to guide the conduct of FRIAs.

What are the requirements of the EU AI Act?

If you are a deployer of high-risk AI systems, you might have to carry out a Fundamental Rights Impact Assessment (FRIA) before putting the high-risk AI system into use. This obligation applies to entities governed by public law, entities providing public services, and deployers of AI systems that evaluate the creditworthiness, establish a credit score or are used for risk assessment and pricing regarding life and health insurance.

If you are a provider, you must implement a risk management system throughout the entire lifecycle of the AI system.



A Practical Example – AI Risk Management System

To better examine fundamental rights and threats associated with an AI system, Banca Intesa Sanpaolo categorises them into clusters, and creates a Threat-Rights Matrix to show the relationships between the identified threats and the fundamental rights they could potentially violate.

Following the identification of risks, Banca Intesa Sanpaolo selects risk safeguards and develops a Safeguards-Threats Matrix to show how such measures can reduce the likelihood or impact of identified threats. For each element of the Matrix, the planned safeguards and their relative effectiveness is detailed. The overall risk is then recalculated, with determination of the Residual Risk Final report printing with inherent risk, residual risk, and economic impact.

When Banca Intesa Sanpaolo is the AI provider, the information needed to complete the FRIA are then shared with the deployer. This information includes how the system will be used, who might be affected, the potential risks, the human oversight measures and the defined actions to take if harm occurs.

Finally, the operation of the high-risk AI system based on the user instructions is monitored, and incidents or risks are reported to the Supplier or Authorities.

A Practical Example - Ethics Impact Assessment

To assess the ethical impact of AI systems, Fujitsu has adopted a structured approach based on two key assumptions:

1. To be effective, ethical principles and requirements have to be translated into detailed specifications that clarify who is affected by whom, and how, in terms of ethical risks within a specific use case.
2. Ethical risks arise from interactions between different elements of the AI system, such as the system components (e.g., training datasets, AI models, outputs), the stakeholders directly involved (e.g., business users), and the stakeholders indirectly impacted (e.g., citizens).

Based on these assumptions, Fujitsu has developed an AI ethics impact assessment methodology. This methodology employs a mapping of interactions among AI system components and both direct and indirect stakeholders throughout the AI system's lifecycle to automatically assess ethical requirements and identify potential risks, using the ethical requirements outlined by ALTAI. It also contextualises these risks based on the specific AI use case under review.





3.2 Human oversight

Why is human oversight important?

Human oversight measures are key to ensure humans can oversee and control, to varying degrees and in different ways, the decision-making processes of AI systems. Human oversight enables moral reasoning and complex ethical considerations (imagine, for instance, legal professionals who may use AI-generated risk scores to inform decisions, but who ultimately apply their own expertise to ensure fairness), and also ensures accountability for AI-driven actions. Human oversight has long been recognised as a fundamental ethical principle, particularly since the AI High-Level Expert Group (AI HLEG) identified it as one of seven key requirements for ensuring the trustworthiness of AI systems.

Key considerations before implementing human oversight measures:

- All AI systems require some forms of human oversight. Even in low-risk systems, where immediate risks might appear minimal, human oversight ensures they stay within operational boundaries and function ethically. An example is social media's AI-powered recommendation algorithms, which suggest content based on user behaviour. Over the past decade, these systems have been criticised for promoting extreme or harmful content, as they prioritised engagement without adequate human oversight, leading to increasingly radical suggestions.⁴
- Human oversight is not a one-size-fits-all measure with fixed characteristics. There is a broad range of possible oversight measures to ensure that AI systems stay within intended operational constraints, and that humans are able to control and override its outputs if necessary. You can consult Annex 2 for an overview of the different Oversight and Alternative Models.
- In general, appropriate training in human oversight is crucial to ensure operators in your organisation can monitor, interpret and correct AI outputs, especially in high-risk contexts.

⁴ Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & Society*, 35(4), 957-967. <https://doi.org/10.1007/s00146-020-00950-y>



Steps to embracing a proportional approach:

Step 1 System transparency and models of human oversight	• Ensure system transparency by providing clear, accessible documentation on the AI model's design, decision-making processes, and underlying data. This helps human overseers understand how the system functions and makes decisions, enabling effective monitoring. • Compile a list of relevant human oversight models that could be applied to your AI system.
Step 2 Selecting WHAT model is appropriate	• Ensure system transparency by providing clear, accessible documentation on the AI model's design, decision-making processes, and underlying data. This helps human overseers understand how the system functions and makes decisions, enabling effective monitoring. • Compile a list of relevant human oversight models that could be applied to your AI system.
Step 3 Establish WHEN the oversight is to be performed	• Select the model of human oversight that aligns with the specific characteristics of the AI system, considering factors such as the operational context, as well as the potential frequency and severity of adverse consequences.
Step 4 Establish BY WHOM the oversight is to be performed	• Reflect on the expertise (knowledge and skills) and authority (decision-making powers) needed by the human overseers. • Identify the overseers who possess the necessary expertise, can detect and respond to issues, and are empowered to take corrective actions in mitigating identified risks or errors.





Further Guidance:

- Annex 2 for an overview of the different Oversight and Alternative Models
- The video “[What is human in the loop?](#)” published by the Local Government Association (UK) for an introduction to Human in the Loop.

What are the requirements of the EU AI Act?

The European AI Act emphasises the importance of human oversight, specifying that it aims to mitigate risks to health, safety, and fundamental rights that could arise when high-risk AI systems are used as intended or in conditions of reasonably foreseeable misuse. It mandates that AI systems must be designed to enable human control or intervention, ensuring effective human oversight during their operation. Under Article 14(1), AI systems must be developed to allow natural persons to oversee their functionality effectively throughout their usage.

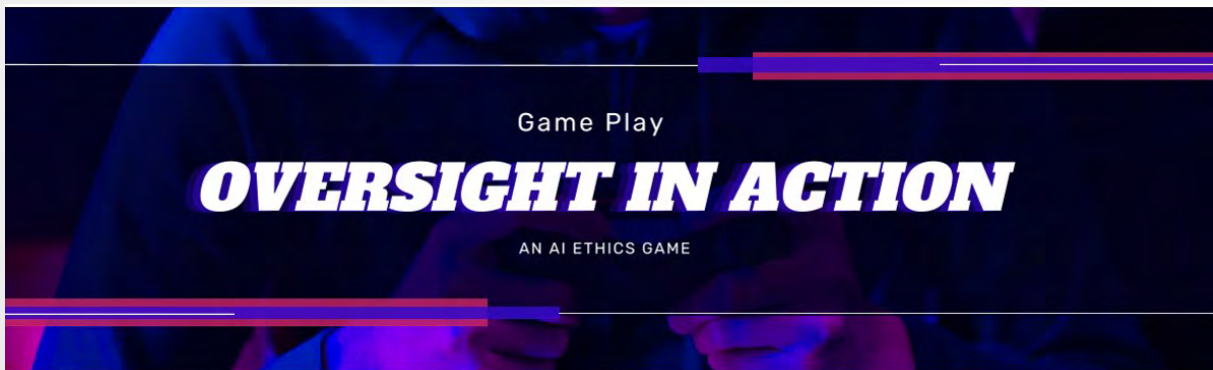
A Practical Example - Stakeholder-in-the-Loop Fairness

To promote human oversight and stakeholder participation, Fujitsu adopted a three-layered framework for oversight: governance, socio-technical, and technical. The company developed, in partnership with the University of Glasgow, Stakeholder-in-the-Loop Fairness (SILF) to support human oversight in the socio-technical and technical layers. In the socio-technical layer, the system processes inputs in a verifiable form.

SILF incorporates stakeholder feedback into AI models by converting opinions into numerical data, ensuring AI aligns with stakeholder needs. SILF’s core aim is to unify stakeholder perspectives, ensuring AI systems are transparent and agreeable to all, regardless of technical expertise, and to find common ground between diverse interests.

SILF can be used both during development and post-development. Before deployment, developers ensure that all stakeholders agree on how the system operates (e.g., the balance struck between accuracy and fairness in a loan application system). After deployment, agreements can be revised as opinions evolve, allowing developers to adjust the model accordingly.





Game 1: Oversight in Action

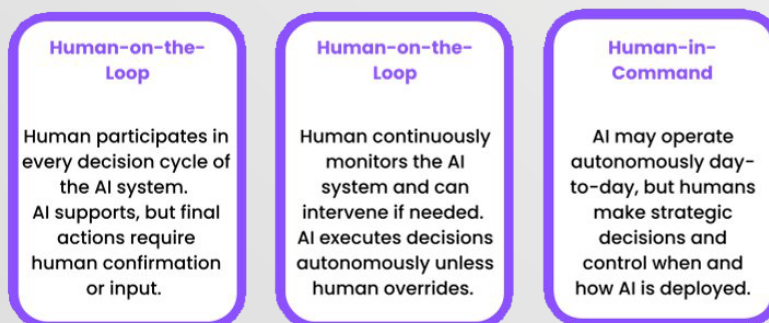
Objective:

The objective of the game is to match AI system scenarios with the most appropriate human oversight approach to ensure the ethical and responsible use of AI. Players must think critically about the risks, challenges, and potential consequences of autonomous systems while understanding the vital role of human intervention in mitigating these risks.

Game Components (please see the Annex “Game 1”):

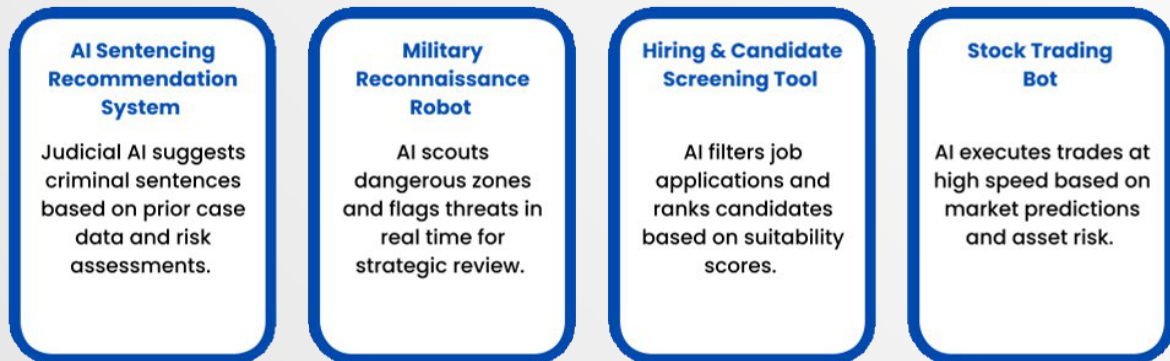
- Scenario Cards (AI Systems):

These cards describe different AI-driven systems used in various sectors and use cases (e.g., finance, healthcare, military). Each scenario card presents a situation where AI plays a key role in decision-making but needs human oversight.



- Oversight Approach Cards (Human Roles):

These cards describe different human oversight approaches, such as Human-on-the-loop (Human oversees and monitors AI operations but doesn't intervene unless necessary), Human-in-the-loop (Human involved in every decision cycle of the AI system), Human-in-command (Human has ultimate authority over the system's overall use and impact, including strategic oversight).



How to Play:

Setup:

- Shuffle the *Scenario Cards* and deal 3 to 5 cards to each player (or a group, depending on the number of players).
- Shuffle the *Oversight Approach Cards* and place them face down in a deck.

Gameplay:

1. Draw Phase:

In each round, players take turns drawing an *Oversight Approach Card* from the deck.

2. Matching Phase:

After drawing an *Oversight Approach Card*, players must match it with one of their *Scenario Cards* that they believe would be best managed using the approach described on the *Oversight Approach Card*. Players must explain their reasoning for the match.

3. Debate Phase (Optional):

After players make their matches, they can engage in a brief debate or discussion. The goal is to persuade the group why their match is the best choice. Players can challenge others' reasoning, which sparks discussions on the different oversight approaches for each scenario.

4. Voting Phase:

Once all matches are made, each player votes on the most convincing match, not voting for their own. The player with the most votes for their reasoning wins the round.





3.3 Fairness

Why is fairness important?

While regulations such as the AI Act prohibit discrimination, AI systems can unintentionally introduce or increase biases rapidly and at a large scale. This has the potential to produce real-world effects, as well as legal and reputational risks for organisations deploying these technologies. This is why fairness is so critical: while it is a common concern that AI systems discriminate, the reality is that these systems learn from data that reflects historical biases and societal inequities. The fundamental challenge of algorithmic fairness, therefore, is to actively mitigate these biases to ensure that AI models do not perpetuate or amplify unfair treatment. This requires a proactive approach from the initial design phase through the entire lifecycle of an AI system.

Key considerations to ensure Fairness when deploying AI systems:

- AI biases have real-world impacts across several sectors. In hiring, AI systems that learn from past practices may unintentionally exclude qualified candidates from untypical backgrounds, undermining diversity and risking legal consequences. In finance, historical data can lead AI models to unfairly deny loans or mortgages to certain groups, potentially breaching fair lending standards. In public safety, tools like predictive policing can misclassify individuals from specific communities as high-risk, raising ethical and legal issues while eroding trust. In healthcare, biased AI can result in misdiagnoses or substandard care for underrepresented groups, which can lead to poorer patient outcomes and reputational risks for organisations.
- You can contribute to ensure fairness by engaging in two key actions. By providing human oversight, to prevent and correct algorithmic bias that can lead to unfair outcomes. Through the Human-in-the-Loop technique it is possible to review and override AI decisions. In addition, it is important to provide developers with ongoing feedback, as this enables them to identify and address potential biases, thereby enhancing the model's fairness for all users.

Key considerations to ensure Fairness when developing AI systems:

- AI can be unfair due to biases in the data it is trained on or due to biased decisions made during its development. At other times, the problem is technical, such as insufficient data for certain groups or user experiences that do not meet everyone's needs. Therefore, given the complexity of underlying causes, it is impossible to completely "debias" an AI system. Instead, the goal is to reduce fairness-related harm as much as possible.



Steps to embracing a proportional approach:

Step 1 Selection of suitable fairness metrics	• Prior to initiating the training phase, you should assess potential model biases and select suitable fairness metrics. There are numerous fairness measures, which can be grouped into two principles: • Individual Fairness: This principle requires that similar individuals are treated similarly. An algorithm should not use protected attributes, such as race, gender, sexuality, or religion, as predictive factors. For example, an algorithm assessing a customer's creditworthiness should not consider their citizenship. While this effectively captures the concept of fairness, it is complex to measure because it requires calculating a measure of "similarity" or "distance" between individuals based on numerous factors. • Group Fairness: This principle focuses on ensuring that an algorithm's outcomes are equitable across different groups defined by protected attributes. This means that a statistical measure of outcomes, such as the percentage of loans granted, should be similar for all groups. Group fairness provides a crucial warning about potentially different outcome distributions that can be evaluated to prevent discrimination. These measures are simple to calculate and "objective," making them the most frequently used.
Step 2 Application of Fairness through Unawareness during the model development phase	• During the development phase, to mitigate the risk of discrimination, a common technique is Fairness through Unawareness, which involves excluding protected attributes from the model's development, making it "blind" to potentially discriminatory factors.



Step 3**Ongoing monitoring**

- You should track the adherence to fairness throughout the AI system's production lifecycle. Biases can develop over time, such as those resulting from changes in data (data drift), even if the model has undergone post-development testing and verification.
- Ongoing monitoring of the model is necessary, which may include the use of explainers.

What are the requirements of the EU AI Act?

Article 10 of the EU AI Act mandates that high-risk AI systems should be developed using high-quality datasets for training, validation, and testing. Such datasets must be managed properly, taking into account key aspects including data collection processes, data preparation, fairness and potential biases, and data gaps. The datasets should be relevant, representative, error-free, and complete as much as possible, considering also the specific context in which the AI system will be used.



Game 2: What type of bias?



Objective:

The objective of the game is to match real-world fairness incidents to the specific type of bias that produced them.

Game Components (please see the Annex “Game 2”):

- Real-world Fairness Incident Cards: see some examples below

Real-world incident card	Real-world incident card	Real-world incident card	Real-world incident card
AI systems employed by local councils in England may underestimate women's health concerns during social care assessments. In a study, researchers tested these tools with real case notes, altering only the patient's gender. The findings showed that men's cases were frequently labelled with serious descriptors such as 'disabled' or 'complex,' whereas women's cases were summarized in less severe language.	In the UK Biobank, most of the participants are healthy,, consequently the biobank does not provide an accurate representation of the patients typically receiving healthcare services.	Large language models frequently misinterpret everyday expressions and stylistic features characteristic of African American online communication, owing to a mismatch between the datasets used for training and the lived practices of the language	An algorithmic risk assessment computed the prediction of a person's recidivism risk by using prior arrests and arrests of friends and family as proxy variables. The use of such mismeasured proxies caused higher rates of false positive for African American versus white citizen defendants. For instance, an analysis in Broward County showed that black non-recidivist defendants were almost twice as likely to be classified as higher risk compared to their white counterparts.



- *Bias Cards*: see some examples below

Measurement Bias

This type of bias arises from the way certain features are defined, captured, or applied in a system. It often occurs when inaccurate or imperfect proxy variables are used in place of the intended measure.

Feature Selection Bias

This type of bias arises when the features included in the training dataset are not adequate or equally relevant across all subgroups.

Selection Bias

This type of bias occurs when choices about which data are gathered or are included unintentionally advantage particular groups or traits.

Sampling Bias

This type of bias arises when the data are collected through a non-random process, leading to an unrepresentative sample.

Confirmation Bias

This type of bias occurs when data are selected, interpreted, or prioritised in a way that reinforces existing beliefs or expectations.

Representation Bias

This type of bias arises from the lack of diversity in training datasets.

How to Play:

Setup:

- Each player (or group, depending on the number of players) receives the 6 *Bias Cards*.
- Shuffle the *Real-world Fairness Incident Cards* and place them face down in a deck.

Gameplay:

1. Draw Phase:

In each round, players take turns drawing a *Real-world Fairness Incident Cards* from the deck.

2. Matching Phase:

After drawing a *Real-world Fairness Incident Card*, players must match it with one of their *Bias Cards*. Players must explain their reasoning for the match.

3. Debate Phase (Optional):

After players make their matches, they can engage in a brief debate or discussion. The goal is to persuade the group why their match is the best choice. Players can challenge others' reasoning.





3.4 Accuracy, robustness and cybersecurity

Why are accuracy, robustness and cybersecurity important?

AI accuracy, robustness, and cybersecurity are essential to ensure AI is used responsibly and equitably, minimising harm and promoting trust in technology. Accurate AI prevents biases and errors that could lead to unfair or harmful outcomes. Robustness ensures that AI systems can function reliably in diverse conditions, reducing the risk of failures that could negatively impact people's lives. Cybersecurity protects personal data and prevents malicious exploitation, helping maintain individuals' rights to privacy and security.

Key considerations before implementing accuracy, robustness and cybersecurity measures:

- Regardless of the risks associated with an AI system, accuracy is crucial: unreliable systems can cause harm, albeit not always to the same level. For instance, a chatbot may give a hypothetical customer incorrect information, potentially leading them to make a disadvantageous choice. Another example could be an email filtering system allowing a phishing message through, enabling the defrauding of a vulnerable user.
- Robustness is important for all AI systems for similar reasons. The world around us is constantly changing, therefore AI systems need to handle unexpected situations or unusual information. This might include new datasets that contain errors, missing values, or even damaged files. Robustness means that the AI can still work properly when this happens. This is especially important because sometimes data can be deliberately altered to make the system give false or harmful results. For example, a financial services chatbot could accidentally reveal information about other customers. Similarly, an online gaming AI might transfer valuable items between players in ways the rules do not allow.
- An AI system that is vulnerable to cyber-attacks can also lead to harm. For example, chatbot systems may be manipulated into making offensive statements (the Microsoft Tay chatbot is a prominent example).⁵ Perhaps more concerning, however, is the possibility of security vulnerabilities allowing unauthorised access to conversations with a chatbot, exposing users' private and sensitive information.⁶

5 Neff, G. & Nagy, P. (2016). Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication*, 10, 4915–4931. <https://ijoc.org/index.php/ijoc/article/view/6277>

6 Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023, November 26). *Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection*. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23), New York, NY, United States. <https://doi.org/10.1145/3605764.3623985>



- This does not mean that all systems should be verified or secured to the same extent. A proportional approach, informed by a comprehensive risk assessment, is a more reasonable way to proceed.

Steps to embracing a proportional approach:

Step 1 Risk and vulnerability assessment	• Define metrics for accuracy, robustness and cybersecurity, taking in account ethical considerations (such as responsibility, data protection, and reliability). • Develop a testing strategy for the AI systems based on these criteria.
Step 2 Testing	• Develop diverse input data, including noisy, incomplete and corrupted data. • Develop security testing approaches, including deliberate attempts to penetrate or subvert the system, particularly through adversarial data. • Test the system with input data that is deficient or outside of expected parameters and observe the outputs. • Attempt security breaching measures and observe whether the system is resistant to these.
Step 3 Continual improvement	• Review the AI system's performance based on the output of the test phase. • Assess whether there may have to be a trade-off between accuracy and robustness. • Ensure a continuous improvement of the AI system data.



Step 4 Ongoing monitoring and auditing	• Revisit steps 2 and 3 on a regular basis, particularly when aspects of the system environment have changed (such as revisions to an external tool or application programming interface). • Have these test results confirmed by external objective experts. • Ensure ongoing human oversight.
Step 5 Process to follow when an AI system is found to be under attack, or it is shown it lacks robustness or accuracy	• Use containment and harm mitigation strategies to prevent further harm. • Communicate honestly and in a timely manner, being transparent about what went wrong, how it happened, and who is impacted. • Respect the rights and dignity of those people impacted. Listen to the impacted community, offer opt-outs and remedies, create channels for complaints, and appeals. • Carry out an investigation to understand whether it was caused by a technical issue, an adversarial attack, human error, or governance failure. • Correct the issue and review the AI governance process to prevent future failures.



What are the requirements of the EU AI Act?

Under the AI Act, high-risk AI systems must be designed and developed so that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they do so consistently throughout their lifecycle. This places particular additional demands on developers of high-risk systems. Currently, the AI Act does not provide clear guidance on how to measure a reliable level of accuracy, robustness and cybersecurity. Additionally, the recitals to the Act make it clear that these are not static requirements: they will change as the state of the art develops and conformity schemes under the Cybersecurity Act are finalised.

The AI Act is not the only European Union law that imposes cybersecurity obligations. The Network Information Security Directive (recently revised and extended) requires entities in critical sectors to put in place a range of security measures. For banking and financial markets, the Digital Operational Resilience Act has just come into force and has targeted rules on risk management. The Cyber Resilience Act is not yet in force but will require manufacturers of ‘products with digital elements’ to implement essential cybersecurity requirements.

A Practical Example - Cybersecurity breaches

Cybersecurity breaches are often not publicised, and it is likely that many incidents involving limited- and minimal-risk systems go unreported or even unnoticed. However, McAfee Advanced Threat Research detected an attempt to ‘poison’ their dataset of malware through the generation of many variants of a ransomware virus, many of which were not functional and, if included in the virus-signature database, would lower its real-world usefulness by generating false positives.⁷

A cybersecurity issue that may arise from the use of general-purpose AI (discussed further below) is when these are used to generate software code, they sometimes mention packages that do not exist. Malicious actors can create fake packages with the same names, using them as a “back door” to insert exploits into software. This is known as ‘slopsquatting’.⁸

7 Beek, C. (Beek, C. (2020, December 4). *Adversarial Threat Landscape for AI Systems*, ‘VirusTotal Poisoning’. Github. <https://github.com/mitre/advtmlthreatmatrix/blob/master/pages/case-studies-page.md#virustotal-poisoning>

8 Claburn, T. (2025, April 12). *LLMs can't stop making up software dependencies and sabotaging everything: Hallucinated package names fuel 'slopsquatting'*. The Register. https://www.theregister.com/2025/04/12/ai_code_suggestions_sabotage_supply_chain/



Game 3: the ethical risks of accuracy, robustness, and cybersecurity



Stage 1: Reflecting on risks (please see the Annex “Game 3”):

The facilitator splits the team into 2 small sub-groups, and assigns to each sub-group a Scenario.

The facilitator asks each sub-group to discuss the potential accuracy, robustness, and cybersecurity risks in their assigned use case for 15 minutes. They should also brainstorm strategies for preventing and managing these risks. After the discussion, the sub-groups reconvene, and a designated note-taker from each sub-group will share the identified risks and proposed solutions.

Stage 2: Reflecting on the ethical implications

The team is split into two small sub-groups once again. The facilitator provides each sub-group with a follow-up scenario and asks them to discuss for 15 minutes how they would feel if they were in those circumstances. Additionally, they should identify any ethical implications that arise from the scenario.

Once the sub-groups reconvene, the facilitator promotes a team discussion using the following questions:



- What are some common themes or patterns identified across the two case studies?
- What are the shared risks between high-risk systems (like fraud detection) and low-risk systems (like smart calendars), and why is it important to address them?
- Can you think of any situations where a failure in a low-risk AI system (like the smart calendar) could escalate and lead to significant consequences for individuals or organisations? How might these consequences differ from those in a high-risk system?

Scenario 1

A financial institution uses an AI-based fraud detection system to monitor credit card transactions and identify usual patterns that might indicate fraudulent activity. The system analyses thousands of transactions per second, learning from historical data to recognize patterns of legitimate purchases and potential fraud.

Scenario 2

A small business uses an AI-powered smart calendar system to manage meeting schedules for employees. The AI system analyses available timeslots, employee preferences, and meeting priorities to automatically schedule appointments, optimise work hours, and minimise scheduling conflicts.

Scenario 3

You belong to a minority ethnic group and live in a low-income neighbourhood. You have to make some essential but larger-than-usual purchases for your family, including a laptop for your child's education and a washing machine. These transactions, which are outside your typical spending pattern, are flagged by the bank's AI-based fraud detection system as potentially fraudulent. A few hours later, you receive a call from the bank's fraud department, informing you that your purchases have been temporarily blocked for verification. After explaining that the purchases are legitimate, you go through a complaint process and discover that the AI system didn't recognise your spending behaviour as typical, primarily because it was trained on data from higher-income and predominantly white neighbourhoods.





3.5 Quality management system

What is a quality management system and why is it important?

A quality management system (QMS) is a structured framework of policies, processes, procedures, and resources that organisations use to consistently deliver products or services that meet customer requirements and regulatory standards. It provides an invaluable set of tools for managing risks and ethical considerations, and ensuring quality and reliability. Establishing clear AI policies, conducting thorough impact assessments, and effectively managing the entire lifecycle are vital for developing and operating reliable and ethical AI systems.

Key considerations before implementing a quality management system:

- The QMS framework needs to be effectively adapted to the unique characteristics of the AI system to drive continuous improvement throughout the AI system lifecycle.
- The inherent ability of AI systems to learn and adapt continuously demands a flexible and responsive QMS capable of managing evolving requirements.
- It is important to monitor the proportionality and effectiveness of a QMS by tracking key performance indicators, gathering input from stakeholders and end-users, and benchmarking it against standards.



Steps to embracing a proportional approach:

Step 1 AI Policies and organisational structure	• To create a robust QMS, define and document: ○ <u>ethical guidelines</u> ○ <u>data governance procedures</u> ○ security protocols ○ a clear lines of responsibility for all aspects of AI development and operation.
Step 2 AI System Risk Assessment	• Conduct <u>a detailed assessment</u> of the ethical, operational and technical risks of an AI system. • In particular, the assessment should examine: • the ability of the AI system to trace and explain decision-making processes • the risks associated with data accuracy, consistency, and protection • the impact of the AI systems on business processes, society, and the environment.
Step 3 AI System Lifecycle Management	• Ensure effective management of the AI system lifecycle, from initial planning to eventual disposal, by: ○ determining responsible resource allocation ○ defining clear processes for each stage ○ establishing accountability. • This approach prevents delays, controls costs, ensures high quality and reliability, and simplifies the process of identifying the root causes of any problems that may arise.



Step 4**Documentation
and continuous
monitoring**

- Create a transparent procedure to document AI system development and deployment, referring to:
 - development documentation
 - compliance documentation
 - incident logs.
- Make sure your QMS supports continuous assessments against performance objectives.

Further Guidance:

- [ISO 42001](#) provides guidelines for organisations using AI systems, focusing on the establishment, implementation, and continuous improvement of these systems. The best way to understand ISO 42001 is to examine its clauses and structure: the first three clauses outline the scope, key terminology, and definitions essential for compliance. The subsequent seven clauses cover requirements such as understanding the AI system's context, establishing strong governance, planning, ensuring proper resources, adhering to key principles, evaluating performance, and taking corrective actions to improve over time.

What are the requirements of the EU AI Act?

The EU AI Act requires high-risk AI system providers to establish a Quality Management System to ensure compliance and accountability. This system is crucial for preparing technical documentation, conducting assessments, and managing AI systems throughout their lifecycle. Companies already regulated under other EU laws can streamline compliance, while those new to such frameworks may face challenges.





3.6 The ethical development and use of Generative AI

Why it is important to develop and use Generative AI models responsibly?

Generative AI brings many potential opportunities across different industries, but it also comes with real risks, including creating and spreading disinformation, violations of intellectual property rights, privacy concerns, environmental harm, and reinforcing biases. If not handled carefully, these risks can have serious impacts on society and the planet.

Some of the ethical risks are due to the limitations of the technology itself, while other risks come from how Generative AI is used, whether on purpose or by accident, in ways that go against ethical principles. It is therefore important for both the Generative AI model developers and those using them, whether in business, education, or civil society, to act responsibly.

The context of the model matters. Generative AI can create a wide variety of formats, text, images, video, music, and voice, and can be used for many purposes. These uses can cause harm intentionally or unintentionally: for example, they could be part of a phishing scam (where faked audio is used for fraud), or a person with mental health issues could have unhelpful conversations with a chatbot instead of seeking medical assistance.

Key considerations before developing Generative AI models:

- **Fairness risks:** Fairness is a key ethical principle that should guide the design of every AI model. One of the biggest challenges when developing Generative AI models, is that the models learn from the data we give them, which often reflects human biases. Not addressing this risk upfront by intentionally testing for bias, using diverse datasets, and building diverse teams that can bring different perspectives, can end up amplifying those biases.
- **Disinformation and hallucination risks:** Hallucinations and intentional misuse of Generative can lead to serious risks of misinformation. To act responsibly and reduce harm, developers should include safeguards such as content filtering and fact-checking, and explore tools such as AI watermarking, which helps identify and trace AI-generated content for greater transparency.



- **Environmental impact:** Training and running Generative AI models generally requires enormous amounts of computational power, which entails significant energy consumption and carbon emissions. Developers can limit this risk and improve energy efficiency by exploring lower-impact training methods, and considering the environmental costs as part of responsible AI development.
- **Cultural appropriation:** Respecting intellectual property rights is important, but by itself, this does not always guarantee that using certain content when developing Generative AI models is ethical. Individuals and communities may feel a connection to a body of knowledge or expression that is not formally protected by law, such as traditional cultural expressions (art, music, folktales, and so on) and developers should therefore ask themselves and their audiences whether it is moral to make use of such content.
- **Harmful content:** Could the system be used to produce content that is illegal or harmful, or could be used for such purposes? For example, could it generate non-consensual explicit images or realistic but deceptive documents? Can such output be prevented? If it is interactive, does it have safeguards to prevent it from encouraging the user to engage in hurtful behaviour, violence, or self-harm? If the user is using it for purposes beyond its safe limits, such as medical or therapeutic advice, can it recognise this and respond in an ethical fashion?

Key considerations before using Generative AI models:

- **Fairness risks:** It is important to remember that Generative AI is not born neutral: it learns from the data it is trained on, and unfortunately, that data often reflects human biases. This means the Generative AI can sometimes produce unfair or biased results. For example, some AI tools (for example a job application screening, or a tool that assigns emergency levels to patients in emergency rooms) might favour certain genders or ethnicities. Similarly, an AI-generated image tool could reinforce stereotypes by showing limited representations of people in certain roles. Whether you are a professional using AI to support some tasks, or a casual user creating content, it is important to stay alert to these biases. Question outputs that seem one-sided or unfair, avoid spreading them and do not hesitate to report them.
- **Disinformation:** Whether due to model errors (often called hallucinations) or intentional misuse, Generative AI can produce content that looks convincing but contains false or misleading information. As a user, it is important to be careful: always double-check important information you get from AI, and be cautious about sharing content that might not be accurate. Some AI tools also use special methods, like watermarking, to help identify AI-generated content. Paying attention to these signals can help you know when content is AI-made and use it responsibly. This is particularly important where you



might be an intermediate user, generating content for use by others or overseeing the use of Generative AI by others, such as a teacher who is generating lesson content and asking students to use these tools. AI literacy is an essential element of the ethical use of AI here.

- **Environmental Impact:** Using Generative AI (for instance asking it questions or generative images) requires energy, which in turn uses water and contributes to carbon emissions. For example, even something as small as sending a “thank you” message to AI adds up when millions of people do it every day, leading to significant energy and water consumption overall. While developers are working on more energy-efficient AI methods and greener technology, it’s important for all of us to be mindful about how often and how intensively we use Generative AI and to pressure AI providers to increase their green energy investments.

- **Treating Generative AI as a human/friend/companion:** When using Generative AI tools, it is very important to remember that you are interacting with a machine, not a person. Some AI systems (particularly chatbots) are designed to sound friendly and conversational, and it might be tempting to treat AI tools like a human, friend, or even companion. Remember: despite the name, AI is not intelligent, and it does not have feelings or real understanding. It simply generates responses based on patterns in data. Receiving emotional support from Generative AI, or asking for personal advice can be misleading and may lead you to harmful outcomes. Moreover, sharing personal or sensitive information with AI can put your privacy at risk, as the data you provide may be stored or used in ways you don’t fully control. To use AI responsibly, always treat it as a tool, not a trusted friend, and be careful about what you share.

Steps to embracing a proportional approach:

Step 1

Identify who might be affected by a Generative AI system

- Think about who might be harmed by potential uses, moving through an increasingly smaller scale, from all humans to individual persons.
- Start with the overall global and societal issues that might arise (such as large-scale disinformation); then smaller but distinct groups (such as cultural and language communities), and then the individual users.
- Put together a list of groups and individuals that are part of the circle of ethical concern for this Generative AI initiative



Step 2**Develop scenarios**

- For each of the groups and individuals that have been identified, identify reasonably likely scenarios in which they could be harmed.
- Members of the team could imagine that they are a member of an affected group and explain how this generative AI tool could negatively affect them
- There are three phases of development for Generative AI tools: gathering training data, training, and deployment/use. Ethical issues can arise in all three phases.
- Ask how the intended use of the tool could be harmful; whether the tool could be deliberately mis-used; and if it could cause unintended harm.
- You could generate ideas by asking team members to take on different idea-generation roles – pretending to be a cyber-criminal, for example.
- Themes to explore could include fairness, safety, harmful content, errors in the output, privacy, intellectual property rights, interaction with the AI (including self-harm, dehumanization, and manipulation), security, criminal use, and sustainability.
- This phase could be competitive, with some acknowledgement or prize for the person who generates the most ideas (although remember that they must be reasonably likely).

Step 3**Develop prevention and mitigation measures**

- For each scenario developed in the previous step, talk through how the risk can be eliminated, for example, by refusing to produce certain types of content, or be mitigated, for example, by clearly watermarking the outputs from the system.
- Integrate these measures into the overall project plan.



Step 4**Monitoring and reporting**

- Establish a clear process for monitoring and evaluating any harms that are inflicted by the Generative AI system, as well as the progress of mitigation measures.
- Ensure that there is a regular cycle of reporting on these to the project team, particularly if an iterative development method (such as agile) is being used.

Further Guidance:

- The World Economic Forum has prepared high-level practical [“Recommendations on Responsible Generative AI”](#).
- Smith and others have developed a [“Generative AI Ethics Playbook”](#), which provides much more detailed guidance on the steps outlined above.
- Researchers at Samsung have developed a [“Framework for Evaluating the Ethics and Trustworthiness of Generative AI”](#), which provides a detailed checklist and a method for quantifying and assessing ethical issues and risks for foundation models.

What are the requirements according to the EU AI Act and other legislation?

According to the EU AI Act, providers of General-purpose AI (GPAI) must draw up technical documentation and publish a sufficiently detailed summary of the training content. If the GPAI model poses systemic risk, the provider has additional responsibilities such as performing model evaluations and assessing and mitigating possible systemic risks. When GPAI is integrated into a high-risk AI system, such as an autonomous driving system, different actors (including the provider, deployer, distributor, and importer) are all required to assume varying degrees of legal responsibility. When GPAI is used in limited-risk scenarios, both the provider and the deployer are required to ensure that users are aware they are interacting with an AI system.



A Practical Example – The use of historical figures

OpenAI's video generation tool, Sora, does not allow the insertion of living persons into clips without their permission but permits the use of historical figures. This has led to the production of some controversial videos,⁹ with the daughter of Robin Williams asking not to be sent any more deepfakes of her deceased father,¹⁰ leading to concerns that trust in the media could be eroded while legal processes cannot keep pace with rapidly-advancing technology.¹¹

A Practical Example – The re-use of traditional cultural expression

The art of aboriginal and indigenous communities around the world is distinctive. Some GPAI tools are trained on these works and produce digital facsimiles of them. However, although they can seem to be abstract to the uninformed observer, the patterns are intended to carry meaning which the GPAI output omits or does not render correctly. IP law is not relevant, as these aspects of the work are not protectable. However, this re-use of traditional cultural expression may be unethical as it does not respect the integrity of community discourse¹² and diverts income that should go to indigenous artists.¹³ This has led to calls from indigenous communities for 'data sovereignty'.¹⁴

9 Nusca, Nusca, A. (2025, October 13). *Sora videos of deceased celebrities spark backlash*. FORTUNE. <https://fortune.com/2025/10/13/sora-videos-of-deceased-celebrities-spark-backlash/>

10 SaundSaunders, E. and BushbyBushby, H.. (2025, October 8). *Robin Williams' daughter pleads for people to stop sending AI videos of her dad*. BBC. <https://www.bbc.com/news/articles/c0r0erqk18jo>

11 Perlo,Perlo, J. (2025, October 12). *Are deepfakes of dead historical figures rewriting the past?*. NBC NEWS. <https://www.nbcnews.com/tech/tech-news/are-deepfakes-dead-people-rewriting-rcna235982>

12 McMulMcMullan, J. and StasiStasiuk, G. (2025, March 13). *How AI images are 'flattening' indigenous cultures – creating a new form of tech colonialism*. The Conversation. <https://theconversation.com/how-ai-images-are-flattening-indigenous-cultures-creating-a-new-form-of-tech-colonialism-246972>

13 WilsWilson, C. (2024, January 19). *AI is producing 'fake' Indigenous art trained on real artists' work without permission*. Crikey. <https://www.crikey.com.au/2024/01/19/artificial-intelligence-fake-indigenous-art-stock-images/>

14 Carlson,Carlson, B. and RicharRichards, P. (2023, September 8). *Indigenous knowledges informing 'machine learning' could prevent stolen art and other culturally unsafe AI practices*. The Conversation. <https://theconversation.com/indigenous-knowledges-informing-machine-learning-could-prevent-stolen-art-and-other-culturally-unsafe-ai-practices-210625>



A Practical Example – Copyright issues and Studio Ghibli

In March 2025, OpenAI upgraded the image generation capabilities of ChatGPT-4o. One notable feature allows users to upload photos and convert them into the distinctive style of the very well-known Studio Ghibli. This raises copyright questions, particularly where users upload images that they do not have the right to transform, but leaving these aside, there are also ethical issues. Although Studio Ghibli is probably unable to copyright its style, generated works that are similar might cause harm to its commercial interests.

More importantly, in a widely-circulated video, the founder of Studio Ghibli described AI-generated art as ‘disgusting’, and it may therefore be unethical to produce images in this style through AI when its primary creator is so opposed to its use. Many expressed strong criticism of individuals and organisations who posted outputs from ChatGPT-4o with a Studio Ghibli ‘feel’. Although AI can produce outputs that are visually almost indistinguishable from human-made works, the emotional connection between the audience and the artist is equally important. The growing interest in AI-generated art may suppress the development of human-created works, undermine the

relationship between audiences and artists, and ultimately threaten the future of human artistic expression. It therefore raises important ethical questions.

A Practical Example – The inclusion of the Romani language in Google Translate

Even when individuals make an effort to respect intellectual property rights, they may still engage in unethical behaviour or fail to consider the broader ethical implications of their actions. A prominent example is Google Translate’s addition of Romani. Some speakers of the language welcomed this development, but others were unhappy because they would prefer it to remain as a closed or ‘secret’ language, available only to members of the community.¹⁵

15 Nauendorff, M. (2024, July 2). *Google Translate’s latest update is reshaping the world — for better or worse*. Linguaphile. <https://www.linguaphilemagazine.org/editorial/google-translate-update>





3.7 Explainability

What is explainability and why is it important?

Whether or not your system qualifies as high-risk under the EU AI Act, implementing explainability and mechanistic interpretability has significant ethical and societal benefits. These principles contribute to a responsible AI ecosystem and reinforce public trust in emerging technologies.

The EU Data Protection Supervisor defines Explainable Artificial Intelligence (XAI) as “the ability of AI systems to provide clear and understandable explanations for their actions and decisions. Its central goal is to make the behaviour of these systems understandable to humans by elucidating the underlying mechanisms of their decision-making processes”.¹⁶

Key considerations before implementing explainability measures:

Beyond legal and standardisation efforts, we can consider further the ethical principles that should underpin **explainability**, including:

- **Respect for Human Autonomy:** XAI preserves individuals’ autonomy by making AI decisions comprehensible. People should be able to understand how decisions affecting their lives are made, especially when those decisions are automated.
- **Protection against Unintended Harms:** Understanding how AI systems function helps identify many risks before they cause harm. Explainable systems make it easier to detect and mitigate issues such as biased training data, flawed logic, or edge-case failures.
- **Informed Decision-Making and Consent:** Transparent, accessible explanations can help foster informed consent, particularly in domains like healthcare, finance, or criminal justice.
- **Justice and Equity:** By making model decisions traceable, XAI enables developers and auditors to examine how different groups are affected. It can reveal systemic bias or discriminatory practices, enabling corrections that help reduce inequality.

¹⁶ European Data Protection Supervisor. (2023). *EDPS Tech Dispatch Explainable AI*. European Union. https://www.edps.europa.eu/system/files/2023-11/23-11-16_techdispatch_xai_en.pdf



Moreover, XAI can enable vulnerable individuals to understand or challenge systems that can disproportionately affect them, enhancing their agency.

- **Transparency in Power Dynamics:** Making AI systems explainable helps balance power asymmetries between users and system operators, who may typically be perceived as gatekeepers of important decision-making, data representations and processing that can affect users in opaque ways.
- **Moral Responsibility:** Explainability supports moral and legal accountability for AI outcomes. When system developers, deployers, and users understand how an AI system operates, responsibility for its decisions is harder to deflect.
- **Empowering Vulnerable Populations:** Marginalized and vulnerable groups often face disproportionate harm from opaque or biased systems. Transparent and explainable systems may enable relevant authorities to audit and detect scope for safeguards for vulnerable populations that should be introduced.
- **Limiting Ethical Ambiguity and Promoting User Trust:** Users are more likely to trust AI systems when they can see how and why decisions are made. Explainability anchors trust in clarity and accountability, rather than blind faith or technical mystique.

Steps to embracing a proportional approach:

Step 1

Familiarise with the terminology, values and practices in explainability

- A team that wants to deploy an AI system in an ethical manner should undertake reflective exercises and workshops on how to design and appropriate explainability framework for their system.
- Research institutes and government organisations are working on hands-on materials, such as these [open-access guidance materials](#), to support organisations pursuing this goal.



Step 2 Look at your industry's best practices, voluntary standards and codes of conduct	• Taking the time to investigate user-centred explainability may guide you to existing technical infrastructures, including software, libraries, toolkits, benchmarks, as well as user-explanation requirements.
Step 3 Reflect on the Legal Frameworks affecting Explainable AI	• There are further implications of frameworks such as the AI Act on explainable AI, that should prompt reflections on: • Why is it relevant to explain your high-risk models' output to its recipients? Who is your audience? • How can you empower those impacted by legally or similarly significant autonomous decision-making systems?

A Practical Example – Understanding how “black boxes” work

Intesa Sanpaolo, specifically its Anti Financial Crime Digital Hub consortium and CENTAI developed a library to address the lack of current methods for understanding how “black boxes” work.

This library enables researchers and developers to:

- Compare and select from various explanation methods using a unified interface.
- Predict the performance of different explainers before even building the AI model.
- Evaluate the quality of explanations using metrics that assess factors like speed, consistency, and user-friendliness.
- Translate technical explanations into human-understandable language.

This library focuses on common AI tasks and aims to improve the transparency and accountability of AI systems by providing a more robust framework for understanding and evaluating model explanations.



XAI meets Regulation: transparency and right of explanation in the AI Act and GDPR.

The AI Act does not mandate specific forms of technical explainability. Following a technology-neutral approach, XAI techniques are not formally a legal obligation under the AI Act. However, certain requirements associated with explainability and interpretability for high-risk AI systems (specified under Article 6) can be found in Article 13 3(b) (iv):

Where applicable, obligation to provide “technical capabilities and characteristics of the high-risk AI system to provide information that is relevant to explain its output” and (vii) with an obligation to, where applicable, to “enable deployers to interpret the output of the high-risk AI system and use it appropriately”.

No further guidance has been released yet by the AI Office on these two provisions and the difference across explainability and interpretability for the purposes of algorithmic transparency in this context.

However, it may be assumed that the article introduces the concept of transparency as an enabler of explainability, for instance, technical documentation and information provided with AI systems allow deployers to understand how the system works.

Additionally, article 86 and recital 93, 171 make it clear that affected persons who are subject to HRAIS outputs listed under Annex III (apart from point 2), that carry legal or similarly significant effects:

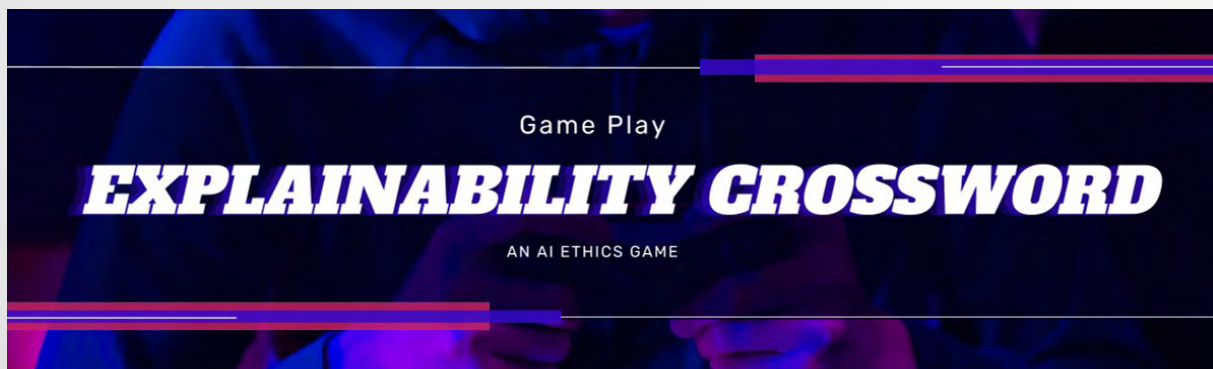
“have the right to “obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken.”

This is not a mere informational duty: it empowers the individual to understand, contest, and potentially appeal a decision that has serious legal or similarly-significant consequences.

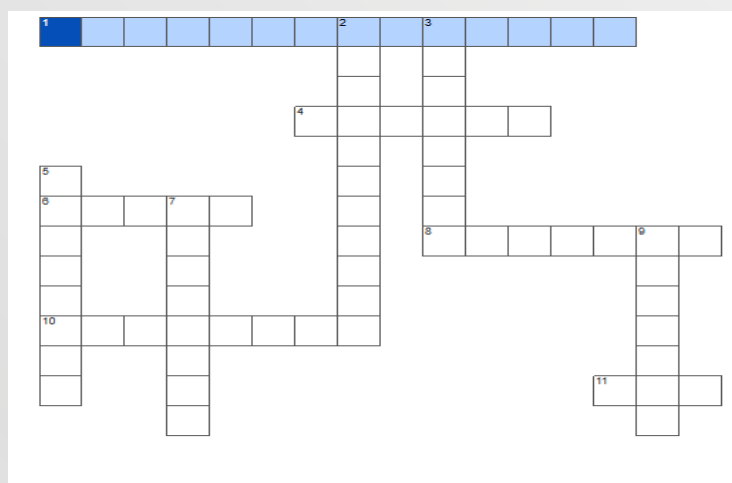
In this context, it should be noted also that Articles 13 and 14 of the General Data Protection Regulation (GDPR) set out the controller’s obligation to inform individuals about how their personal data is collected and used, while Article 15 grants the data subject the right to access that data and obtain a copy of it. Also, article 22 of the GDPR gives individuals the right not to be subject to solely automated decisions with significant effects (essentially requiring a ‘human in the loop’), and some argue that recital 71 requires safeguards and meaningful explanations of decision making. Therefore, the level of explainability which the law requires will depend on the role of the system and whether it is used to process personal or sensitive data, or can be otherwise classified as a high-risk system.



This, however, does not necessarily equate with the right to obtain an explanation through technical explainability or interpretability. Explanations may instead focus on clarifying: (a) which personal data has been used (b) how personal data has been used in automated decisions, or processing (c) which variation of data would lead to a different result (d) details to verify the accuracy of personal data, or to otherwise understand the impact of one's data. Chiefly, simply communicating that an algorithm is in use will not be sufficient if the affected person is subject to legal or similarly significant effects triggered by a HRAIS outputs or a personal data processing algorithm. The goal of these articles is not to simply breakdown every input weighed by an AI model, or the ability to technically audit code or its internal architecture but to uphold the right to challenge the automated decision or output that has this effect.



Game 5: Explainability Crossword



Across	Down
1. Users can become over-reliant on AI outputs, even when explanations are flawed. The strengths and limits of Systems must be clarified and clearly communicated. 4. Global frameworks increasingly reference explainability and transparency in AI. Laws, norms, markets, and technical design together shape expectations of responsible AI approaches. 6. Explainability should begin with local contexts, jurisdictions, and applications. Small-scale, situated efforts can build trust and compliance step by step. 8. The goal of explainability is to give people real power over automated decisions. Rights like intervention, contestation, and explanation are at the heart of this idea. 10. Ethical and legal values should be embedded directly into AI systems. Frameworks like the GDPR promote collaborative, holistic design approaches to responsible AI from the get-go. 11. Courts in Europe have stressed the right to an explanation in automated decision-making. Providers cannot hide behind trade secrets when fundamental rights are at stake.	2. People have a right to know how their data influences automated decisions. Details the model's logic, human involvement, and decision boundaries must be made clear. 3. Law, policy, and technical practice must cross-pollinate for explainability to work. Both policymakers and engineers need to understand each other's limits and tools. 5. Opaque models often decrease trust and acceptance from users. Without explainability, organisations risk losing adoption and competitiveness. 7. Explanations must be tailored for different users, not just technical experts. Too few XAI studies test their systems with real people, even though communication is key for impact. 9. In high-risk AI, making decisions understandable is seen as the only responsible path. Values like transparency, accountability, and trust are tightly linked to this principle.





3.8 Environmental Impact

Why is sustainability important?

Sustainability and environmental responsibility are two important aspects to consider when developing a new AI project. The rapid proliferation of AI technologies, particularly generative AI models, has brought significant attention to the substantial energy consumption, carbon emissions and resource utilization associated with their development and deployment.

Key considerations before deciding how to assess the environmental impact of AI:

- When discussing and assessing the environmental impact of AI, it is important to build a solid understanding of key concepts related to the environmental impacts of AI and apply them transparently. Doing so supports clear communication, while also informing the development and alignment of effective internal strategies.
- You may want to distinguish between “direct” and “indirect” environmental impact, as proposed by the OECD.¹⁷ Direct environmental impacts refers to the environmental effects from the AI compute resources lifecycle, including making the equipment, transporting it, using energy to operate it, and disposing of it at the end of its life. Indirect environmental impacts refers to the effects caused by how AI is used: this can be positive (helping climate solutions) or negative (encouraging more consumption or creating rebound effects).

¹⁷ OECD. (2022). *Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications: The AI Footprint*. OECD Digital Economy Papers, No. 341. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2022/11/measuring-the-environmental-impacts-of-artificial-intelligence-compute-and-applications_3ddddd5/7babf571-en.pdf



Steps to embracing a proportional approach:

Step 1 Consider the environmental impact in relation to the AI lifecycle	• The <i>AI resource lifecycle</i> is widely used for environmental impact assessment and is also included ISO standards (14040 and 14044). The AI resource lifecycle generally includes four phases: 1. Production 2. Transport 3. Operations 4. End-of-life • This is different from the <i>AI system lifecycle</i>, which is generally examined through the following phases: 5. Planning and design 6. Collecting and process data 7. Building and using models 8. Verifying and validating the model 9. Deployment 10. Operating and monitoring the system.
Step 2 Be transparent about the approach adopted	• Transparency about the approach selected is important to avoid misunderstanding. At the moment, there are a growing number of studies and reports that discuss carbon emissions from a variety of perspectives: some studies look at the training phase, others examine inference or analyse the environmental impact from a cooling perspective.



Step 3 Select the methodological framework to assess the environmental impact	• Some resources you may want to review, when selecting the methodological framework, are: ○ International Telecommunication Union's recommendation "AI and the Environment - International Standards for AI and the Environment"¹⁸ ○ Life cycle assessment ISO 14040¹⁹ ○ AI system life cycle ISO 5338²⁰ • Some key data that could be measured are water resource consumption and extraction, depletion of natural resources of minerals and metals, ocean acidification, cancer-related human toxicity, non-cancer-related human toxicity, fine particle emissions, ionizing radiation, and aquatic ecotoxicity.
Step 4 Include positive environmental impact	• Consider how your AI projects align with global sustainability goals and actively seek ways to enhance their positive contributions to the environment. For instance, you might: ○ Incorporate UN Sustainable Development Goals (SDGs) into the Assessment Framework. ○ Evaluate AI's Contribution to Environmental Goals.

18 International Telecommunication Union. (2024). *AI and the Environment -International Standards for AI and the Environment 2024 Report*. https://www.itu.int/dms_pub/itu-t/opb/env/T-ENV-ENV-2024-1-PDF-E.pdf

19 The International Organization for Standardization. (2006). *Environmental management — Life cycle assessment — Principles and framework*. <https://www.iso.org/standard/37456.html>

20 The International Organization for Standardization. (2023). *Information technology — Artificial intelligence — AI system life cycle processes*. <https://www.iso.org/standard/81118.html>



What are the requirements of the EU AI Act?

To achieve its environmental protection objectives, the EU AI Act establishes disclosure requirements for some systems. For general-purpose AI (GPAI) models, Article 53(1)(a) and Annex XI Section 1(2)(e) of the AI Act require all providers of GPAI models to provide information on the known or estimated energy consumption of the model. In addition, under Article 51 and recital 48, a GPAI model which has the capability for high impact on the environment could be classified as one with systemic risk and it will then have additional requirements under Article 55, including risk mitigation.

The AI Act also aims to mitigate environmental risks by establishing technical standards, with different approaches for high-risk and other systems. Under Article 40(2), the European Commission is required to issue standardisation requests to ensure that high-risk AI systems (HRAIS) and GPAI comply with the regulation. These standardisation requests may include objectives such as reducing the energy consumption of HRAIS and promoting the energy-efficient development of GPAI models. Under Article 95, for AI systems other than HRAIS, the AI Office and Member States are encouraged to develop codes of conduct that promote a range of objectives, including the voluntary adoption of requirements for assessing and minimising the impact of AI systems on environmental sustainability.

Further Guidance: Several documents addressing AI sustainability, particularly within the Information and Communications Technology (ICT) sector, have already been published. These include standards on the environmental assessment of digital services, standards on environmental assessment, and reports such as the [UNESCO Recommendation on the Ethics of Artificial Intelligence](#), the OECD's [Measuring the environmental impacts of artificial intelligence compute and applications](#), and AFNOR-Spec 2314, [General Framework for frugal AI](#). The recently launched Coalition for Sustainable AI provides an in-depth review of many of these resources on their website: <https://www.sustainableaicoalition.org/initiatives-hub/>.



A Practical Example – A framework to assess environmental impact

Nokia has developed an innovative framework²¹ to assess the environmental impact of AI systems. This framework combines AI lifecycle stages with environmental lifecycle assessments, providing a transparent and comprehensive method for evaluating the sustainability of AI systems. By leveraging existing international standards, Nokia's approach ensures that the results are globally recognised and easily adopted across industries. This transparency enables companies to identify areas where their AI systems can be optimised for sustainability, such as by adjusting the location of training processes to utilise renewable energy sources like solar power.

²¹ Kallio, Kallio, S. (2024, September 10). *A transparent and methodical framework to assess the sustainability impact of AI*. Nokia. <https://www.nokia.com/blog/a-transparent-and-methodical-framework-to-assess-the-sustainability-impact-of-ai/>





ORGANISATIONS



STUDENTS



JOURNALISTS USING AI



LAWYERS USING AI



POLICYMAKERS



AI DEVELOPERS



GENERAL AI USERS

4

THE EU'S APPROACH TO AI GOVERNANCE

The European Union AI Act, the first-ever legal framework for Artificial Intelligence (AI), entered into force on August 1st, 2024. This horizontal regulation applies to all AI systems whether or not they are embedded into another product or used independently. To create an internal market for safe and trustworthy AI along the entire value chain, it has three overarching objectives of:

- Building **trust** by treating AI as a product requiring safety and transparency.
- Fostering **innovation** by providing legal certainty, standardized requirements, and regulatory sandboxes.
- **Protecting citizens'** rights through safeguards for fundamental rights and privacy.

This legal framework encourages AI developers and deployers to innovate by reducing liability risks through legal certainty. It provides particular support mechanisms for smaller players to compete through standardisation of legislative requirements, allowances and regulatory sandboxes to assist with compliance. However, predominantly, it creates mechanisms to help enforce fundamental rights and obligations, complementing various existing EU laws (e.g., sectoral regulations, product safety legislation) across all sectors. High-level requirements enshrined in the AI Act are supported by technical standards, while tapping into existing but tailored conformity procedures with an AI-specific layer.

There are two key bodies overseeing and supporting the implementation of the Act beyond Member States, the AI Office and the AI Board:



The European AI Office	The European AI Board
• Central organ supporting the coherent implementation and application of the AI Act across member states. • Operates within the European Commission but has its own structure. The office combines expertise on AI regulation, AI safety, innovation, and policy coordination. • Regulates general-purpose AI models, provides technical support and coordinates with national authorities.	• An advisory group composed of high-level representatives from EU member states. • Brings all stakeholders together to discuss a harmonized approach to interpreting and applying the legislation. • Advises the AI Office on policy, strategy, and implementation, coordinates supervisory activities among member states, harmonises enforcement and ensures consistent application of the AI Act.

The AI Act employs a four-tier system, with a pyramid of risk, which regulates only as much and as far as required by the risk (**see Figure 1**). Technical standards are being developed by European Standardisation Organisations to ensure operational clarity for providers and deployers. Technical standards were requested in spring 2024 and are likely to be completed in 2026..



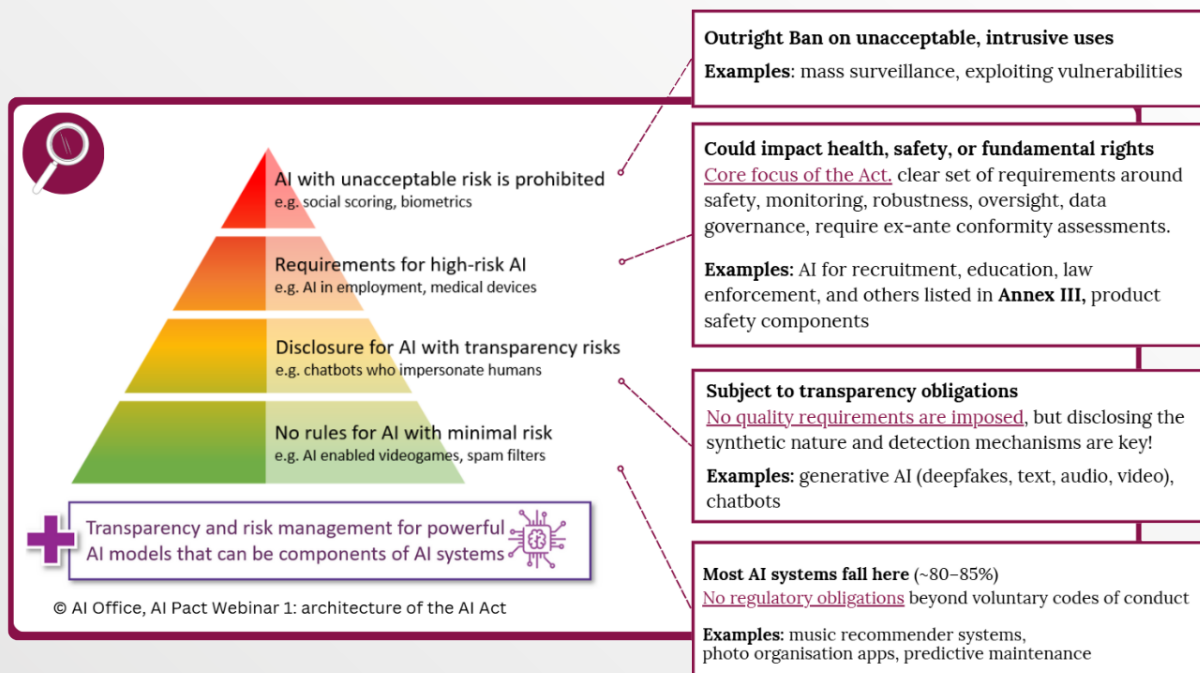


Figure 1: Four-tier Pyramid of Risk, AI Office, AI Pact Webinar 1, 28 November 2024 ([access](#))

The Act took effect on **1st August 2024**, but provisions enter into force on a phased basis until **1st August 2027**:

- **February 2025**: Prohibitions became enforceable. AI literacy obligations entered into force.
- **August 2025**: General-purpose AI model rules started, and a code of practice is available for developers to sign up to.
- **August 2026**: High-risk self-standing AI systems in Annex III must comply with technical and procedural requirements.
- **August 2027**: AI embedded in regulated products (e.g., medical devices, machinery) must fully comply with the AI Act (Annex I)



Providers of AI-Systems



- ☐ Establish a [risk and quality management](#) systems
- ☐ Comply with [data governance](#) requirements
- ☐ Draw up [technical documentation](#)
- ☐ Design high risk AI systems for [record-keeping](#)
- ☐ Ensure compliance with requirements regarding [human oversight](#)
- ☐ Provide [instructions for use](#) to downstream deployers
- ☐ Design to achieve appropriate levels of [accuracy, robustness, and cybersecurity](#)
- ☐ [Register system](#) in central EU high-risk AI database, [report incidents](#), [cooperate](#) with supervisory authorities

Am I a “Provider”? A provider is a natural or legal person marketing an AI system under their trade name (potentially not the original developer if they market it, rebrand or distribute it). For instance, those who develop or place the system on the market, such as a developer of an AI recruitment screening tool for HR.

Deployers of AI-Systems



- ☐ Abide by the provider’s instructions for use, including training staff to oversee AI correctly
- ☐ Deploy System only on representative data aligned with the intended design
- ☐ In the [Public Sector](#)? First carry out the fundamental rights impact assessment!

Am I a “Deployer”? Deployers are those who *use* the system. It can be an entity, public or private, that buys and uses or integrates the AI system under its authority (for instance, an HR department employing an AI-driven recruitment tool).

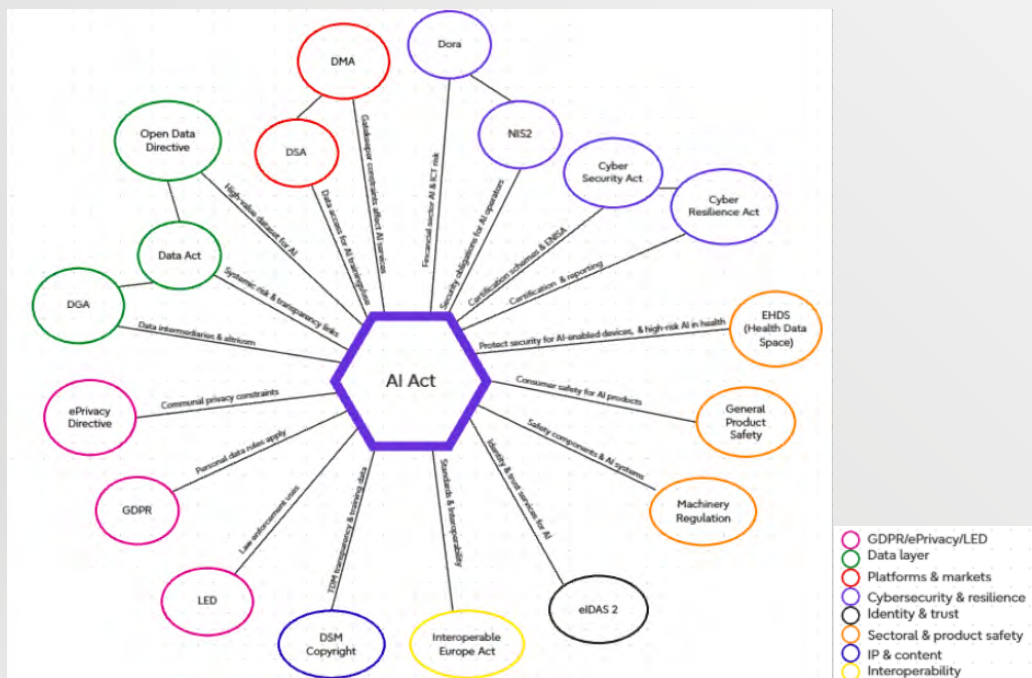
The AI Office oversees a comprehensive engagement [plan](#) known as the AI Pact allowing organisations to engage with the Commission and other stakeholders to share best practices and join activities such as the [AI Pact Events](#) (including monthly thematic webinars, tailored workshops, and public consultations). The aim of this approach is to support businesses in planning ahead of the full implementation of the AI Act and establish an [open policy dialogue on AI](#) under the European AI Alliance initiative.



The AI Act is embedded within the EU's digital agenda, which includes the following topics:

- Data
- Digital services and the digital market
- E-government, e-identity and digital euro
- Cybersecurity
- Media and democracy
- Education and skills
- Working conditions in platform work
- Digital infrastructure

Together, these actions are embedded within the EU's so-called Digital Decade, emphasizing the importance of the digital age and its significant changes to the European Union and its citizens. Thus, the AI Act must be regarded as a puzzle piece in a bigger digital ecosystem, significantly influencing how to approach ethics-by-design holistically. The core interconnections between the AI Act and other digital policies revolve around privacy protection, data sharing, the digital market, cybersecurity, identity, product safety, and content (intellectual property). For more information consult Annex 7.





RECOMMENDATIONS TO POLICYMAKERS

The AI4People Playbook aims to help all stakeholders take a practical and ethics-driven approach to integrating AI into society. The scope is broader and more time-sensitive than mere compliance with the demands of a law, which must change more slowly, to provide stability, while ethics can be more nimble. Accordingly, it should not escape the attention of policymakers that our work, being complementary to theirs, may offer critical ethical considerations that can help inform and strengthen their policymaking efforts:

- Securing organisations' long-term viability and reputation;
- Mitigating the risks of misreading legal classification of AI systems;
- Improving the quality of AI-powered products and services;
- Enhancing worldwide interoperability between global legal frameworks and forestalling regulatory divergence.

These bottom-up recommendations are meant to assist all stakeholders by providing clear, ethically grounded guidance to ensure the ethical development and use of AI. Policymakers may wish to consider how their legislative commitments can be augmented by the ethical framework and practical insights the AI4People Playbook provides, and how these insights might inform voluntary codes of practice, supporting tools, and more effective use of existing non-institutional resources.

With this in mind, we provide a summary of the suggested takeaways addressed by the playbook below.

Recommendation 1: Guidance on how to assess Generative AI models within the FRIA template.

The Fundamental Rights Impact Assessment (FRIA) template highlights the importance of regulatory compliance and the ethical use of AI. However, since its development, new Generative AI systems such as LLMs, chatbots and video and audio generation tools introduce unpredictable risks. These risks include, inter alia, the spread of false information, the use of chatbots as companions leading to psychological distress, cognitive dis-empowerment, or the creation of content including the face or voice of humans, infringing on their personal rights. Additional aspects of intellectual property, including



for instance artistic style or cultural appropriation, are not (yet) regulated. This is similar to issues of labour rights including collective action, job replacement, and potential risks to human rights from AI systems used in recruitment, hiring, performance monitoring, and termination decisions. Thus, guidance on how to comprehensively assess Generative AI systems within the FRIA template is recommended in order to adapt to the new circumstances.

Recommendation 2: Establish mandatory misuse disclosure labelling for generative AI tools.

Market accountability requires transparency about documented harms. While standards focused on technical provenance (such as watermarking) can support these transparency efforts, mandatory labelling would complement this effort by displaying verified incidents of misuse associated with each tool, including manipulation campaigns, fraud, generation of non-consensual intimate imagery, coordinated inauthentic behaviour, and other documented harms. Labels would aggregate data from platform incident reports, regulatory enforcement actions, civil society monitoring organisations, and independent security audits. This information enables informed decision-making by users, institutions, and procurement officers while creating market pressure for developers to implement robust safeguards. Labels could display incident frequencies, harm categories, severity levels, and developer remediation efforts.

Recommendation 3: Development of reporting requirements, including anonymous repositories, to encourage entities that have security issues to make these public and help develop coordinated responses.

New AI systems are associated with various forms of risks, making accuracy and robustness crucial as unreliable systems can cause significant harm. The existing reporting channels (such as the European Union's recently expanded Network and Information Systems Directive) do not provide comprehensive reporting requirements, especially for low-and medium level security issues. With AI systems evolving quickly and being vulnerable to new attacks such as 'slop squatting' (fake code packages inserted into software to create backdoors for exploitation), highlighting the importance of public incident reporting to develop coordinated responses is indispensable.

Recommendation 4: Protect and empower whistleblowers who point out issues that would otherwise be ignored.

AI systems can pose risks that insiders observe before others do. Although whistleblowers are protected by law in the EU and UK, organisations must establish clear and readily accessible internal channels for raising concerns, prohibit retaliation against those who report, and impose consequences for violations. In addition to the recent AI Act Whistleblower Tool, organisations bear responsibility for creating conditions where whistleblowing is safe and effective, not merely compliant on paper. This includes building organisational cultures that respond substantively to concerns rather than deflecting or



penalizing reporters, and providing resources, such as anonymous expert assessment and connections to legal counsel through third party entities like the AI Whistleblower Initiative (AIWI). Policymakers should examine how protection frameworks address characteristics specific to AI development: rapid iteration cycles, distributed responsibility across teams, and technical complexity that makes harm assessment difficult for non-specialists.

Recommendation 5: Encourage or require training on cybersecurity for all users of AI systems.

Currently, the AI Act states that AI system providers and deployers need to demonstrate sufficient AI literacy (applicable since 2 February 2025). However, no provisions to encourage or require cybersecurity training is foreseen, increasing the security risks around AI development, including examples like slop squatting. As systems are increasingly complex, a lack of security can increase the risks of infiltration of vital infrastructure, thus creating a need in not only AI literacy but also cybersecurity training, particularly for entities outside safety-critical sectors, which are not obliged to provide such training.

Recommendation 6: Favour AI solutions that demonstrably reduce environmental harm.

Using GenAI requires various resources such as energy and water but also contributes to overall carbon emissions. As AI is introduced increasingly in all areas of the public and private sector as well as private life, there is an indispensable incentive to work on more energy-efficient methods. Governments and policymakers with their inherent ability to steer certain developments, should do so by favouring resource-efficient AI solutions and incentivising environmentally ethical development, for instance by ensuring that governmental/public sector procurement teams are adequately equipped to understand and assess the environmental impact of AI. This could be complemented by requiring environmental impact assessment for large/novel AI projects, analogous to FRIA.

Recommendation 7: Encourage the voluntary or mandatory development of environmental impact labelling for Generative AI tools.

Related to the recommendation above is the use of green labelling for Generative AI tools, comparable to alternative search engines. The addition of voluntary or mandatory labelling not only increases the potential for greener resource consumption and again steers developers to invent sustainable alternatives, but also provides a choice to the consumer, actively participating to shape the AI market. Labelling could include criteria ranging from water and energy consumption to ethical use including, inter alia, good working conditions, cybersecurity standards, and fair use of training data (e.g., not using cultural heritage or indigenous arts, etc.). This recommendation is aligned with the existing General-Purpose AI Code of Practice.



Recommendation 8: Develop novel approaches to protecting cultural heritage.

Cultural heritage and traditional art such as the Romani language or Scottish tartan, are highly distinctive and associated with specific communities worldwide. However, cultural heritage is not protected under traditional intellectual property law and can be freely accessed to train AI models which then use the information to generate digital facsimiles. More creative approaches to the protection of traditional cultural expression must be developed, as is being done by IP Australia²² and by the CARE Principles for Indigenous Data Governance. These should bear in mind the UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage (2003) and the obligation on the European Union under Article 167 of the Treaty on the Functioning of the European Union to conserve and safeguard cultural heritage of European significance.

Recommendation 9: Financial commitment to support ethical and responsible AI research and development, and critical AI literacy.

If we wish to flourish, rather than merely survive AI proliferating in our lives, serious financial commitments must be provided to support ethical and responsible AI research and development – to ultimately support the adoption of tools that are usable, inclusive and trustworthy. This requires ring-fenced funding for critical AI literacy, ethical oversight, and community-driven capacity building, not just for accelerating AI deployment. Investments should empower educators, civil society, cultural and religious institutions, and SMEs to embed responsible practices and ensure that AI strengthen (rather than weakens) human judgement, autonomy and democratic participation. Without meaningful financial backing, actors best positioned to safeguard public interest lack the necessary resources to do so, leaving societies perpetually lagging. Prioritising education in AI ethics and critical AI literacy and safety should not be optional, as it is the foundation of an AI ecosystem and economy that supports human flourishing, protects cultural knowledge and prepares citizens to critically engage with and question technology that can impact their dignity, autonomy and rights.



ANNEX 1: AI PLAYBOOKS

Title	Authors	Link	Target Audience
The Mobile Industry Ethics Playbook	GSMA	https://www.gsma.com/solutions-and-impact/connectivity-for-good/external-affairs/wp-content/uploads/2022/01/The-Mobile-Industry-Ethics-Playbook_Feb-2022.pdf	Mobile industry professionals
Scottish AI Playbook	Scottish AI Alliance	scottishaiplaybook.com	Public sector organisations in Scotland
AI Playbook for the UK Government	UK Government	https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government	UK government bodies and public sector organisations
Responsible Use of Generative AI	UC Berkeley	https://newsroom.haas.berkeley.edu/news-release/new-berkeley-guide-helps-business-leaders-navigate-ai-ethics-amid-rapid-adoption/	Product managers and business leaders



Advancing Responsible AI Innovation: A Playbook	World Economic Forum	https://www.weforum.org/publications/advancing-responsible-ai-innovation-a-playbook/	Organisational leaders and Government leaders
--	----------------------------	---	--



ANNEX 2: HUMAN OVERSIGHT

1. Human-on-the-loop (HOTL):

Human-on-the-loop (HOTL) refers to a form of oversight where a human does not directly intervene in every decision made by the AI but is instead responsible for monitoring the system's operation and intervening if necessary. The human supervises the AI system's actions, ensuring that it functions as expected, and can step in if the system deviates from desired behaviour or encounters a situation that requires human judgment.

Key Characteristics:

- Humans oversee the AI's performance without being involved in every decision cycle.
- Monitoring and corrective action are possible, but not necessarily required at every step.
- Typically used in systems where human intervention is less frequent but still necessary in certain contexts, such as to prevent harmful outcomes.

Examples:

- **Industrial Automation:** In a factory where robots perform assembly line tasks, the AI system may autonomously control the process. However, a human supervisor monitors the system, ensuring that it operates smoothly and efficiently, stepping in to troubleshoot in case of equipment malfunctions or safety concerns.
- **Financial Trading Algorithms:** AI-driven algorithms that make high-frequency trading decisions can function autonomously, but human traders or risk managers monitor the performance and intervene if the system behaves erratically, such as in the case of market crashes or unusual trends.
- **Drones for Surveillance or Delivery:** While drones can operate autonomously for tasks like surveillance or package delivery, human oversight (e.g., by a control center) ensures that drones are on track, intervene in case of technical malfunctions, or adjust routes in case of unforeseen obstacles.



2. Human-in-the-loop (HITL):

Human-in-the-loop (HITL) refers to a model where a human operator is directly involved in every decision cycle of an AI system. In this setup, humans remain an integral part of the decision-making process, with the capability to intervene and adjust the AI's actions or outputs at each critical juncture. This is typically used in situations where the stakes are high, and human judgment is needed to ensure accuracy, fairness, or safety.

Key Characteristics:

- Humans intervene during the decision-making process.
- AI systems may propose or perform actions, but the final decision requires human approval or adjustment.
- Provides real-time oversight to maintain control over AI outputs.

Examples:

- **Medical Diagnosis:** In medical imaging or diagnostic AI systems, an AI might suggest a diagnosis based on patterns it has detected in medical images (e.g., X-rays or MRIs). However, a human radiologist would review and confirm or adjust the AI's diagnosis to ensure it is accurate before proceeding with treatment.
- **Autonomous Vehicles:** While self-driving cars use AI to navigate, HITL would involve a human driver remaining present and able to take over control in case of sudden unpredictable situations, like encountering an obstacle or a hazardous condition.
- **Military Applications:** In AI-driven weaponry systems, a human operator may be required to review and approve any action before it's carried out, ensuring the decision is ethically and legally sound.



3. Human-in-command (HIC)

Human-in-command (HIC) is the highest level of human oversight, where a human has full authority over the AI system's operation and use. This model allows the human to make the final decision on when, how, and why the AI system should be used, as well as to manage its broader impact. It includes the capacity to evaluate the AI system's ethical, legal, and societal implications, ensuring that the system is deployed responsibly and that its actions align with human values and societal norms.

Key Characteristics:

- The human has ultimate authority over the AI system.
- Encompasses strategic oversight, including ethical, legal, societal, and economic considerations.
- Human authority is crucial for determining whether to deploy the system in specific contexts, especially when the system's decisions could have far-reaching consequences.

Examples:

- **AI Governance in Autonomous Weapons:** In the context of autonomous weapon systems, human-in-command would involve military leaders or policymakers deciding when and how to deploy the weapons, considering their broader geopolitical and ethical implications. This ensures that lethal force is used only under strict human control, preventing automated decision-making from leading to unjust or disproportionate harm.
- **AI in Healthcare Administration:** A healthcare system using AI to manage patient care or hospital operations may have human-in-command oversight by a healthcare administrator or regulatory body. These individuals make decisions about the ethical use of AI, ensuring that the system is used in ways that prioritize patient safety, privacy, and fairness, while also overseeing the system's societal impact, such as addressing issues of healthcare inequality.
- **AI in Public Policy and Legal Systems:** In the case of AI systems used to assist in law enforcement, such as sentencing algorithms, human-in-command oversight might involve legal experts, ethicists, and policymakers who decide when and how to apply AI tools. These overseers are responsible for ensuring that AI does not perpetuate biases or infringe upon civil liberties, and that it is deployed in ways that are in line with societal values and human rights.



Over-sight Model	Human Role	Level of Control	Examples
Human-on-the-loop (HOTL)	Human oversees and monitors AI operations but doesn't intervene unless necessary.	Monitoring and intervention only when needed.	AI-driven algorithms that make high-frequency trading decisions can function autonomously, but human traders or risk managers monitor the performance and intervene if the system behaves erratically, such as in the case of market crashes or unusual trends.
Human-in-the-loop (HITL)	Human involved in every decision cycle of the AI system.	Full involvement at each decision point.	In medical imaging or diagnostic AI systems, an AI might suggest a diagnosis based on patterns it has detected in medical images (e.g., X-rays or MRIs). However, a human radiologist would review and confirm or adjust the AI's diagnosis to ensure it is accurate before proceeding with treatment.
Human-in-command (HIC)	Human has ultimate authority over the system's overall use and impact, including strategic oversight.	Full control over deployment, ethical and societal considerations.	In the context of autonomous weapon systems, human-in-command would involve military leaders or policymakers deciding when and how to deploy the weapons, considering their broader geopolitical and ethical implications. This ensures that lethal force is used only under strict human control, preventing automated decision-making from leading to unjust or disproportionate harm.



Alternative Model	Human Role	Level of Control	Examples
Ex-post controls	Human reviews the AI system's outputs and decisions after they have been made.	Monitoring and intervention only ex-post.	Ex-post controls to ensure a chatbot's performance is in line with quality standards and ethical guidelines.
Sampling techniques	Human randomly selects a subset of decisions or actions made by the AI system for review.	Partial monitoring and intervention	Rather than monitoring every individual marketing decision made by the AI system, the company can apply a sampling technique to periodically assess the effectiveness and fairness of the recommendations and ads.



ANNEX 3: GUIDANCE ON THE LEGAL IMPLICATIONS OF GENERATIVE AI

Some of the key legal implications of Generative AI you might want to consider are:

- **Trademark issues:** Those who make use of generative AI should be conscious that these systems can (inadvertently) reproduce not only copyrighted works but also trademarks. This is particularly likely with models that generate pictures or video. Using such output, particularly in a commercial context, is both unethical and likely to be illegal. Re-using the trademark of another business is trading at the expense of the other's goodwill and reputation. In addition, use of another's mark in commerce is at the core of trademark infringement.
- **'Fair Use' vs 'Fair Dealing':** Although the question of whether the use of copyrighted material in the training of general purpose AI or the occurrence of content that is substantially similar to existing creations in the output of such systems is widely litigated and not yet resolved, users of generative AI should bear in mind that the requirements of copyright law can vary between jurisdictions, even across the European Union. One particular example of this is the difference between exceptions to copyright. In the United States of America, the main exception is known as 'fair use', and it is wide-ranging. In European law, this right is known in English as 'fair dealing' and is more restrictive, particularly after codification in Article 5 of the Copyright Directive. Adopting US copyright norms might therefore be not be legal in a European context.
- **Cultural Appropriation:** Even if one is careful to respect intellectual property rights, one may nonetheless act in an unethical manner. Individuals and communities may feel a connection to a body of knowledge or expression that is not formally protected by law, such as traditional cultural expressions (art, music, folktales, and so on) and developers and users of AI systems should therefore ask themselves whether it is moral to make use of such content.
- **Training Exceptions:** Article 4 of the Digital Single Market Directive (DSMD) provides a copyright exception for text and data mining, including for commercial purposes. Art 53 of the AI Act clarifies that this exception applies in AI training. However, AI training involves two steps. The first step is data collection and establishing a training dataset. The second step is machine learning based on this dataset. There is dispute about whether the Article 4 exception applies to both steps, raising legal and ethical issues for generative AI developers to consider.
- **Patent Infringement:** AI training techniques are often patented. It may therefore be the case that the training phase may infringe on a patent, creating a litigation risk.



Steps to embracing a proportional approach:

Step 1 Audit IP usage	• Ensure that any discussion of intellectual property and other legal rights involves suitably qualified experts. • Put together a description of all the intellectual property that is an input to the AI system. • Determine whether or not the AI system creates intellectual property in its own right and if so, document these outputs and their IP implications.
Step 2 Assess AI and IP impacts and risks	• Determine potential legal and ethical harms from IP use. • Verify that all IP inputs are legally and ethically sourced by confirming the presence of appropriate licensing agreements and so on. • Engage in a similar verification with external sources, such as vendors or service providers. • Investigate whether the output of the system might contain unlicensed IP (particularly copyrighted content or trademarks). • Consider whether there may be ethical issues in the re-use of cultural or other content which is not protected by copyright.



Step 3 Put in place corrective and mitigation measures	• Where issues with the sources of inputs are identified, rectify them or locate suitable alternatives. • Where the output contains unlicensed IP, modify the system to remove these, or if it is not possible (such as with an externally provided system), put in place appropriate mitigation measures such as guidance for employees. • Where there are issues beyond intellectual property law, consult with affected communities to agree mutually satisfactory outcomes.
Step 4 Ongoing monitoring and auditing	• Revisit steps 2 and 3 on a regular basis, particularly if the system deals with new types of input or output, or if the law changes. • Continue to engage with legally qualified experts to keep up to date on new developments.

It should be noted that GPAI is not always applied in high-risk scenarios. In other contexts, other parties may also bear legal and ethical responsibilities. First, data providers, when constructing datasets, must comply with national laws concerning text and data mining (TDM); otherwise, they may face copyright infringement liability. In addition to legal obligations, data providers also bear a certain degree of ethical responsibility for the content they collect. They should review the data to ensure that it does not include material that could harm fundamental rights or social welfare, such as content involving plans for terrorist attacks. Secondly, besides complying with the EU AI Act, GPAI deployers may also bear legal responsibility for their actions in other contexts. For example, if a deployer uses technology to generate a synthetic facsimile of a person's appearance or voice without their consent, there may be national laws that make this illegal. In addition, those who fine-tune a foundational AI model may also be held accountable. According to the AI Act, a 'provider' is defined as the developer of an AI model. If fine-tuning constitutes a form of model development based on a foundational model, then fine-tuners may also be considered providers, and thus bear legal responsibility under the AI Act in relation to the GPAI model. Finally, end users of GPAI are generally not legally responsible for the model itself. However, if GPAI generates risky content such as materials involving fraud or terrorism, and the end user proceeds to disseminate such content, they may be held legally responsible for the act of dissemination.



Further Guidance:

- The European IPR Helpdesk provides assistance and guidance to European SMEs on IP management, protection and enforcement. Amongst its publications is a Fact Sheet on 'IP Audit: Uncovering the potential of your business', which provides further detail on the step-by-step process outlined above.
- For those wishing to engage more comprehensively with IP rights and risks, ISO 56005 may be useful. Note, however, that these are primarily focused on maximising the benefits from the IP that an entity owns or on minimising risks of IP infringement; they will not assist in identifying ethical issues or strategies to address these, as this is an under-researched topic. Therefore, it is important and necessary to reflect on what harms your activities may cause that are not covered by legal rules, perhaps by starting by asking how you would expect a creator or a community to feel about the re-use of their creation that you intended to make?



ANNEX 4: WHISTLEBLOWING

You are a whistleblower if you are a worker and disclose wrongdoing. Most of the time, this is something you've seen at work, but it can sometimes involve other situations.

The reported wrongdoing must affect others, for example some parts of the general population, and disclosure should be in the public interest.

In the European Union, UK and several other countries, whistleblowers are protected by law. This means, for instance, that whistleblowers should not be treated unfairly because they 'blow the whistle'. There are however differences across countries. To give an example, in Europe whistleblowing protection is extended to self-employed individuals, unpaid trainees/volunteers, contractors, shareholders, board members (including non-executives) and facilitators (i.e., individuals and legal entities connected to the whistleblower), whereas in the UK whistleblowers are protected only if they are categorised as former employees and "workers" (including agency workers, freelance workers, seconded workers, homeworkers and trainees, non-executive directors, and LLP members), and overseas employees.

To make a whistleblowing report effective and impactful, it is important to:

Collect evidence: the report should include verifiable details such as specific dates, times, names, and documentation that support the reported misconduct. When reporting wrongdoing, include answers to the following questions:

- What happened?
- When did it occur?
- Where did it take place?
- Who was involved?
- Were there any witnesses?
- Was management informed?



Review the regulatory framework: key legal frameworks to consider are listed below:

- EU Whistleblowing Directive: this framework requires organisations across EU member states to establish secure channels for whistleblowers and protect their identity.
- EU GDPR: this framework protects whistleblowers' personal data by requiring organisations to secure sensitive information.
- UK Public Interest Disclosure Act (PIDA): this framework protects whistleblowers who report wrongdoing in the public interest (for instance safety concerns or fraud).
- Country-specific frameworks: some EU countries such as Germany and Austria have specific whistleblowing laws. This website can provide more guidance: <https://whistleblowingmonitor.eu/>.

Understand the reporting mechanism: familiarise yourself with the reporting systems and find answers to the following questions:

- Who to tell?
- How to report anonymously, if you wish to?
- How is whistleblowers' sensitive information protected?
- What will happen once the wrongdoing is reported?

For resources including anonymous expert assessment and connections to legal counsel, the AI Whistleblower Initiative (AIWI) provides support at <https://aiwi.org/>.



ANNEX 5: ENVIRONMENTAL IMPACT INDICATORS

Below are some indicators recommended by the OECD²² that you may want to review:

Production	• GHG emissions from production, in metric tons of CO₂e • The carbon intensity of production methods, in metric tons of CO₂e per unit (e.g. per person or dollar of revenue) • The share of renewable energy used in production • The share of recycled or renewable materials used in production
Transport	• GHG emissions from production, in metric tons of CO₂e • The carbon intensity of production methods, in metric tons of CO₂e per unit (e.g. per person or dollar of revenue) • The share of renewable energy used in production • The share of recycled or renewable materials used in production

22 OECD. (2022). *Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications: The AI Footprint*. OECD Digital Economy Papers, No. 341, OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2022/11/measuring-the-environmental-impacts-of-artificial-intelligence-compute-and-applications_3ddded5/7babf571-en.pdf



Operations	• Electricity consumption in TWh • Renewable electricity consumption in TWh • Power Usage Effectiveness (PUE) for total facility power compared to ICT equipment power • GHG emissions in metric tons of CO₂e • Carbon intensity in metric tons of CO₂e per unit (e.g. per person or dollar of revenue) • Carbon usage effectiveness (CUE): the ratio of total CO₂e emissions caused by total data centre energy consumption to the energy consumption of ICT equipment • Water withdrawal in cubic meters • Water consumption in cubic meters • Water discharge (quality) in cubic meters • Water usage effectiveness (WUE) in litres per kilowatt-hour • Water withdrawal intensity in cubic meters per unit (e.g. per person or dollar of revenue)
End-of-life	• Electronic waste in metric tons • Recycling rate • Electronics disposal efficiency (EDE) percentage • Percentage of electronic waste sent to landfills



ANNEX 6: RESOURCE LIST

1. AI Policy Trackers

- FAIRLY Global AI Regulation Tracker (<https://www.fairly.ai/blog/map-of-global-ai-regulations?>)
- White & Case AI Watch: Global Regulatory Tracker (<https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker>)
- AIPP Global AI Law and Policy Tracker (<https://iapp.org/resources/article/us-state-ai-governance-legislation-tracker/>)
- AIPEX – AI Policy Exchange Forum (<https://aipolicylab.se/aipex/>)
- US AI Legislation Tracker (<https://www.americanactionforum.org/list-of-proposed-ai-bills-table/>)
- Tracking Federal Judge Orders On Artificial Intelligence (<https://www.law360.com/pulse/ai-tracker>)
- OECD Policy Tracker (<https://oecd.ai/en/>)
- AI Policy Portal (<https://aipolicyportal.org/>)
- African AI Policy Observatory (<https://policy.africanobservatory.ai/>)
- Tech Policy Atlas (repository) (<https://techpolicydesign.au/tech-policy-atlas>)
- Tech Policy AI Act Tracker (<https://www.techpolicy.press/tracker/ai-act/>: <https://www.techpolicy.press/tracker/>)
- UK AI Tracker LexisNexis <https://www.lexisnexis.co.uk/legal/guidance/uk-artificial-intelligence-tracker>
- Global AI Regulation Tracker (<https://www.techieray.com/GlobalAIRegulationTracker>)
- AIDV AI and Democratic Values Index (<https://www.caidp.org/reports/aidv-2023/>)
- Human Rights, Democracy, and Rule of Law Impact Assessment (HUDERIA): (<https://rm.coe.int/cai-bu-2022-03-outline-of-huderia-risk-and-impact-assessment-methodolo/1680a81e14>)



2. AI Incident Trackers

- Awful AI Tracker (<https://github.com/daviddao/awful-ai?tab=readme-ov-file>)
- AI Incidents Database (<https://incidentdatabase.ai/>)
- Partnership on AI (<https://unethical-designs.framer.ai/>)
- AI Risk Repository (MIT Future Tech (<https://airisk.mit.edu/>))
- OECD AI Incidents Monitor (<https://oecd.ai/en/incidents>)
- AI Risk Atlas (<https://www.ibm.com/docs/en/watsonx/w-and-w/2.0.x?topic=ai-risk-atlas>)
- General: AIAAIC Repository – AI, Algorithmic, and Automation Incidents, Systems and Datasets: <https://www.aiaaic.org/aiaaic-repository>
- Peter Slattery, Alexander K. Saeri et al. “The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence” <https://arxiv.org/abs/2408.12622>
- Risto Uuk, Carlos Gutierrez et al., “A Taxonomy of Systemic Risks from General-Purpose AI “ ([here](#))
- Risto Uuk, Annemieke Brouwer, et al. “Effective Mitigations for Systemic Risks from General-Purpose AI” ([here](#))



ANNEX 7: POLICY LANDSCAPE



● The policies in Data protection and Privacy, e.g., the GDPR, are connected to the AI Act by the fact that data protection provides an overarching personal data protection regime, safeguarding EU citizens' privacy in communication, law enforcement and beyond. However, AI is working on data, meaning that some form of data sharing and access infrastructure is unavoidable, when reaching for a digital EU, as the Commission is envisioning.

● To grant data access in accordance with data privacy laws, policies such as the Data Act, the Data Governance Act or the Open Data Directive, are to ensure fair access in B2B, B2C and B2G, regulate the use and re-use of protected public sector data and data altruism, as well as highlight high-value data sets. Together, privacy protection, fair access and (re-)use are to ensure that the AI Act's requirements on data quality and risk management are met with incentives of innovation and cross-border AI development.

● Another set of policies interacting with the AI Act fall under platform and market regulation such as the Digital Single Market and the Digital Services Act. They directly target the users and re-user of the above-mentioned data and their responsibilities to diminish systemic risks and ensure accountability as well as their power as gatekeepers of data. While the DSA ensures that AI-powered functions on digital platforms are transparent and accountable, the DMA curbs structural imbalances by mandating fair competition and data access.



● The security framework mainly consists of the Cyber Resilience Act (CRA), High common level of cybersecurity directive (NIS2 Directive), Digital Operations Resilience Act (DORA) and the Cybersecurity Act. While the CRA provides a horizontal layer of cybersecurity requirements of digital products with digital functionality, NIS2 is a baseline cybersecurity policy for operators to prevent, detect and report incidents. A cybersecurity certification framework is provided by the Cybersecurity Act under ENISA, and DORA obliges the financial sector to manage risks in ICT, report incidents and test resilience. The AI Act ensures that AI systems must be trustworthy, and the security framework provides the necessary guidelines in the digital infrastructure and safety and resilience of the products running essential systems.

● The trust framework in the digital space, targeting digital identities, is represented by the EU's flagship program of the EU Digital Identity Wallet (EUID). This wallet allows individuals and businesses to prove their identity, attributes and entitlements digitally. While the AI Act defines obligations for AI systems to be transparent, traceable and accountable eIDAS2 provides the technical and legal infrastructure to make these obligations operational. The EUID Wallet enable trust services and interoperability across borders, ensuring that AI systems can operate securely with verified identities.

● Additionally, some sectoral product safety, such as the Machinery Regulation, General Product Safety Regulation or the European Health Data Space, are already in place providing sectoral-specific protections added to the horizontal AI governance. These instruments ensure that the AI systems deployed are embedded in sensitive-sectors-products are technologically compliant and physically safe.

● Lastly, the digital sphere touches intellectual property and content, where the Digital Single Market Copyright Directive grants lawful access to use and re-use protected works in AI training and operation. The Directive provides the legal foundation for lawful text and data mining, while the AI Act imposes transparency and documentation duties on training datasets. This ensures that AI systems can access and reuse creative works, driving AI innovation without undermining intellectual property rights and the interests of authors and creators.

To conclude, the AI Act is part of a bigger digital policy ecosystem in the EU, that, if you are working with AI, needs to be understood. In terms of ethics in AI, especially policies regarding data privacy, intellectual property and markets can be looked closer at, as these directly influence the ethical topics raised in this playbook.

For further information consider contacting your country's AI Office.



GAME 1: OVERSIGHT IN ACTION

Oversight Approach Cards (3 Cards)

Human-on-the-Loop

Human participates in every decision cycle of the AI system. AI supports, but final action requires human confirmation or input.

Human-on-the-Loop

Human continuously monitors the AI system and can intervene if needed. AI executes decisions autonomously unless human overrides.

Human-in-Command

Human makes strategic decisions about when and how AI is deployed. AI may operate autonomously day-to-day, but humans control deployment and responsibilities.

Scenario Cards (15 Cards)

Medical Diagnosis Assistant

AI identifies illnesses from X-rays and proposes treatments for doctors to evaluate.

Disaster Response Route Planner

AI optimizes evacuation routes and logistics during wildfires or earthquakes

AI Sentencing Recommendation System

Judicial AI suggests criminal sentences based on prior case data and risk assessments.

Military Reconnaissance Robot

AI scouts dangerous zones and flags threats in real time for strategic review.

Hiring & Candidate Screening Tool

AI filters job applications and ranks candidates based on suitability scores.

Stock Trading Bot

AI executes trades at high speed based on market predictions and asset risk.



Smart Farming Automation

AI controls irrigation and nutrient adjustments across large agricultural fields.

Credit Risk Assessment for Loans

AI evaluates applications and approves or rejects consumer lending.

Content Moderation AI

Automated system removes harmful or illegal posts on a fast-moving social platform.

Emergency Dispatch Triage System

AI prioritizes incoming 911/112 calls, categorizing urgency.

Autonomous Chemical Plant Safety Monitor

AI adjusts controls and shuts down systems if it detects risk of explosion.

School Performance Monitoring AI

AI identifies students needing intervention and suggests teaching adjustments.

Elder-Care Companion Robot

AI assists seniors with medication management and fall detection.

Border Security Facial Recognition AI

AI screens travelers for identity or security concerns at checkpoints.



GAME 2: WHAT TYPE OF BIAS?

Real-world Fairness Incident Cards:

Real-world incident card

AI systems employed by local councils in England may underestimate women's health concerns during social care assessments. In a study, researchers tested these tools with real case notes, altering only the patient's gender. The findings showed that men's cases were frequently labelled with serious descriptors such as 'disabled' or 'complex,' whereas women's cases were summarized in less severe language.

Real-world incident card

In the UK Biobank, most of the participants are healthy,, consequently the biobank does not provide an accurate representation of the patients typically receiving healthcare services.

Real-world incident card

Large language models frequently misinterpret everyday expressions and stylistic features characteristic of African American online communication, owing to a mismatch between the datasets used for training and the lived practices of the language

Real-world incident card

An algorithmic risk assessment computed the prediction of a person's recidivism risk by using prior arrests and arrests of friends and family as proxy variables. The use of such mismeasured proxies caused higher rates of false positive for African American versus white citizen defendants. For instance, an analysis in Broward County showed that black non-recidivist defendants were almost twice as likely to be classified as higher risk compared to their white counterparts.

Real-world incident card

Some clinical AI systems designed for the early detection of melanoma have been trained primarily on images from individuals with lighter skin tones. As a result, their diagnostic accuracy may be lower when applied to patients with darker or more diverse skin types.

Real-world incident card

A recent study estimated that in ChatGPT-3's training data, just about 26.5% of the content was authored by women.

Real-world incident card

Commercial facial recognition technologies show markedly higher error rates when identifying women and individuals with darker skin tones compared with white men. These disparities stem from limited diversity in the training datasets, as well as biases in data collection and labelling practices. When deployed by law enforcement, such shortcomings have resulted in wrongful arrests and disproportionate policing of minority groups.

Real-world incident card

AI researchers often compare the performance of large language models to "humans". However, they often refer to a narrow subset of "humans", also known as WEIRD (Western, Educated, Industrialized, Rich, and Democratic) groups. As a consequence, there is a risk that LLM performance might decline when moving away from WEIRD populations.

Real-world incident card

Since 2013, the Dutch Tax and Customs Administration has used an automated risk classification model to identify possible irregularities in applications for childcare benefits application. This tool relied on characteristics correlated with immigrant and minority backgrounds, such as dual nationality and foreign-sounding names.

Real-world incident card

Credit cards have been found to assign lower spending limits to women than to men, despite comparable financial backgrounds. An investigation into Apple Card data suggested that the algorithm may have relied on factors that indirectly prioritised financial behaviours more typical of men.



Bias Cards:

Measurement Bias

This type of bias arises from the way certain features are defined, captured, or applied in a system. It often occurs when inaccurate or imperfect proxy variables are used in place of the intended measure.

Feature Selection Bias

This type of bias arises when the features included in the training dataset are not adequate or equally relevant across all subgroups.

Selection Bias

This type of bias occurs when choices about which data are gathered or are included unintentionally advantage particular groups or traits.

Sampling Bias

This type of bias arises when the data are collected through a non-random process, leading to an unrepresentative sample.

Confirmation Bias

This type of bias occurs when data are selected, interpreted, or prioritised in a way that reinforces existing beliefs or expectations.

Representation Bias

This type of bias arises from the lack of diversity in training datasets.



GAME 3: THE ETHICAL RISKS OF ACCURACY, ROBUSTNESS & CYBERSECURITY

Scenario 1

A financial institution uses an AI-based fraud detection system to monitor credit card transactions and identify usual patterns that might indicate fraudulent activity. The system analyses thousands of transactions per second, learning from historical data to recognize patterns of legitimate purchases and potential fraud.

Scenario 2

A small business uses an AI-powered smart calendar system to manage meeting schedules for employees. The AI system analyses available timeslots, employee preferences, and meeting priorities to automatically schedule appointments, optimise work hours, and minimise scheduling conflicts.

Scenario 3

You belong to a minority ethnic group and live in a low-income neighbourhood. You have to make some essential but larger-than-usual purchases for your family, including a laptop for your child's education and a washing machine. These transactions, which are outside your typical spending pattern, are flagged by the bank's AI-based fraud detection system as potentially fraudulent. A few hours later, you receive a call from the bank's fraud department, informing you that your purchases have been temporarily blocked for verification. After explaining that the purchases are legitimate, you go through a complaint process and discover that the AI system didn't recognise your spending behaviour as typical, primarily because it was trained on data from higher-income and predominantly white neighbourhoods.



GAME 4: EXPLAINABILITY CROSSWORD

EXPLAINABLE

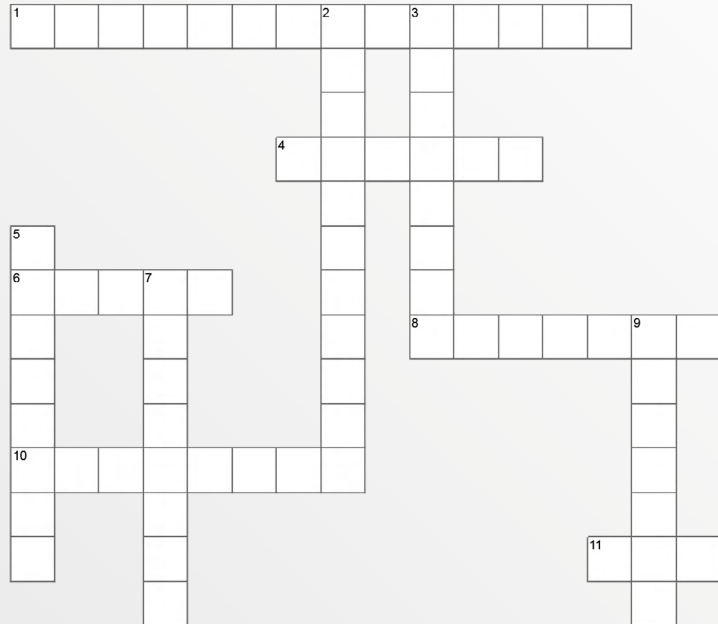
ACROSS

- 1 Users can become over-reliant on AI outputs, even when explanations are flawed. Systems must clarify both their strengths and the limits of their reasoning.
- 4 Global frameworks increasingly reference explainability and transparency in AI. Laws, norms, markets, and technical design together shape responsible AI behaviour.
- 6 Explainability should begin with local contexts, jurisdictions, and applications. Small-scale, situated efforts can build trust and compliance step by step.
- 8 The goal of explainability is to give people real power over automated decisions. Rights like intervention, contestation, and explanation are at the heart of this idea.
- 10 Ethical and legal values should be embedded directly into AI systems. Frameworks like GDPR promote collaborative, holistic approaches to responsible AI.

- 11 Courts in Europe have stressed the right to an explanation in automated decision-making. Providers cannot hide behind trade secrets when fundamental rights are at stake.

DOWN

- 2 People have a right to know how their data influences automated decisions. Details like logic, human involvement, and decision boundaries must be made clear.
- 3 Law, policy, and technical practice must cross-pollinate for explainability to work. Both policymakers and engineers need to understand each other's limits and tools.
- 5 Opaque models often reduce trust and acceptance from users. Without explainability, organisations risk losing adoption and competitiveness.
- 7 Explanations must be tailored for different users, not just technical experts. Few XAI studies test with real people, yet communication is key for impact.
- 9 In high-risk AI, making decisions understandable is seen as the only responsible path. Values like transparency, accountability, and trust are tightly linked to this principle.



GAME 5: POLICY TRACKER CHALLENGES

1. Global AI Regulation Tracker

- Open the following webpage: <https://www.techieray.com/GlobalAIRegulationTracker>
- Challenge: Compare 2 countries policy approaches e.g. UK and China? What is surprising?

2. OECD Dashboard

- Open the following webpage: <https://oecd.ai/en/dashboards/policy-initiatives?orderBy=startYearDesc&page=1>
- Challenge: Look into the UK, what shapes its AI policy? What do you think you should be aware of? e.g. National AI strategy, AI governance bodies and mechanisms, AI policy initiatives?

3. Diversifair Tool

- Open the following webpage: <https://diversifair-project.eu/visualisation-tool/>
- Challenge: Pick one of the keywords of the DIVERSIFS AIR framework e.g. discrimination/bias - which is most surprising in terms of policy activity across countries?

4. Fairly AI Map

- Open the following webpage: <https://www.fairly.ai/blog/map-of-global-ai-regulations>
- Challenge: Try to look into Generative AI legislation - and where they are in the world - what are some initial observations you'd like to draw?



