



High-risk EU AIAct Toolkit



UNIVERSITY OF
CAMBRIDGE



LEVERHULME CENTRE FOR THE
FUTURE OF INTELLIGENCE





Contents

<u>About</u>	00
<u>Team</u>	02
<u>Acknowledgments</u>	03
<u>European Artificial Intelligence Act</u>	04
<u>High risk EU AIAct Toolkit</u>	10
About space	11
Space 1: Business, team and principles space	12
Space 2: Impact, risks and mitigation space	34
Space 3: Stakeholder engagement space	48
Space 4: User-centered design space	60
Space 5: Data governance space	76
Space 6: Model governance space	96
Space 7: Monitoring, evaluation and care space	106



About

Welcome to the **High risk EU AiAct Toolkit - HEAT**, a toolkit created by the **Leverhulme Centre for the Future of Intelligence** at the University of Cambridge and the AI service provider **Ammagamma**.

HEAT helps providers of high-risk AI comply with the EU AI Act. It's primarily directed towards **product managers** or those with an oversight role over an AI project. We guide you through the mandatory documentation processes, from risk management and data governance to the Declaration of Conformity. Ultimately, we also hope to make it easier for citizens to understand what high-risk AI is and how it affects their rights.

By completing HEAT, you're not just making compliance easy, but

also upskilling yourself with the latest AI ethics technique. You're concretizing project aims, evaluating product use cases, and discovering what the EU's fundamental rights and values really are. You can show customers and the rest of your business that you believe ethics is about more than just ticking boxes.

It's about bringing the EU AI Act to life through your business' values. HEAT turns the requirements of the EU AI Act into simple step-by-step instructions. Completed documents can be downloaded from the tool.

We also include workbook sections where you can demonstrate your compliance with the EUAI Act and prepare for future audits.

Our values

1. Pro-Justice approach to AI Act compliance

Our toolkit goes beyond compliance, focusing on fulfilling the AI Act's intent by addressing often-overlooked areas such as environmental impact, accessibility, consent and redress. This pro-justice approach ensures AI aligns with both the spirit of the Act and societal well-being.

2. Responsible AI as a continuous commitment

Our toolkit supports continuous, iterative improvements in AI, helping teams stay responsive to emerging ethical challenges and evolving user needs. Responsible AI isn't a one-time goal but a sustained commitment to ethical standards.

3. Collaborative effort across boundaries

Our toolkit encourages collaboration and engagement across teams, stakeholders and communities. By engaging internal teams, external experts and affected communities, we foster an inclusive approach that respects and serves society, making ethical AI a shared responsibility.



Team

Leverhulme Centre for the Future of Intelligence

Eleanor Drage
Senior Research Fellow

Tomasz Hollanek
Research Fellow

Dorian Peters
Senior Research Fellow

Yulu Pi
Research Assistant

Ammagamma, part of Accenture

Cosimo Fiorini
Ind & Func AI Decision Science Manager

Virginia Vignali
PhD Student, University of Bologna

Acknowledgments

We would like to thank the **Advisory Board members** for their contributions, feedback and knowledge shared throughout design process:

Caroline Bassett

Cynthia Bennett

Toju Duke

Connor Dunlop

Federica Frabetti

Kevin Guyan

Os Keyes

Merve Hickok

Michael Madaio

Kerry McInerney

Bogdana Rakova

Saurabh Tiwari

Kshitiz Verma

Sharon Webb

Thanks to *BT, Mainly.ai* and *students of the MSt in AI Ethics and Society from the University of Cambridge* for their help with user testing.

The project has been developed with the *Accenture Center of Excellence in AI and generative AI* that will make the toolkit freely available.

This project was made possible by funding from *Stiftung Mercator* as part of the *Desirable Digitalisation project*.



FIRST REGULATION ON ARTIFICIAL INTELLIGENCE

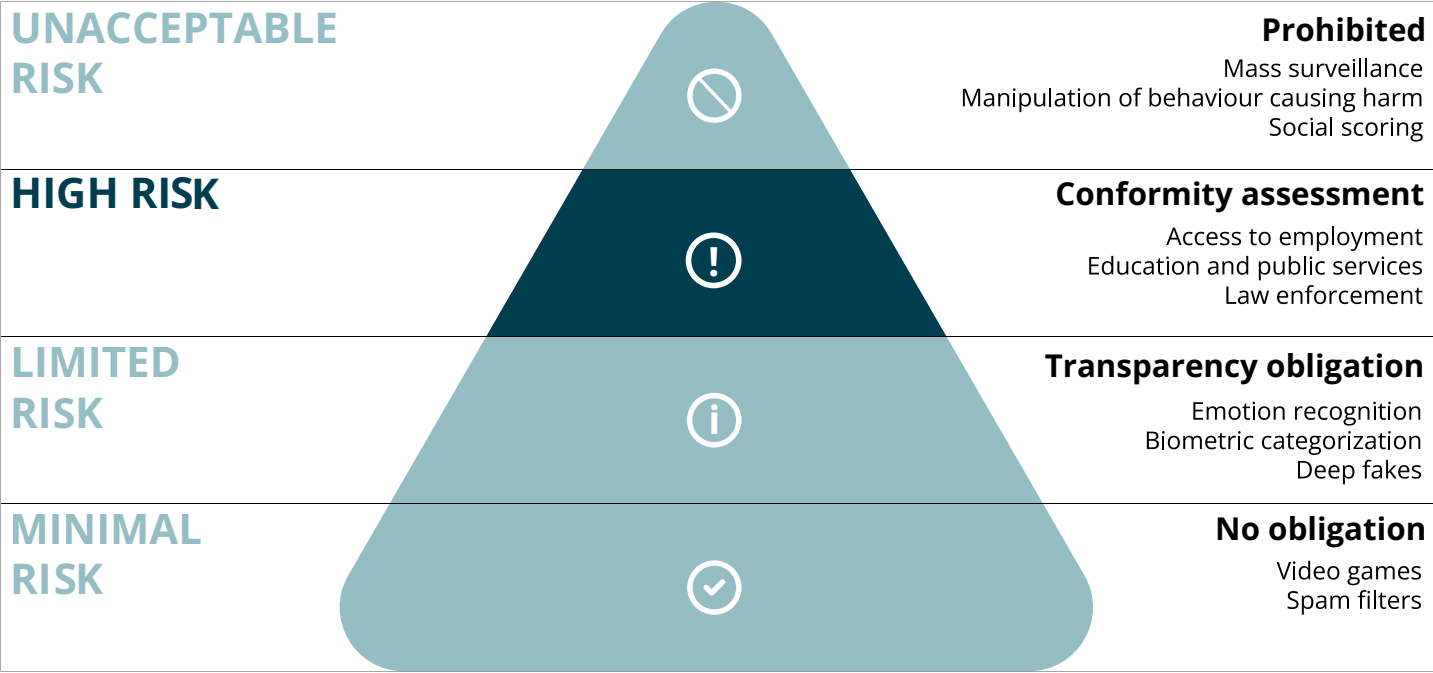
European Artificial Intelligence Act

HOW DOES THE AI ACT REGULATE AI?

The EU AI Act, the world's first comprehensive AI law, categorizes AI systems into different risk tiers based on their potential risks and level of impact, each with its own set of rules to ensure safety and responsibility.

Eight specific AI practices will be strictly prohibited, including the use of emotion recognition in workplace and educational settings, social scoring systems, predictive policing and AI designed to manipulate human behaviour or exploit vulnerabilities. While most AI encountered in our daily lives, like recommender systems and spam filters, will get a free pass, high-risk AI will face stringent rules to mitigate risks.

Our toolkit is designed to help high-risk AI providers comply with the EU AI Act.



Key definitions

AI system

Machine-based system that is designed to operate with varying levels of autonomy, which may exhibit adaptiveness after deployment; and for explicit or implicit objectives, infers from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments.

AI provider

Natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model; has an AI system or a general-purpose AI model developed and places it on the market; puts the AI system into service under its own name or trademark, whether for payment or free of charge.

AI deployer

Natural or legal person, public authority, agency or other body using an AI system under its authority; except where the AI system is used a part of a personal non-professional activity.

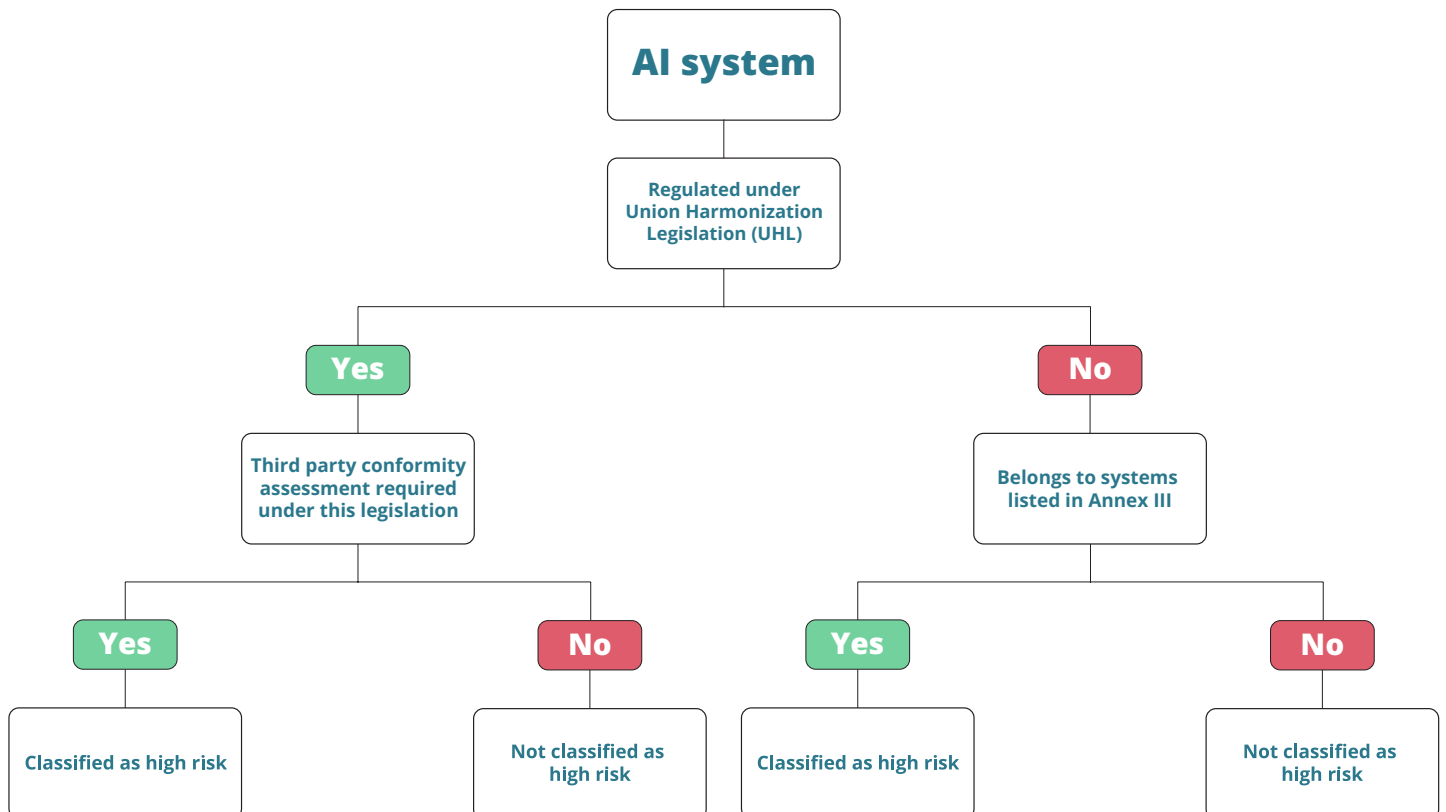
Does my AI fall into the high-risk category?

The Act delineates high-risk AI Systems as those possessing substantial potential to inflict harm, impact safety, or impinge upon fundamental rights.

High-risk AI systems fall into two main categories:

- [AI systems listed in Annex III of the EU AI Act.](#)
- AI systems categorized as products or safety components under specific EU product safety regulations (Union harmonization legislation) listed in [Annex I of the EU AI Act.](#)

You can also check if your AI falls into this category by using a [free online tool](#).



ANNEX III

High-risk AI system

Biometrics

- Remote biometric identification system
- Biometric categorization
- Emotion recognition AI system

Critical infrastructure

- Safety components for critical infrastructure operation (e.g. road traffic, digital infrastructure, water, gas, heating, electricity, etc.)

Education and vocational training

- Determining access/admissions
- Evaluating learning outcomes
- Monitoring prohibited behaviour

Employment and workers management

- Recruitment, hiring and job ads
- Promotion, termination and salary
- Task allocation and performance

Essential service and benefits

- Public benefits eligibility
- Life and health insurance pricing
- Emergency services response
- Credit scoring

Law enforcement AI usage

- Victim risk assessment prediction
- Polygraph tests
- Evaluating reliability of evidence

Migration, asylum and border control

- Asylum and visa application
- Identifying individuals
- Assessing security risks
- Polygraph tests
- Evaluating reliability of evidence

Judicial and democratic processes

- Researching and interpreting facts and applying the law
- Influencing outcome of elections and referendums

Products or product safety components

Machinery

- Directive 2006/42/EC

Toys and toy safety

- Directive 2009/48/EC

Medical devices

- Regulation (EU) 2017/745
- Regulation (EU) 2017/746

Vehicles

- Regulation (EU) 2019/2144
- Regulation (EU) No 168/2013

Aviation

- Regulation (EU) 2018/1139
- Regulation (EU) 300/2008

Marine Equipment

- Directive 2014/90/EU

Rail system

- Directive (EU) 2016/797

Appliances burning gaseous fuel

- Regulation (EU) 2016/426

Personal protective equipment

- Regulation (EU) 2016/425

Cableway installations

- Regulation (EU) 2016/424

Pressure equipment

- Directive 2014/68/EU

Personal watercraft

- Directive 2006/42/EC

Lifts

- Directive 2014/33/EU

Radio equipment

- Directive 2014/53/EU

What are the obligations of high-risk AI providers?

1. Ensure your high-risk AI systems are compliant with the requirements set out in Article 8-15, including keeping the logs which are automatically generated by your high-risk AI systems.
2. Implement a quality management system (Article 17).
3. Maintain documentation (Article 18).
4. Conduct the conformity assessment before market placement (Article 43). Conformity assessment procedure must include:
 - Verifying that their quality management system complies with Article 17.
 - Reviewing the technical documentation to ensure the AI system meets the essential requirements in Chapter III, Section 2.
 - Ensuring the design, development process, and post-market monitoring plan of the AI system align with the technical documentation.
5. Drawing up an EU Declaration of Conformity (Article 47).
6. Affixing the CE marking to the system or its packaging/documentation, indicating regulation compliance (Article 48).
7. Registering in the EU database (Article 49).
8. Taking Corrective Actions and a Duty of Information when appropriate after placing your AI system in EU market (Article 20).
9. Demonstrating system conformity upon request by a national authority.
10. Ensuring compliance with accessibility requirements according to Directives (EU) 2016/2102 and (EU) 2019/882.



TOOLKIT

High risk EU AiAct Toolkit

WHAT CAN YOU EXPECT FROM HEAT?

- A pro-justice guide to understanding and complying with EU AI Act obligations for high-risk AI providers.
- An invitation to challenge unspoken assumptions: is AI really the best solution for your problem?
- Practical methods and tools to support your compliance journey.
- A flexible, non-linear exploration: select and move between different spaces in the toolkit.

Remember: Completing this toolkit doesn't guarantee full and indefinite compliance. We believe AI ethics and compliance is an ongoing, iterative process that requires continuous engagement and reflection.

[Try and use](#)
[High risk EU AiAct Toolkit](#)

ABOUT SPACE

HEAT is organised into 'spaces'. There is no requirement to follow a linear progression to complete these spaces; in fact, you should frequently revisit and reevaluate spaces as contexts change and new concerns emerge.

In each task, we provide guidance and action points, including the 'Description' [of the task], 'How to Approach' [the task], 'Expertise and Engagement' [the people with which to engage], 'Go Further' [optional and deeper details] and 'Workbook' [to enter your input].



SPACE 1 BUSINESS, TEAM AND PRINCIPLES



SPACE 2 IMPACT, RISKS, MITIGATION SPACE



SPACE 3 STAKEHOLDER ENGAGEMENT SPACE



SPACE 4 USER-CENTERED DESIGN SPACE



SPACE 5 DATA GOVERNANCE SPACE



SPACE 6 MODEL GOVERNANCE SPACE



SPACE 7 MONITORING SPACE



SPACE 1

Business, team and principles space



OVERVIEW

This Space is where the HEAT journey begins. You'll define project aims, evaluate product use cases, justify AI implementation, reflect on guiding values, and discover the EU's Fundamental Rights and Values which are central to the AI Act.

The EU AI Act defines [high-risk AI](#) as systems with the potential to cause significant harm to EU citizens or the environment. Examples include AI used in critical infrastructure, transportation, and healthcare, and decision-making processes with legal implications, such as credit scoring or hiring decisions (see the full list in [Annex I](#) and [Annex III](#)). You can also check if your AI falls into this category by using a [free online tool](#).

Complying with the AI Act can be a daunting task for providers of high-risk AI, as it requires extensive documentation: including a risk management system, data governance and declaration of conformity. Citizens also lack an understanding of what

high-risk AI tools are and what their rights are under EU regulation.

HEAT streamlines mandatory EU internal control and documentation while providing crucial guidance on ethical and social considerations related to their products. This tool goes beyond mere legal compliance by prompting product managers to think critically about how AI relates to structural inequality and engage meaningfully with the ethos of the Act.

HEAT integrates and reorganizes the requirements of the AI Act into step-by-step instructions. This means that completed documents can be downloaded from the tool. In this way, users can demonstrate what steps they have taken to ensure they do more than just conform to the legal minimum required by the Act.



1.1 ORGANIZATION

Task: Provide organisation title and contact details

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Annex VIII: Information to be Submitted upon the Registration of High-Risk AI Systems \(Article 49\)](#)

RATIONALE

You need to register your company and your high-risk AI system in the EU database

WORKBOOK | Enter your input

Organisation Title

Address

Contact Phone Number

Email

Other (the size of your organizatio, etc)

1.2 ACCOUNTABILITY

Task: Document responsibilities across your team

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 17: Quality Management System](#)

RATIONALE

Making accountability explicit ensures that everyone knows what they are responsible for, and nothing falls through the cracks.

DESCRIPTION

Define the responsibilities of each team member. Specifically, who will be accountable for the different aspects of the project.

HOW TO APPROACH

Please assign one or more team members to cover the responsibilities in the workbook section, noting that some duties may overlap.

- Consider how you will clarify roles and address any areas of overlap. While it is crucial to assign names to these roles, this alone is not enough to ensure responsibility is fully integrated into your process. Clarify how responsibility is distributed throughout the team, with each member understanding exactly where and how they should act.

GO FURTHER

Reflection point: How is work valued and distributed?

The value given to certain tasks or jobs is often entangled with race, gender, and disability. For example, work considered 'high-value', such as management or system development, is disproportionately assigned to men; while tasks considered 'low-value', such as data cleaning, is disproportionately delegated to women or other minoritised groups. Although 80% of the work involved in data analysis is spent on the process of cleaning and preparing the data, a 2014 *New York Times* article equated the task of cleaning data to the low-wage maintenance work of 'janitors' or cleaners, thus associating it with low-value work. Consider how work is valued and distributed across your team and aim to ensure low-value and high-value work are distributed fairly.

We encourage you to ask these questions from [Data Feminism](#), including:



- Who is doing the work of data science/AI (and who is not)?
- Whose goals are prioritized in data science/AI (and whose are not)?
- Who benefits from data science/AI (and who is either overlooked or actively harmed)?

Please note that some roles may have overlapping responsibilities. You can assign multiple staff members to one role, or one staff member to multiple roles.

WORKBOOK | Enter your input

AI system owner

Drives accountability for all roles and ensures the effective execution of responsibilities; Oversees product design and deployment to achieve intended results and takes responsibility for any harms or malfunctions.

Enter name, job title and contact

Data Protection Officer

Ensures compliance with GDPR in all data handling and processing activities relating to product design.

Enter name, job title and contact

AI Robustness Checker

Handles the technical aspects of evaluating robustness of your AI system.

Enter name, job title and contact

AI Bias Mitigator

Focuses on identifying and mitigating bias in your AI system. It can be someone or a group of people from the data, legal or ethical teams. Make sure this person has proper knowledge of [European non-discrimination law](#).

Enter name, job title and contact

AI Transparency Officer

Creates transparency reports to enhance the transparency of the design process and AI system for users and stakeholders.

Enter name, job title and contact

Complaint Oversight Officer

Ensures effective mechanisms are in place for users and stakeholders to issue a complaint and have that complaint responded to and acted on.

Enter name, job title and contact

Stakeholder Liaison Officer

Ensures the active involvement of stakeholders and citizens in the system design and development process.

Enter name, job title and contact

1.3 DIVERSITY

Task: Plan for diversity across your team

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 69: Codes of Conduct for Voluntary Application of Specific Requirements](#) - Although the Act does not mandate team diversity, Article 69 states that the Commission and the Member States are entrusted with encouraging codes of conduct specifically aimed at promoting the voluntary adoption of requirements related to diversity within development teams in all AI systems.

RATIONALE

Lack of diversity leads to ill-fitting solutions, unanticipated harms, and misrepresentation of users and their needs.

DESCRIPTION

Reflect on diversity across your team. The goal of this task is to design teams that include, encourage and actively engage with diverse perspectives.

What is meant by diversity?

Diversity means changing men-only, abled-bodied only, white-only, English-speaking only, and privileged-only organisations into ones led by people from all of society. Often people are also doubly discriminated against because they are both, for example, female and disabled. **Understanding how people can be treated badly because of more than one criterion is called 'intersectional feminism'.** This is vital knowledge in AI, where homogenous teams of people who have not experienced discrimination because of race, gender, disability or class (even if they are from different countries) end up producing tools that can be harmful to or not work for those who are different from them.

Tokenism

Having a token women or person of colour on the team doesn't solve the problem, because:

- No one wants to feel like the token diversity hire
- One person does not represent an entire demographic
- One person's race or gender doesn't make a work culture anti-racist or feminist
- a person may be considered diverse by an organisation - but may be privileged in terms of class or have previously lived in somewhere they were not discriminated against, and therefore has no knowledge of discrimination. It's not their job to represent diversity; their job is simply to do their job.

Pushback

Diversity measures are always imperfect and can leave people feeling isolated. The trick is to help employees understand where they sit in the growing tapestry of a diverse organisation. People often feel their jobs are at risk because of reverse discrimination. To get buy-in from all employees, they need to be reassured of their importance to the organisation, made aware that tech companies are still predominantly white and male, made to feel like they understand the point of making companies more diverse, and are cognisant of how new kinds of people may also make the organisation more receptive to their own quirks and idiosyncrasies.

How diverse teams reduce AI harm: Real World Example

The United Nations estimates that 20% of the older population live alone. A big tech company sells Internet of Things devices designed to give families, caregivers or healthcare professionals information about, for example, an elderly person's changes in behaviour around the house or alert them in case of a fall. However, these devices have also been used by domestic abusers to track and monitor the activity of their partners. The big tech company in question did not put any safeguards in their products until a woman in their engineering team who had experienced domestic abuse pointed out how their products were being used. The World Health Organization (WHO) state that globally, almost one third (27%) of women aged 15-49 report that they have been subjected to some form of physical and/or sexual violence by their intimate partner. And that's just the women who report it.

HOW TO APPROACH

- With external experts, assess what gaps in diversity may exist on your team. Who is designing and developing this AI product? Does the range of disciplines, backgrounds and identities represented on your team reflect the diverse requirements of your stakeholders? You also may not be aware of what perspectives are missing, so it's crucial to bring in either internal or external DEI experts to review potential gaps.
- Consider whose perspectives might be missing based on the goals of your AI product.



Consider various ways to incorporate more diverse perspectives as needed.

How to diversify your team

- Actively seek to diversify your team through inclusive recruitment strategies. This may involve reaching out to a broader pool of candidates and ensuring a fair, unbiased hiring process.
- Leverage stakeholder engagement as a powerful tool to bring diverse insights into your project. See more in the **Stakeholder Engagement Space** ([Space 3](#)).
- Put in place a diversity and inclusiveness policy on recruitment of staff.

EXPERTISE AND ENGAGEMENT

One approach to filling in representational gaps and improving diversity is to engage with relevant advocacy groups that represent key stakeholder groups. For example, consider organisations that represent the interests of people based on age, disability, background, etc. However:

- The relationship should be mutually beneficial - don't expect to email an organisation and for them to provide this service for you.
- Equally, this is no substitute for gaining that knowledge within your organisation, i.e. it should support rather than hinder internal diversity schemes.
- Do your due diligence! Make sure the organisation you consult with is led by people who are representative of the group and does not merely act on behalf of that demographic.

GO FURTHER

Don't outsource AI data work to platforms/companies that exploit workers

Often low-paid, monotonous and repetitive data tasks, termed 'AI data work', is outsourced to generalised platforms like MTurk or more specialised companies, where workers receive lower than minimum wages, manage unpredictable income streams and the frequent non-payment of tasks, and work without the standard labour protections of an employment contract. If you can, use a BPO company with employees based in an office with a traditional hierarchically organised managerial structure, rather than geographically distributed, crowdsourced labour from digital platforms such as Amazon Mechanical Turk.

When outsourcing AI data work, ask your potential third-party suppliers:

- Are their workers classified as employees or as self-employed?
- Is the prevailing minimum wage in the employee's location applied to all workers?
- Do workers have legally binding ways to make their needs and desires heard to platform operators, through union membership, collective bargaining, and, in countries with such structures, works councils and co-determination rights?
- Are workers paid the full amount for which clients are billed, in real currency?
- Are workers ever penalized for declining to accept some offered tasks or declining to work at certain times of day?
- In the event of technical problems with task or platform, do workers ever pay the cost for

lost time or work?

- Are platform terms – including the terms for payment, work evaluation and dispute resolution – presented in a readable format that is clear and concise?
- Are worker evaluations and ratings ever based on non-payment rates and are workers given reasons for any negative ratings?
- Do they have published procedures for workers to raise concerns and do they demonstrate that they are enforced?

To learn more, read:

- [International Labour Organisation](#)
- [Hidden Workers powering AI](#)

Meritocracy is a myth

There is lots of misunderstanding about what positive action is, i.e. hiring according to diversity. People assume, for example, that positive action is the opposite of (or undermines) meritocracy. This is a myth: all hires are made on whether the recruiter deems a candidate a good fit with the organisation. Historically, able-bodied white males have aligned more with what an organisation expects an ideal employee to look like. Even if a candidate that fits this traditional demographic is good, this isn't more meritocratic than hiring a good candidate that fits the diversity brief.

The difference is that positive action means being aware that how someone performs academically, behaves, and whether they possess work-related skills is influenced by factors including the education they could afford, whether they were able to study at home, the behavioural and cultural norms they grew up with, and health outcomes. In other words, positive action just means being aware that the past wasn't meritocratic either.

Activity: 'Our stories'

This can be oral/written/shared/private: engage in a reflective activity where team members share personal stories of when they didn't feel included. This could relate to issues such as accent, race, religion, or other aspects. Additionally, participants can share instances when their experiences directly influenced the development or design of a product.

1.4 THE PROBLEM, APPROACH AND JUSTIFICATION OF AI

Task: Think about the problem you're trying to solve and domain expertise required to approach it. Consider existing best practices in your area of application and whether the use of AI is justified.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Annex IV: Technical Documentation Referred to in Article 11\(1\)](#)

RATIONALE

There is often ambiguity around the problem an AI product intends to solve, including why and how it is appropriate for solving it. Thinking this through will help you ensure that you can provide all the necessary information in the next subtask.

DESCRIPTION

The justification for choosing AI as a solution typically involves key benefits like cost reduction and efficiency increases. However, we encourage you to think beyond these immediate gains when considering the three questions below. Reflect on the broader impact of AI on both your business and our society.

HOW TO APPROACH

1. *Is AI the best solution to your problem?*

When approaching problem-solving with AI, it's crucial to first acknowledge that the optimal solution may not necessarily involve AI. This is especially important when your system is intended to replace proven non-AI solutions to specific problems.

Consider established best practices to your problem, both technical and non-technical and evaluate whether and how AI provides better solutions.

Also think about common pitfalls in applying AI to solve problems. For instance, relying on seemingly convenient but inappropriate proxies for actual outcomes (see real world example below).

The US healthcare system uses commercial algorithms to guide health decisions. Evidence shows that Black patients assigned the same level of risk by one such algorithm are sicker than white patients. As health costs were used as a proxy for health needs, the algorithm reduced

the number of Black patients identified for extra care by more than half. Less money was therefore spent on Black patients with the same level of need. The model designers identified healthcare cost as a proxy for health needs without considering the influences of legacies of inequality that also affect cost. This suggests that the choice of convenient, seemingly effective proxies for ground truths can be an important source of algorithmic bias in many contexts. Find more at [AI Fairness in Practice](#).

2. How has AI been used in your domain? What were the criticisms?

There may have been criticisms of the use of AI in your domain – in particular, if it's a high-risk area of AI application. You should be aware of these criticisms and decide how your system and approach will account for them.

Search for relevant case law, academic papers and media reports that critique the use of AI in your domain.

You can use the following datasets and repositories to facilitate your search:

- [AI, Algorithmic, and Automation Incidents and Controversies Repository](#)
- [AI Incident Database](#)
- [AI Litigation Database](#)

See below an example of applying AI in a specific domain – recruitment – and existing critiques and justifications of using AI to solve specific issues:

AI is being used in video interviews to identify personality traits in candidates. The idea is that AI can 'get rid of' race and gender by assessing candidates' keywords and micro-gestures. However, this presumes that race and gender have no bearing on how people speak and move. These attributes are always part of someone's personality. Therefore, these tools don't do what they claim. Worse, they can take away funding from tried-and-tested diversity initiatives that don't involve AI. They also don't support companies in improving candidate experience once they enter the organization, by, for example, resolving issues like childcare, access, power and pay inequalities, or experiences of discrimination in the workplace.

[Read more](#)

3. Having considered existing best practices and criticisms of the use of AI in your domain, do you feel confident that you want to proceed with your design?

Consider the expertise and engagement section below to decide how to proceed – especially if you feel uncertain about this question.

EXPERTISE AND ENGAGEMENT

It is vital that your stakeholders – future end-users and other impacted groups – can rule out AI as a solution if they do not believe it is the right way of addressing the stated problem. You can use [Space 3](#) to consider engaging your stakeholders and revising your design.



GO FURTHER

Here are a few anonymized real-world examples illustrating how companies determined that AI was not the right solution for their needs.

“We decided that the product didn’t align with our values”

We were asked to create a recommendation engine for a gambling product (scratch cards). This was against our principle of using technology to help people make more informed decisions rather than selling them something that can be harmful. This kind of game can exploit someone’s weakness - gambling - and also become an illness, so optimising these products isn’t aligned to our values.

“We pivoted from workforce assessment to bias detection”

We were asked to build a workforce assessment tool for HR analytics, but the people who proposed the project didn’t have a clear idea of what they wanted to do. We knew that evaluating the performance of people neutrally doesn’t work; you can’t train a model with company data without incorporating the biases that the company has created over time. So instead of abandoning the technology, we worked with what AI can actually do, which is highlight the biases present in workforce assessments. This tool made these biases explicable and analysed them so the company could be better acquainted with its own biases.

We could also use this tool to evaluate whether employee satisfaction corresponded with payroll schemes and other improvements. This would enable the company to better address employee dis/satisfaction.

We also pivoted from creating a churn analysis that would predict whether an employee would be successful in the company (again, inevitably biased) to instead making a system that demonstrates how biases were affecting the model. This explainability system helped employees better understand the correlation between a systems features and its outputs. This was so successful that we proposed this to other companies too.

“We reconsidered the optimal use of AI for employee feedback”

Employees often struggle to write feedback for each other. We were asked to create an HR platform that used generative AI to automate employee feedback. The idea was that employees would just need to add some keywords to the system, select negative, neutral, or positive language, and generative AI would generate feedback for them. We decided not to create this for two reasons.

First, it would mean using an algorithm to substitute a human in a process, yet the value of doing this process of giving feedback to colleagues is mainly to improve the interaction between colleagues. Second, automating this process suggests that the feedback process is too time consuming, and it would be better for employees to save their time for other things. This was actually against the company’s principle of taking time for feedback in order to spread value. This is a classic example of AI doing something that actually contradicts the

company's values. Instead, we suggested that AI be used to do a kind of value spellcheck on feedback. The system could suggest improvements based on best practice for feedback, including examples of positive behaviours, specific examples of their colleagues' actions and how the employee feels about those.



1.5 JUSTIFYING THE APPROACH AND IDENTIFYING THE INTENDED USE

Task: Summarize the effort and research conducted in 1.4 to justify the use of AI for the defined problem and identify intended use cases.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Annex IV: Technical Documentation Referred to in Article 11\(1\)](#)

RATIONALE

Failing to summarize the effort and research conducted in the previous task would lead to overlooking critical risks, non-compliance with the AI Act, and the potential deployment of an ineffective or harmful AI solution.

DESCRIPTION

In Task 1.4, we asked you to specifically consider the problem you are developing AI solutions for, identify the best technical and non-technical practices available, and examine the known risks associated with using AI to address this problem or similar ones.

You chose to continue developing an AI system for your defined problem. We want to remind you again that you should consider discontinuing the AI approach if you realize that it is not the right solution for your problem.

In this task, you will summarize the effort and research you conducted in the previous task to justify the use of AI.

WORKBOOK | Enter your input**1. Problem definition**

Start by clearly defining the problem you want to solve using AI. Consider whether the problem you are trying to solve is influenced by historical and social inequalities. Describe how AI specifically addresses and solves this problem. How are you taking measures to avoid your AI reproducing and reinforcing these inequalities? Explain how AI solves the problem, including if and why its capabilities and advantages are unique.

Template: The problem we are trying to solve is [clearly state the issue]. AI solves this problem by [detailed explanation]. We are taking measures to avoid reproducing and reinforcing inequalities by [explain measures].

2. Role of AI

Break down how you will develop AI functionality by giving a description of the functions and describing the training methods, input, and output for each function. Consider how these AI functions contribute to solving the defined problem.

3. Existing best practices

*What are the non-technical and technical best practices to solve the problem?
What is this baseline performance?*

How does AI improve current solutions?

What elements of AI improvement cannot be measured using technical metrics (e.g., broader social impact)?

What technical metrics are appropriate ways of measuring, for example, fairness, transparency, and inclusivity, and why?

4. Past incidents and case studies

What are the biggest critiques of the use of AI in the problem you are trying to solve?

What are the most known/severe incidents of using AI to solve your problem?

How does your project respond to these concerns and criticisms? (e.g., previous failed or harmful attempts to solve similar problems with AI).

How is your AI solution different?

5. Intended use cases and expected outcomes

Make a list of intended use cases and expected outcomes.

Template: We expect [insert stakeholder name] to use our AI system at [context: include when, where, how]. We expect our AI system can help them [describe expected outcome].

1.6 VALUES AND PRINCIPLES

Task: State the values or principles that will guide your project.

RATIONALE

Clear communication of the values and principles guiding your AI system's design is key to building positive relationships with customers. It helps set boundaries and expectations, ensuring alignment with customer values and ethical standards.

DESCRIPTION

Corporate AI ethics values and principles are often vague and difficult to implement. As Alex Hanna (ex-Google) has explained, “be fair” or “don’t be biased” can be meaningless in engineering contexts, where bias and fairness have (multiple) mathematical definitions. [Google’s ethical principles](#) have been criticized for being an example of corporate ethics-washing.

HOW TO APPROACH

Principles guiding your AI development should:

- Outline exactly how to go about ethical engineering. [Example](#)
- Express a company’s unique culture and goals. [Example](#)
- Avoid the use of universals like ‘do no harm’ and instead express specific incentives (e.g. we are trying to combat inequality) and politics of the organisation. [Example](#)
- Outline your approach to combating inequality through your technology. This means working out how your tool makes society less unjust for minoritised groups, as opposed to merely trying to avoid additional harm. We call this an ‘actively reparative approach’. The Distributed AI Research Institute (DAIR) do this through, for example, ‘co-constructing knowledge together with the collaborating communities and fostering collective sensemaking’. [Example](#)

GO FURTHER

Explore further on the consensus over eight fundamental principles that AI systems ought to adhere to: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values. [See here: Principled Artificial Intelligence](#)



WORKBOOK | Enter your input

Please draft your organization's ethical principles, drawing inspiration from the examples above, or provide links to existing statements of values/principles.

Please give specific examples of AI applications that adhere to and diverge from your principles and values.

If your organization has an ethics committee, please list its members and describe their roles in the design, development, and governance of your AI products.

1.7 THE EU'S FUNDAMENTAL RIGHTS

Task: Get acquainted with the EU's list of Fundamental Rights and Values

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 27: Fundamental Rights Impact Assessment for High-Risk AI Systems](#)

RATIONALE

The AI Act aims to protect our Fundamental Rights and Values (FRVs). By understanding how your AI system might impact these FRVs, you can not only comply with the Act but also help ensure it respects people's rights, avoids potential risks, and promotes the ethical use of AI technologies.

DESCRIPTION

The AI Act underscores the importance of ensuring that AI systems do not undermine, directly or indirectly, the Fundamental Rights and Values (FRVs). You will assess the impact of your system on the FRVs in Space 2. Now is a good time to get acquainted with the list of FRVs and how these relate to technology development.

HOW TO APPROACH

The FRVs in the EU include:

- The right to the protection of personal data
- The right to freedom of expression
- Equality before the law (including the prohibition of discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation).

You can explore the full list [here](#). You can also learn about EU values in a short video [here](#). Try out the quizzes to test your understanding of The EU's Fundamental Rights.

Fair trial and due process

A criminal justice system relies on AI algorithms to determine sentencing and parole decisions. The algorithms consider factors like the defendant's zip code and family background, leading



to biased outcomes against certain communities. Which human right is being violated?

1. Right to a Fair Trial
2. Right to Non-Discrimination
3. Right to Security of Person
4. Right to Freedom of Movement

Privacy and data protection

A company implements an AI-powered employee monitoring system without informing its staff. The system tracks employees' keystrokes, screen activity, and even records conversations during virtual meetings. This monitoring process is conducted without obtaining the employees' consent. Which human right is being violated in this corporate scenario?

1. Right to Privacy
2. Freedom of Speech
3. Right to Work
4. Right to Assembly

Prohibition of discrimination

In a job recruitment process, an AI system is used to analyse facial expressions and voice tone during video interviews to assess candidates' emotional intelligence. However, this system systematically assigns lower scores to female candidates regardless of their actual qualifications. Which human rights are potentially violated in this situation?

1. Right to Work
2. Right to Non-Discrimination
3. Right to Equality
4. Right to a Fair Trial

Freedom of expression

AI has the capacity to undermine individuals' freedom of expression, a fundamental aspect of engaging in public debates.

For instance, AI use for content moderation can help filter out self-harm, violence, sexual content, and hate speech online. However, research has shown that AI content moderation systems have suppressed content from disabled, queer, and fat creators, contributing to isolation and restricting opportunities for their work.

To safeguard people's freedom of expression, AI's use for content moderation should be implemented with oversight and a clear process to protect people's freedom of expression.

Reference: [Contestability For Content Moderation](#)

GO FURTHER

- Explore how the [Alan Turing Institution](#) is developing and applying inclusive and responsible AI to safeguard human rights and address humanitarian challenges.
- Read the report of [Artificial Intelligence & Human Rights: Opportunities & Risks](#) to grasp the implications of AI on human rights, acknowledging both the potential risks they pose and the opportunities they offer to enhance the fulfillment of these rights.



SPACE 2

Impact, risks and mitigation space



OVERVIEW

In this Space you'll explore and document the possible positive and negative impacts your system may have on individuals, communities, and the environment as you are legally required to establish, implement, document and maintain a [risk management system](#). You'll take special care to pre-empt any possible risks of harm. Once you identify risks, you should work towards making concrete plans for mitigating these and promoting justice through your system's design.

This space applies throughout all stages of development and deployment to ensure continuous consideration and integration of these critical points. You'll need to return to this space as the project evolves, contexts change, and new risks surface. This Space is closely linked with [Space 3](#) as proper impact assessment relies on engaging with stakeholders.

Consequence Scanning is a useful method of assessing potential negative consequences of innovation and developing mitigation mechanisms. You can use the [Consequence Scanning Kit](#) to guide your work in this Space.

Farsight: You can describe your AI project idea using prompts in the [Farsight](#) tool to help identify real-world AI incidents that might relate to your project and reflect on potential harms associated with these. However, please note that this tool CANNOT replace actual critical thinking and stakeholder engagement.

2.1 IMPACTS ON INDIVIDUALS, GROUPS AND COMMUNITIES

Task: Consider both intended and unintended consequences of your product for individuals, groups and communities.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 9: Risk Management System](#)

RATIONALE

If you don't systematically assess the potential impacts of your system, you'll be caught off guard when these crop up after release. By that point, it may be too late to prevent harms to stakeholders and your reputation.

DESCRIPTION

Evaluate and document the potential impacts of your system on both individuals and groups, with particular focus on marginalized and vulnerable populations.

HOW TO APPROACH

1. Run one or more [Consequence Scanning workshops](#) with your team and your stakeholders (see Space 3 for more on engaging stakeholders).
2. Ensure inclusion of marginalized and vulnerable groups to assess potential benefits and harms caused by use of AI (see [Space 3](#) for more on identifying vulnerable groups).
3. Based on insights collected during Consequence Scanning, as well as additional insights that may emerge from other forms of Stakeholder Engagement, complete the workbook on this task and go to [Task 2.4](#) to document how you implement mitigation strategies for each potential harm identified (the Consequence Scanning Kit includes a template for this purpose; or you can use your own).

This process should be recurrent: you will need to assess your product's impacts using Consequence Scanning workshops or other appropriate methods regularly as your project evolves, contexts change, and new risks surface.

Themes to consider

As you organise and run workshops, consider potential impacts across the following areas:

- User experience
- Quality of service
- Privacy

- Model Performance
- Security, Reliability and Robustness
- Accessibility
- Fairness and equal treatment
- Job security
- Informed decisions
- Physical and mental health and safety
- Disparate socioeconomic impacts
- Allocation of resources and opportunities (finance, education, employment, healthcare, housing, insurance and social welfare)
- Stereotyping and demeaning outputs

Also consider whether the system disproportionately benefits certain groups based on:

- Varied social backgrounds and education levels
- Different ages
- Different gender and/or sexual orientation
- Different nationalities or ethnicity
- Different political, religious and cultural backgrounds
- Physical or hidden disabilities

EXPERTISE AND ENGAGEMENT

You should draw on the expertise of your stakeholders ([Space 3](#)) with respect to potential benefits and risks, especially marginalised or vulnerable groups affected by AI system. You could also engage with domain experts who can advise on risk ([Task 3.3](#)).

GO FURTHER

Explore the example of unintended negative consequences in an AI tool used for healthcare: [Gender bias revealed in AI tools screening for liver disease](#).

You can also explore more AI-related incidents and controversies through the following datasets and repositories:

- [AI, Algorithmic, and Automation Incidents and Controversies Repository](#)
- [AI Incident Database](#)

WORKBOOK | Enter your input

Enter information about individuals, groups and communities, potential system benefits and potential system harms.

2.2 IMPACTS ON THE ENVIRONMENT

Task: Consider the impact your product has on the environment.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 95: Codes of Conduct for Voluntary Application of Specific Requirements](#); [Article 112: Evaluation and Review](#)

RATIONALE

Beyond the binding rules for high-risk AI system providers and deployers, the AI Act also delineates non-binding ethical principles, including environmental well-being. This principle encourages providers and deployers to develop and use AI in a 'sustainable and environmentally friendly manner. Monitoring the environmental impact of AI is likely to become compulsory in the coming years, so we recommend getting ahead by completing this section.

DESCRIPTION

From the energy used to train a given model to the amount of water needed to cool data centres, the environmental footprint of ML-based products is significant. To start thinking about how to limit the negative impact your AI system has on the environment, you should first explore what form this takes, and how you can measure it.

HOW TO APPROACH

- **Life Cycle Analysis.** Life Cycle Analysis (or Assessment – LCA) is a common method of assessing the impact of a given product/service/process on the environment. If your organization does not conduct such assessments, you can consult [this website](#) to learn how to successfully conduct LCA. You can read more about LCA specifically for AI systems [here](#).
- **Carbon Footprint Calculation.** To calculate the carbon footprint of your model, you can follow the [Hugging Face Method](#).

GO FURTHER

- Minimizing the negative impact of your product on the environment might require redesigning it: do consider this as a possibility.
- Consider the wider impacts of AI on the environment. For example, ChatGPT uses roughly 500 ml of fresh water for every 20 to 50 questions an end-user asks; see more [here](#) and [here](#). You can learn more about such comprehensive assessments [here](#).
- Think about the efficiency of your system. While this measure alone won't solve the problem of AI's environmental impact, improving a given system's efficiency may reduce its carbon footprint. Learn more about the so-called 'Green AI' [here](#).
- Consider the wellbeing of other entities, not only humans, in your design. The allocentric / more-than-human-centered / nature-centric approach to technology design encourages designers to think about non-human animals, and the environment more broadly, as 'stakeholders' in the design process.
- You can follow the design ideation exercise below to start thinking about how this approach could relate to your work. It can be done individually or as a team. Answers can be written or shared orally.

The Environment: an unimaginable user persona?

1. Close your eyes and think about 'the environment' or 'the natural world'.
 - What do you see?
 - Can you see a human in your vision?
 - Is 'the human' a part of 'the environment'?
 - What other living entities can you see when imagining 'the natural world'?
2. Now imagine that you're tasked with re-designing the product you're already working on with 'the environment' in mind as the 'end user'.
 - How is your view of the product changing? With 'the environment' in mind, does the product appear useless?
 - Do you see the needs of 'the environment' as distinct and different from your human end-users?
 - Do you think these needs might be conflicting? If so, how?
3. Now consider 'the environment' as an all-compassing network of interdependencies that humans are a part of.
 - Is catering to human users synonymous with catering to 'the environment'? If not, why?
 - How is the idea for your product changing when you begin to think about human 'user personas' as only some of the many 'personas' that form 'the environment'?
 - Can you think of ways in which your product might improve environmental wellbeing (not to be confused with environmental health of human users)?

The source for the ideation exercise is [here](#).



WORKBOOK | Enter your input

*Take note of the following environmental impact information:
Where the model was trained (in terms of location):*

The hardware used — e.g. GPU, TPU, or CPU, and how many:

Training type: pre-training or fine-tuning:

The estimated carbon footprint of the model (based on the Hugging Face tool):

Have you considered other aspects of the impact of your product on the environment? If so, document your thinking here:

2.3 POTENTIAL MISUSES

Task: Anticipate unintended impacts and misuse cases for your product.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 9: Risk Management System.](#)

RATIONALE

Failing to anticipate and preempt the misuse or malicious exploitation of your AI system not only puts people at risk but also damages the reputation of your product.

DESCRIPTION

Consider and document both unintentional misuses by end-users and intentional abuse by malicious actors. This process should be recurrent: You must anticipate and preempt misuse or malicious exploitation regularly as your system updates, contexts change, and new ways of using your AI system emerge.

HOW TO APPROACH

Communicate Usage Conditions

When watching for malicious intentions, explore and describe below how your product can be used for the following malicious activities:

- Cyber Crime
- Biosecurity Threats
- Politically Motivated Misuse
- Unwarranted surveillance

The following examples demonstrate how people can intentionally misuse the functions of an AI system to do harm.

- Using Facebook Ad Microtargeting for Election Interference or Targeting the Vulnerable: exploiting the extensive user data on Facebook, malicious actors can tailor political ads to specific demographics, potentially swaying election outcomes or targeting vulnerable populations with misinformation.

- Using 'Find My iPhone' for Stalking:

Despite its intended use for locating lost or stolen devices, the "Find My iPhone" feature can be



abused for stalking purposes, posing a threat to personal safety and privacy.

- **Using Dating Apps for Catfishing:**

The anonymity provided by dating apps can be exploited for deceptive practices like catfishing, where individuals create fake profiles to manipulate and deceive others for personal gain.

- **Using Mental Health Forums to Sell Drugs:**

Online mental health forums, designed to provide support, can also be used by malicious actors to target vulnerable groups. They can become platforms for illegal activities, such as the sale of drugs, posing significant risks to individuals seeking genuine help.

- **Think About Malicious Actors with Bad Intentions:**

Consider who might use your AI with those malicious intentions?

Consider who will be harmed the most if your AI system was used for those malicious intentions?

You can learn more about identifying potential malicious users at [Tech Ethics Toolkit: Tool 6 Think About the Terrible People](#).

EXPERTISE AND ENGAGEMENT

By engaging with diverse teams and effectively involving stakeholders, you increase your chances of identifying misuse as it helps bring various scenarios to the table. Proceed to [Space 3](#) for staging effective stakeholder engagement.

GO FURTHER

Check out this report for insights into: [The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation](#).

WORKBOOK | Enter your input

Describe how your product can be used for malicious activities and what safeguards you have implemented to prevent them:

Consider who might use your AI with those malicious intentions?

Consider who will be harmed the most if your AI system was used for those malicious activities??

Describe if your AI system requires any special conditions to operate.

2.4 RISK MITIGATION

Task: Collate potential negative consequences from previous steps and identify mitigation strategies for them.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 9: Risk Management System.](#)

RATIONALE

Leaving the negative impacts you identified in the previous task unmitigated can lead to significant legal non-compliance, damage to brand reputation, financial losses, reduced product reliability and loss of customer trust.

DESCRIPTION

Identify and document any negative impacts on citizens and the environment or potential misuses in sections 2.1, 2.2, and 2.3. You then need to consider measures to address any issues that arise.

It's essential to recognize that the process of identifying and addressing risks is ongoing and iterative. This means it needs to happen throughout the product lifecycle. Keep revisiting this task when additional risks are identified.

HOW TO APPROACH

- Summarize the potential negative impact on different stakeholders you identified in [2.1](#), the impact on the environment in [2.2](#), and potential misuses you identified in [2.3](#).
- For each harm and misuse, how likely are the risks, and how significant?
- For each harm and misuse, what is your plan to avoid the risks? You can make your plan from, for example, an organizational, legal, business model, technological or design perspective.

For example: Internet of Things devices can be used by domestic abusers to track and monitor their partners. Once this was highlighted out that this was the case, companies have responded by building in functionalities to their devices that reduce the likelihood of this occurring. The result is systems that actively promote safety by design. For more, see [Safety-by-Design approaches](#).

EXPERTISE AND ENGAGEMENT

To identify mitigation strategies you may wish to consult with your stakeholders in [Task 3.2](#) and engage with AI ethics or domain expertise in [Task 3.3](#). They can advise on how the AI tool can be used to support their needs, rather than in a way that might jeopardise their rights.

GO FURTHER

- We encourage you to explore how you can integrate [Design Justice principles](#) into your impact assessment and risk mitigation processes.
- We use design to sustain, heal and empower our communities, as well as seek liberation from exploitative and oppressive systems.
- We centre the voices of those directly impacted by the outcomes of the process.
- We prioritize design's impact on the community over the intentions of the designer.
- We view change as emergent from an accountable, accessible, and collaborative process, rather than a point at the end of a process.
- We see the role of the designer as a facilitator rather than an expert.
- We believe that everyone is an expert based on their own lived experience, and we all have unique and brilliant contributions to bring to a design process.
- We share design knowledge and tools with our communities. We work towards sustainable, community-led and -controlled outcomes.
- We work towards non-exploitative solutions that reconnect us to the earth and to each other. Before seeking new solutions, we look for what is working at the community level.
- We honour and uplift traditional, indigenous, and local knowledge and practices.

WORKBOOK | Enter your input

Summarize the potential negative impact on different stakeholders you identified in [2.1](#), impact on the environment in [2.2](#), and potential misuses you identified in [2.3](#).

For each harm and misuse, how likely are the risks, and how significant?

For each harm and misuse, what is your plan to avoid the risks? You can make your plan from, for example, an organization, legal, business model, technology, or design perspective.

2.5 HUMAN OVERSIGHT

Task: Design effective mechanisms for human oversight of your system.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 14: Human Oversight](#)

RATIONALE

Ineffective human oversight can lead to overlooking the limitations and risks of AI. It can put people at risk and make it harder to identify and rectify mistakes, potentially compromising the safety and reliability of AI.

DESCRIPTION

Any consequential decision made by an AI system that affects people must be reviewed by a human. During this review, additional factors such as contextual information can be considered before reaching a final decision.

The responsible party/human reviewer of AI that you assigned in 1.2 will be responsible for troubleshooting, managing and overseeing the system during and after deployment.

HOW TO APPROACH

1. Specify how you provide any training for human reviewers to enable them to fully understand the capacities and limitations of the AI system and correctly interpret its output. This training aims to empower human reviewers to duly monitor AI system's operation, while remaining aware of the tendency to over-rely on the output produced by the system.
2. Specify which human oversight mechanism you plan to take. The following mechanisms are taken from the [EU Commission White Paper on Artificial Intelligence](#).
 - Human-in-the-Loop (people will make decisions based on output): the output of the AI system does not become effective unless previously reviewed and validated by a human (e.g. the rejection of an application for social security benefits may be taken by a human only)
 - Human-out-of-the-Loop (people will evaluate system output): the output of the AI system becomes immediately effective, but human intervention is ensured afterwards (e.g. the rejection of an application for a credit card may be processed by an AI system, but a human must subsequently review it).
 - Technical features for Human-in-the-loop: monitoring the AI system while in operation and

the ability to intervene in real time and deactivate it (e.g. a stop button or procedure is available in a driverless car when a human determines that car operation is not safe).

- Technical features for Human in Control: in the design phase, this means imposing operational constraints on the AI system (e.g. a driverless car shall stop operating in certain conditions of low visibility when sensors may become less reliable, or maintain a certain distance in any given condition from the preceding vehicle).

GO FURTHER

Human Oversight is not a perfect safeguard against the harms and dangers of AI. Be alert to 'Automation bias', where practitioners uncritically accept the recommended decision of an algorithm, rather than meaningfully engage with that output.

Human oversight serves as a tool for reviewing AI work, but it does not fully address the errors, mistakes, and biases of AI. For more on the limitations of human oversight, read [The flaws of policies requiring human oversight of government algorithms](#).

WORKBOOK | Enter your input

Specify how you provide any training for human reviewers to enable them to fully understand the capacities and limitations of the AI system and correctly interpret its output.

Specify which human oversight mechanism you plan to take.



SPACE 3

Stakeholder engagement space



OVERVIEW

Directly engaging with the people your technology will affect is vital in ensuring your system meets real-world needs, doesn't discriminate, and actively promotes justice. This Space provides guidance on identifying and engaging with a range of stakeholders. You will return to this Space throughout the project to plan (and document) your team's collaboration with an array of stakeholders.

3.1 IDENTIFY STAKEHOLDERS

Task: Identify your stakeholders and select those to consult with.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 95: Codes of Conduct for Voluntary Application of Specific Requirements](#)

RATIONALE

Identifying stakeholders will allow you to explicitly consider and weigh up different peoples' needs and rights – which may be, at times, conflicting – and ensure your system benefits as many people as possible.

DESCRIPTION

Consulting with stakeholders means collaborating with the people who your technology will impact the most. To design and deploy AI systems in a responsible way, you need to consider a variety of stakeholders. A stakeholder, in the broadest sense of the term, means anyone who has a 'stake' in your project, including the design team and the system's end users; but also 'indirect' stakeholders who will be affected by or have an interest in the project.

HOW TO APPROACH

Identify all stakeholders, including system deployers, primary users, and other affected individuals. Some individuals and groups will be more affected than others so pay special attention to those groups that may be negatively affected by your system and/or who may be considered vulnerable within the given context.

You can also move to [Space 3.2](#) and [Space 3.3](#) to work out how to consult these stakeholders, and ensure you've identified all the potential negative effects and adopted the right mitigation strategies.

1. Identify stakeholders

To identify all relevant stakeholders, you can use the following prompts (adapted from this [Microsoft guidebook](#)):

- **End user (DIRECT STAKEHOLDER):** Who will be most directly involved in using the system? Who will have to interpret system outputs to make decisions? E.g., marketing team,

students

- **Evaluation or decision subjects (DIRECT STAKEHOLDER):** who will be evaluated or monitored by the system, whether or not by choice? Who will the system make predictions or recommendations about? E.g., registered customer.
- **System providers and deployers (DIRECT STAKEHOLDER):** who will troubleshoot, manage, operate, oversee or control the system during and after deployment? Who will own and make decisions about whether to employ a system for specific tasks? E.g., your team or the customer development team.
- **Bystanders (INDIRECT STAKEHOLDER):** who in the vicinity of the deployed system may be impacted by its use? E.g., passers-by.
- **Regulators and civil society organizations (INDIRECT STAKEHOLDER):** who may advocate for regulation of this system or be concerned about compliance? E.g., government health entities.
- **Communities (INDIRECT STAKEHOLDER):** which communities may be affected by the short- or long-term use of the system? E.g., communities with low digital literacy.
- **Associated Parties (INDIRECT STAKEHOLDER):** who may have a substantial interest in the system based on their relationship to other stakeholders? E.g., company partners, family members.

2. Distinguish between the deployer, users and affected individuals (in accordance with the AI Act)

▪ Example 1

If you are developing an AI hiring tool for screening resumes:

Deployer: The entity responsible for implementing and utilizing the AI hiring tool is referred to as the deployer. In this scenario, the recruiting company procuring and using your tool is the deployer.

Users: Human Resources (HR) managers who actively interact with and make use of the AI hiring tool are designated as users. They are the individuals responsible for leveraging the tool's capabilities to streamline the resumé screening process.

Affected individuals: Those impacted directly by the outcomes of the AI hiring tool are the affected individuals. Jobseekers whose resumé undergo evaluation by the AI tool fall into this category. The tool's decisions may influence their chances of employment, which means that the jobseekers are directly affected by the tool's application.

▪ Example 2

If you are developing an algorithm to decide whether to accept a person's credit application:

Deployer: In this scenario, the bank integrating and using your tool is the deployer.

Users: Credit officers who actively interact with and make use of the algorithm are designated as users.

Affected individuals: Credit applicants. The tool's decisions may influence their access to financial resources.

WORKBOOK | Enter your input

List your direct stakeholders (including different groups of users, designers, business decision-makers).

List your indirect stakeholders (such as family members of direct users or other groups that the system affects indirectly).

3.2 ENGAGE WITH STAKEHOLDERS

Task: Engage with identified stakeholders.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 69: Access to the Pool of Experts by the Member States;](#)
[Article 95: Codes of Conduct for Voluntary Application of Specific Requirement.](#)

RATIONALE

Directly engaging with Stakeholders, including end users, customers, and affected communities (especially those you've identified as vulnerable) will provide insights into their needs, concerns and requirements. Understanding these perspectives is essential in developing AI solutions that address real-world problems and promote justice.

DESCRIPTION

You can use this task to plan stakeholder participation activities at various stages of the development and deployment process – from defining the problem you're trying to solve or justifying the use of AI, to how you are monitoring the system – to ensure you can learn from those who will be affected by your system and that they have a say in how the technology is developed.

HOW TO APPROACH

Engaging stakeholders can take on different forms and formats. You should distinguish between consultation and inclusion.

Note: this Space adopts a categorisation of participatory AI methods by [Delgado et al., \(2023\)](#).

Important:

- External stakeholders should be compensated for any time they provide (don't expect your users to do work for free. Consider transport burdens and access constraints. Reaching lower-resourced groups may require going to them).
- To enable genuine inclusion, you must avoid tokenism and 'participation-washing': making it appear as if diverse stakeholders have been included in the design process, while they have merely been consulted (see below).



- Working with some communities – in particular vulnerable groups, such as children or refugees – is important but requires expertise and extra levels of care. You might need to partner with community organisations or AI ethics experts to ensure you know what you’re doing, and that inclusion is meaningful.
- Bear in mind that some communities face many invites to be consulted (e.g. trans people, First Nations people) and therefore may have less time and energy available. Trust may also be an issue if they have faced exploitative research practices in the past. Getting advice from relevant community organisations may be helpful.
- Speak your stakeholders’ language - consider the technology and AI literacy of groups you’re communicating with. Exclude jargon and ensure a shared understanding of the terms you use and technologies you refer to.

Consultation versus inclusion

Consultation

This approach involves direct users only: who comment on a prototype or built system, usually through questionnaires or interviews.

- Why is participation needed? To improve the end user’s experience.
- What is on the table? The user-facing elements of the system and it’s performance.
- Who is involved? Direct users of the system.
- What form does stakeholder participation take? Stakeholders share input through questionnaires or interviews.

We provide more information about engagement form in [Space 4](#).

Inclusion

This approach engages all user groups from early on in a project to inform goals and design; relies on more collaborative methods.

- Why is participation needed? To ensure your AI system’s performance aligns with stakeholders’ needs, goals, and values.
- What is on the table? The overall performance, goals, and impact of the system.
- Who is involved? Both direct and indirect stakeholders, including those representing vulnerable groups.
- What form does stakeholder participation take? Stakeholders discuss preferences, goals, and potential harms with the project team during structured design workshops and/or open-ended discussion sessions

EXPERTISE AND ENGAGEMENT

The way you go about stakeholder engagement will depend on how you’re applying AI and which groups are affected. Usually, you will need to:

- Identify the right community representatives
- Build mutual trust
- Develop a shared vocabulary
- Prepare engagement workshops
- Establish a feedback mechanism (to inform how the involvement changed your design)

You can go through the questionnaire below (adapted from [Birhane et al., 2022](#)) to include self-reflexive participatory design practices and projects that aim to include, rather than merely consult, different stakeholders - notably those most often excluded from tech design. This is especially important if the AI is designed to solve a problem 'on their behalf', like a tool designed to support disabled people in some way. Remember: 'Nothing About Us Without Us!' - [Sasha Costanza-Chock](#).

Self-Reflexive Assessment of Participatory Practices

Consider trust and power dynamics

- Do the project goals that motivate a participatory effort seek to support community interests?
- What efforts will go into building trust with the people and communities involved?
- When there is lack of trust, how is it understood in its historical and structural context (e.g., effects of racism and discrimination)?
- What efforts will be taken to mitigate the effects of power imbalances between participants and the project team?
- What efforts will be taken to mitigate the effects of any power imbalances within the participant group(s)?

Create a safe space for challenge and disagreement

- What efforts will be taken to ensure a transparent and open conversation between participants and the project team?
- What mechanisms are put in place to allow participants to question the existence of the product/tool itself, rather than merely helping reduce harms or improve 'benefits'? Can they say no to a tool being built, and have their voice heard? Evidence of this will also help improve public trust.
- In what ways will the participation process allow for disagreement?

Consider reciprocity and the participant experience

- How do participants experience the process of participation?
- What do participants own and how do they benefit?
- Will the participatory effort be a one-off engagement with the community, or a recurring/ long-term engagement?
- Did participants have the opportunity to refuse participation or withdraw from the process without causing direct or indirect harm to themselves or their communities?

Ensure participation shapes the project

- At what phase of the development process will participation with communities take place?
- Will there be flexibility in the process to influence decisions made in prior phases?
- Could the data source or curation decisions be changed, and data be re-collected based on insights from your participatory efforts?
- Could the plans for evaluation of performance or harms be updated based on insights from your participatory efforts?



GO FURTHER

Note: Task 3.2 adopts a categorization of Participatory AI methods by [Delgado et al., \(2023\)](#).

Collaboration and partnerships

Beyond consultation and inclusion, you can aim for true collaboration between your team and external stakeholders.

- Why is participation needed? To shape the system's scope and purpose.
- What is on the table? Whether and why the system should be built.
- Who is involved? Designers and representatives of the affected community working together as equal partners.
- What form does stakeholder participation take? Co-design activities and long-term partnership.

The University of Michigan provides the [Community Partnerships Playbook](#) for guidance on how to create equitable partnerships between technical and community experts.

Designing with vulnerable groups

- [Design from the Margins, Kennedy School of Government, Harvard](#)
- [Research with Potentially Vulnerable People - Guidance from UK Research and Innovation](#)
- [Children and AI, The Alan Turing Institute](#)

WORKBOOK | Enter your input

Describe what measures you have taken to include stakeholders throughout your development process. Include how you managed any participation by vulnerable groups and how participants were compensated.

Document your answers to the Self-Reflexive Assessment of Participatory Practices below.

3.3 ENGAGE WITH AI ETHICS AND DOMAIN EXPERTISE

Task: Engage with experts in relevant areas including AI Ethics and environmental impact.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 69: Access to the Pool of Experts by the Member States;](#)
[Article 95: Codes of Conduct for Voluntary Application of Specific Requirement.](#)

RATIONALE

Failure to consult AI ethics and domain expertise risks AI systems lacking understanding of their impact and contextual information: leading to ineffective, ethically questionable solutions with negative social impacts across sectors.

DESCRIPTION

AI is a horizontal technology that intervenes in many different sectors. Sector experts, sector-specific consumers, and ethics experts should be consulted to ensure the AI product is effective and has a positive social impact.

HOW TO APPROACH

Examples of AI ethics and domain expertise across development:

PHASE	FUNCTION	EXPLANATION	EXPERTISE
Project ideation	Ethics Consultant	What kind of AI ethics expertise you’re looking for depends on the product. This function requires the expertise of people trained in either ethics or diversity and inclusion work	Legal experts, AI ethicists with socio-technical expertise (including gender studies, critical race theory, cultural studies backgrounds)
Project ideation	Domain experts	This class of experts does not necessarily	e.g. education, infrastructure, hiring

		need to have an ethics background, but knowledge of the application/domain area	
Data collection and pre-processing	Data Protection	You may require legal advice regarding the compliance requirements for privacy protection regulations	Legal and privacy experts
Data collection and pre-processing	Data Collection Strategist	This class of experts can help you design effective strategies of gathering relevant data while minimising the collection of people's personal data	Information Science, Statistics
Data collection and pre-processing	Bias Detection	This function requires diverse expertise in understanding bias, ranging from socio-cultural factors to statistical nuances. You should be sure to consult both technical and social experts in AI bias	Information Science, Fairness and statistics experts, AI ethicists with gender studies/race theory, cultural studies backgrounds
Model Training & Validation	Red Team	Engage with red teaming practice to identify and test the AI model's defences and vulnerabilities	Cybersecurity, machine learning, other relevant disciplines given the model's task
Model Deployment	Usability test	Involve experts in user experience to assess ease of use, user experience and overall effectiveness of your product from the perspective of affected individuals	Human-Computer Interaction
Model Monitoring	Audit	These experts should be involved at various	Cybersecurity, machine learning experts

		moments throughout the lifecycle of the project, as monitoring is an ongoing process. Code auditors can verify if there are technical issues in the model.	
External auditors / the public	Audit	The most responsible tools can be community audited. We recommend that companies collaborate with these auditors to improve public trust	The International Association of Algorithmic Auditors (IAAA), O'Neil Risk Consulting & Algorithmic Auditing (ORCAA), National Audit Office (NAO)

EXPERTISE AND ENGAGEMENT

We need to think differently about ‘AI ethics expertise’. Expertise emerges in conversations between experts and other stakeholders rather than from an individual ‘AI ethics expert’. You need to involve multiple people who can troubleshoot together.

In the case of an AI hiring tool, this might include an affected person (a candidate), a domain expert (a recruiter), a domain expert that deals specifically with bias (a recruiter with expertise in recruitment and diversity), an expert in the specific AI technology, a bias expert in that field (an expert in recruitment and AI ethics), a legal expert etc. It’s also important to take note of what expertise you already have in the team so you know where it needs to be complemented by external expertise.

Domain experts will need to be involved throughout the product lifecycle. For example, in the healthcare context, they will likely include healthcare professionals who need to be consulted throughout, including as part of ideation and usability testing but also on model training & validation.

GO FURTHER

- Explore relevant conferences such as [CHI](#) and [FAccT](#) to identify expertise relevant to your project.
- Seek out diverse voices by reaching out through specialist interest groups such as [Women in AI](#), [Black in AI](#) and the [Indigenous AI Network](#) (see others on the list on the [Responsible AI and Journalism Toolkit](#)).
- Utilize the [ACM library](#) to find information/publications that can aid in understanding and assessing project problems and risks.
- Read the [Data & Society policy](#) brief on the relevance of socio-technical expertise to AI governance.



SPACE 4

User-centered design space



OVERVIEW

In this Space, you'll focus on the people who use your systems (your affected individuals) and on the user-facing elements of your product or service. This includes thinking through audience needs, access, transparency, consent and appeal.

4.1 ACCESS AND DISABILITY JUSTICE

TASK: Design for access through inclusive stakeholder engagement, prototyping, testing and design, based on principles from work on disability justice.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 16: Obligations of Providers of High-Risk AI Systems;](#)
[Article 95: Codes of Conduct for Voluntary Application of Specific Requirements](#)

RATIONALE

An organization that fails to ensure access to users with disabilities is at risk of violating EU and national laws (see news coverage on [accessibility law suits](#)). Complying with Accessibility best practice will ensure you don't risk losing a quarter of your users. In fact, enabling broad access allows you to serve the range of devices on which and situations in which people access technologies (remotely, text-only, low light, poor signal, on the train, in the field, etc.).

DESCRIPTION

What is disability and why design for it?

In contemporary society, disability is frequently viewed similarly to an illness. However, it is society itself that creates disability by being inaccessible to various individuals. This means that streets, stairs, revolving doors and other features of our social infrastructure disable people by making it difficult or impossible for them to navigate everyday life.

Disability = society - accessibility

The quickest way to fail your users is to deny them access. Consider this:

- 1 in 4 adults in Europe have a disability
- All humans experience forms of disability as they age
- Many disabilities are invisible (e.g. hearing loss, colour blindness, dyslexia,

cognitive) Ensuring justice for the many users accessing technology means AI systems should be usable for everyone. Here are some additional reasons for prioritising accessibility:

- Designing a technology to work well for disabled people has been proven to improve everyone's experience of it.
- The [European Accessibility Act](#) requires companies operating in Europe to adapt their digital products and services to accessibility standards by June 2025. (Note: this may vary for micro enterprises - check the latest version of the Act).

HOW TO APPROACH

- **Ensure your designers and developers update their knowledge.** Accessibility relies on design (choices about colour, sound, typography) as well as html and code (meta tags, mark-up and output alternatives) so both teams should be familiar with the principles and standards herein. Note: [beware of overlay tools](#) that promise to make your system accessible automatically-they can make things worse.
- **Read the [design justice principles](#).** Design justice goes beyond accessibility. Consider how these might fit in with the values you listed in [Task 1.5](#).
- **Ensure accessible participation and testing.** Often prototypes are built for able-bodied testers and co-designers, excluding disabled stakeholders. How will you ensure disability is accounted for when you engage with users and stakeholders?
- **Comply with the [European Accessibility Act](#).** This [checklist](#) summarises the act requirements.
- **Comply with the [Web Accessibility Directive](#).** This is critical if your system might be used by the public sector. The [European Commission's approach](#) to web accessibility offers a simple guide that draws on international web accessibility standards ([WCAG](#)).

EXPERTISE AND ENGAGEMENT

In addition to ensuring your design and development team have accessibility expertise and that you engage people of diverse abilities in co-design and testing, you should consider seeking advice from experts at disability advocacy organisations or reach out to teams within your broader organisation that specialise in disability and inclusion.

Remember to do this early (before decisions are made) so that the first version of your system complies with the EU Accessibility Act and accessible interactions are encoded. Engaging too late means invited stakeholders may be exploited, as many of their suggestions and feedback will not be possible to implement retroactively. If engaging early is no longer possible, plan for sufficient resources to rebuild or redesign based on feedback.

GO FURTHER

- Representational harms in disability include the reinforcement of stereotypes by AI

generated images. This [paper on representational harms in GenAI](#) suggests solutions.

- To go beyond accessibility, move beyond user-testing and consultation to participation as inclusion. Head back to 3.2 for more information on deeper partnerships with stakeholders; in this case, impacted individuals with disabilities.
- Try this [Design Justice exercise](#) which provides a quick route to insights for you and/or your team.
- [Cards for humanity](#): this simple online card game lets you make a quick start on digital inclusion. Draw two cards and see if your product or system is usable and enjoyable for different personas.

WORKBOOK | Enter your input

Describe what measures you are taking to ensure your system is accessible to people of diverse abilities. Include whether it complies with the EU Accessibility Act and Web Accessibility Directive.

4.2 MAKE USE OF AI TRANSPARENT

TASK: Make sure it's clear to affected individuals that they're interacting with an AI system or AI-generated content, and if it's being used to make decisions about them.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems](#)

RATIONALE

Prioritising transparency means making sure users know when AI is involved, empowering them to make informed decisions, building trust, and upholding integrity in digital content dissemination.

DESCRIPTION

Inform users that they are interacting with AI, are subject to AI decisions, or are interacting with AI-generated/modified content (e.g. via visible watermarking).

HOW TO APPROACH

- Include noticeable indicators in the user interface that **highlight AI-driven functionalities**, ensuring users are aware when they are interacting with AI.
- If your system will be used to **categorise biometrics**, inform them how it works, even if they are not interacting with it directly.

Failure to inform users about such operations may result in legal consequences. [The Hungarian Data Protection Authority fined Budapest Bank €700,000 for carrying out automated decision-making and profiling based on an AI analysis of customers' emotions in customer service calls without their consent.](#)

- If AI is used to generate and modify **images, sounds or videos to mirror reality**, inform users about these alterations.

EXPERTISE AND ENGAGEMENT

When selecting methods of communicating the information outlined above, it's important to consider the following:



1. Identify your target audience
2. Tailor your language and transparency level to match your target audience. Keep your language clear and concise, providing enough information to inform users without overwhelming them
3. Ensure accessibility of the disclaimer for all users, including those with disabilities. Refer back to [Task 4.1](#) for guidance

You can collaborate with UX or behavioral science researchers to design and evaluate the effectiveness of your communication methods.

GO FUTHER

This article from Partnership on AI shows various methods of labelling AI content: [From Deepfakes to TikTok Filters: How Do You Label AI Content?](#)

WORKBOOK | Enter your input

What measures are you taking to inform affected individuals that they are interacting with an AI system?

What measures are you taking to inform people you are using AI to conduct biometric categorization and emotion recognition?

How do you disclose the use of AI-generated or modified content?

4.3 INFORM STAKEHOLDERS ABOUT YOUR AI SYSTEM

TASK: Establish a communication strategy to inform the deployers, users and affected individuals of the general logic behind the model used in your AI products and highlight any known limitations to ensure that the system is used for its intended purposes.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 13: Transparency and Provision of Information to Deployers](#)

RATIONALE

Effective communication with deployers, users and affected individuals is critical if they are to understand the purpose, scope, and limitations of your AI system. This involves clarifying why and when the system is used, along with potential risks and benefits. Failure to communicate properly can lead to misunderstandings, misuse or unintended consequences.

HOW TO APPROACH

- When communicating with deployers, users and individuals affected by your AI tool, it's crucial to offer accessible and comprehensive information.
- Tailor the complexity and depth of the information for the needs of different stakeholders to ensure that the audience gains a meaningful understanding.
- Go to the workbook to document how you communicate your AI system information to deployers, users and affected individuals.
- Keep your stakeholders updated to ensure they are aware of changes in your practices.

EXPERTISE AND ENGAGEMENT

- Collaborate with communication experts such as copywriters and HCI researchers to simplify technical language, ensuring users can easily understand the underlying principles and limitations of the AI model.
- Conduct tests with representatives from the user group and actively engage with individuals to make sure that the communication strategy about your AI model's operations, capabilities and limitations works for them.



GO FURTHER

To understand best practices for AI documentation, you can draw inspiration from [‘How Better AI Documentation Practices Foster Transparency in Organizations’](#).

WORKBOOK | Enter your input

How do you communicate the following with deployers and individuals affected?

The name and contact information of the system provider

Intended use (see [1.4](#)) and foreseeable misuse (see [2.3](#)) of the system

The benefits, limitations, and potential harms of your system to affected individuals

System performance: the level of accuracy and other metrics that AI system has been tested against

Performance for Specific Persons or Groups: specify if the system’s performance varies for specific individuals or groups

Input data requirement for the intended use

Capacities and limitations: emphasize any conditions in the deployment environment that might adversely impact the system’s performance

What is the Model Logic: draws on [6.3](#) (model card logic) to translate technical terms about how the model works into natural or visual language appropriate for the public. Consult GDPR guidance (Recital 58) on lay language explanations (consult with experts as necessary)

How to report an issue or complaint (see [4.4](#) and [7.4](#))

Potential future modifications and monitoring plan

Maintenance and Care Instructions: provide user-friendly instructions on system maintenance and care measures, including software updates, the expected lifetime of the system and the computational and hardware resources needed

Logs Management Mechanisms (see [7.3](#))

Human oversight measures (see [4.5](#))

4.4 OBTAINING MEANINGFUL CONSENT

TASK: Ensure that you obtain clear, informed, meaningful consent before collecting, processing and storing personal data.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[The definition of consent according to GDPR](#)

RATIONALE

Failing to obtain informed consent for processing people's data can erode trust and damage your reputation. People won't engage with you if they can't trust you with their data or control its use. Additionally, it can lead to substantial fines under GDPR.

DESCRIPTION

What is consent?

Consent can be limiting, because it puts us in a passive position, whereby we are always consenting to someone doing something to us.

This task helps you do better than merely obtaining the consent of affected individuals. By helping you co-create systems with affected individuals, we design Spaces - from Stakeholder Engagement to User Centered Design.

Help ensure that affected individuals are at the centre of data and design decisions. For example, user agreements are usually created by companies and consented to by affected individuals. These agreements could instead be drafted with affected individuals through the methods outlined in Spaces 3 and 4.

Consent can be undermined by:

- 'Nudging', which can result in unconscious or unintentional behaviour change
- having to 'opt out'
- Not having access to an alternative (non-AI) system.

Meaningful consent means:

- A system that doesn't coerce behavioural change
- Opting in rather than having to opt out
- Having the option to not use a system
- Empowering affected individuals by protecting their data.

Consider integrating tools that modernise the consent process by empowering rights holders. Affected individuals should feel empowered to opt out of using the tool. There should therefore be a non-AI process that the AI system runs alongside, so they do not feel coerced into using the system (to, for example, retrieve their salary, access food in humanitarian aid contexts, or apply for a job).

Consider whether certain options are more perceivable. Are users required to interact with a consent widget before they can access contents of the system containing information they need before they can give or withhold consent?

HOW TO APPROACH

- **Ensure users are opting in to using the system rather than having to opt out.** It must be clear exactly what affected individuals are consenting to when they consent to have their data collected
- **Consent must be continual.** Affected individuals who have not consented to the use of their data may be willing to do so to the collection of some of it once they have more information about how the system works. Equally, an affected individual may consent early on, then should be able to withdraw their consent later down the line
- **Make it clear at what cost users are giving their consent.** Allow the affected individual to make a cost-benefit analysis based on potential knock-on effects of giving data away. To gain user trust, it's important that products are transparent about costs as well as benefits
- **Improve AI literacy as part of the system.** If the user understands how the system works, they're more likely to enter willingly into an exchange of information, or be able to abstain or contest it
- **Ensure the system is not manipulating users behaviour.** Read this to avoid manipulative ways of obtaining consent: [Dark Patterns: 10 Examples of Manipulative Consent Requests](#)
- **Make the consent layer as granular as possible.** Rather than requiring users to either opt in or opt out entirely, can they choose to consent to specific requests or features of the tool while opting out of others?
- **Ensure a non-AI process is in place.** Discuss with your client which non-AI and/or non-tech alternative pathways can be offered. Affected individuals won't be able to meaningfully consent to using a tool if AI is the only way they can, for example, retrieve their salary
- **Design.** Good design makes consent more meaningful. Consent boxes can encourage complacency
- **Ensure the data you're using has been scraped ethically.** Integrate <https://api.spawning.ai/spawning-apiAPI> to automatically comply with opt-outs not just for image data; but for text, audio, videos and more. For example, eligible datasets hosted on Hugging Face show a data report powered by Spawning's API, informing model trainers what data has been opted out and how to remove it



GO FURTHER

Check out [ToSDR](#), a community-driven effort to analyse websites’ privacy policies and grade their respect for users’ privacy.

WORKBOOK | Enter your input

Describe how you currently obtain (or will obtain) people’s consent to ensure they understand and have control over how their personal information is processed by your AI system?

For each statement below, please describe whether your AI system enables this interaction and how. Provide as much detail as possible.

Statement 1: End users/affected individuals can opt in to receive the AI system’s output

Statement 2: End users/affected individuals can opt out of receiving the AI system’s output

Statement 3: End users/affected individuals can correct or challenge the AI system’s output

Statement 4: End users can reverse the AI system’s output after it has been generated

4.5 STAKEHOLDER FEEDBACK, REQUEST AND APPEAL

TASK: Ensure stakeholders can submit feedback, lodge complaints, or report incidents.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 85: Right to Lodge a Complaint with a Market Surveillance Authority;](#)
[Recital 170.](#)

RATIONALE

Establishing feedback mechanisms not only helps you identify and address stakeholders' concerns but also enhances their experience and trust: enabling sustainable growth and a competitive advantage.

DESCRIPTION

It is essential for companies to actively consider feedback and complaints from those affected. This involves establishing a clear process for individuals to submit their feedback and individual rights requests, ensuring each submission receives a timely response.

HOW TO APPROACH

- **Empowerment through Feedback and Reporting.** Establish a user-friendly feedback and reporting channel, ensuring accessibility and understanding for all stakeholders, regardless of their background or level of technological literacy.
- **Responsive Communication.** Ensure that stakeholders who report issues promptly receive notifications detailing how and when they can expect a response.
- **Address Dissatisfaction.** Inform stakeholders about available steps they can take if they are dissatisfied with how you respond to their feedbacks.
- **Act on Feedback.** Note how you will regularly review and analyse the feedback received. Use this information to identify patterns, address common concerns, and make improvements to your AI product.



EXPERTISE AND ENGAGEMENT

To design user-friendly and accessible feedback and reporting channels, collaborate with stakeholders to understand their needs and conduct usability testing to identify and address any issues. Revisit [Space 3](#) to conduct stakeholder engagement properly. Regularly review feedback channel performance and gather input from users on their experience. Use this information to make continuous improvements, enhancing both the functionality and user satisfaction of the feedback mechanisms.

WORKBOOK | Enter your input

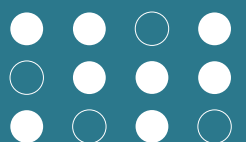
List the channels available for stakeholders to provide feedback, such as email, hotline calls, online forms, or dedicated feedback portals. How do you ensure these channels are easily accessible? Do you support anonymous feedback? If not, provide your reasoning below

Describe the process for handling and responding to feedback and complaints. Explain how stakeholders will be notified about the status of their submissions and when they can expect a response



SPACE 5

Data governance space



OVERVIEW

In this Space, you will engage in tasks designed for effective management, quality assurance and compliance of data governance throughout the data lifecycle - from collection and processing to evaluation.

The focus on quality and improving representation in data is crucial. The first step is becoming aware of the power dynamics at play in your data: from how it is collected and who it represents, to how it is selected, cleaned and used.

We encourage you to handle data in a way that shifts power back towards marginalized and affected individuals.

This section will help you become better equipped to promote equality and justice in the data practices of AI development and deployment.

Start with some key lessons from [Data Feminism](#) and [‘Towards Accountability for Machine Learning Datasets’](#):

- **Consider Data Context.** Data is not inherently neutral or objective.
- **Document Assumptions and Limitations.** Understand and record the assumptions and limitations of the dataset, especially regarding whose interests and goals it serves.
- **Verify Data Quality.** Treat datasets with scepticism until proven reliable. Conduct independent quality analyses and checks throughout the dataset’s development.
- **Acknowledge Invisible Data Labor.** Recognize and account for the unseen labour involved in data collection and processing, and how it shapes the dataset.

5.1 UNDERSTAND HOW POWER OPERATES IN DATA

TASK: Explore power dynamics and their impact on data practices.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 10: Data and Data Governance.](#)

RATIONALE

Data aren't neutral; they reflect social contexts. Understanding power in data is critical, going beyond just technical aspects. Understanding how power operates in data can help us grasp the consequences of collecting and processing data.

DESCRIPTION

Data originates from an unequal world, and AI is created and implemented within this same framework. Therefore, it's not possible to eliminate bias from your datasets and AI. Striving to collect, prepare and utilize data in ways that specifically benefit historically marginalized communities can counterbalance structural inequality and pave the way for more inclusive AI.

HOW TO APPROACH

- **Reflect on the negative impacts and potential misuse of your AI system that you identified in Space 2.** Consider how this relates to your data practices: which data is collected, whose data is under-represented or excluded, and who has access to and analyses the data.
- **Establish community-centric data collection mechanisms.** Prioritize collecting data from diverse communities and demographics to ensure the affected individuals are well represented in the data. Think about the effects of excluded or under-representing in your dataset.

Read the UK government's recommended best practices for collecting, analysing, and reporting [ethnicity data](#).

- **Foster trust and collaboration with local communities through transparent communication and involvement in the data collection process.** It's often risky for diverse groups to be represented in datasets. They need to be aware of these risks. For instance, the [Gender Shades](#) project developed a Pilot Parliaments Benchmark (PPB) to achieve better representation of women of colour in datasets used for facial recognition. However, critics

noted that facial recognition disproportionately harms women of colour (through, for example, surveillance). Therefore, this more representative dataset was also making it easier for facial recognition to see - and harm - women of colour.

- **Recognize the ethical implications of data cleaning, tidying, and processing, keeping record of rationales behind your decisions and ensuring careful integration of both technical and ethical considerations throughout the procedure.** Datasets frequently contain missing values, often seen as empty fields representing unanswered survey questions. Handling missing data typically involves excluding entries with missing values or imputing new ones through a process called imputation.

However, both approaches can lead to significant bias, as missing data might contain essential insights into social issues. For instance, in medical data, missing values might be more common among low-income patients who could refuse costly medical tests. Similarly, certain groups may face questionnaire challenges, like elderly individuals coping with small fonts or non-native English speakers grappling with complex language. For further information, read [Fairness without Imputation: A Decision Tree Approach for Fair Prediction with Missing Values](#).

EXPERTISE AND ENGAGEMENT

- Acknowledge and incorporate the multidisciplinary skills needed for data preparation, possibly engaging experts such as data ethicists, statisticians and sociologists.
- Consider the benefits and risks of choosing cleaner data at larger scale that is relatively easy and quick to purchase, or more accurate data at a local scale for which one must engage and build trust with community groups.
- Understand the importance of the data ‘tidying’ and selection work, which is crucial to the proper maintenance of a system.
- Recognize that data ‘tidying’ involves a series of decisions about what to include and exclude. These are ethical decisions.

GO FURTHER

- Read [The Seven Principles of Data Feminism](#).
- Read [Potential sources of bias across the AI lifecycle](#) section of ‘Fairness in the AI lifecycle’ published by UK Information Commissioner’s Office.
- Read [A Survey on Bias and Fairness in Machine Learning](#).
- Read [A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle](#).

Familiarize yourself with the provided resources and consider how they can be applied to your data collection and processing efforts.

5.2 GDPR COMPLIANCE

TASK: Complete GDPR compliance assessment.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[General Data Protection Regulation \(GDPR\)](#).

RATIONALE

The GDPR governs the use of personal data within the European Union. If your AI system is involved in the collection, utilization, or storage of individuals' personal data, it is imperative to ensure compliance with the GDPR. Failure to adhere to these regulations may result in serious repercussions, including financial loss, identity fraud and a loss of trust.

HOW TO APPROACH

Conduct [GDPR Compliance assessment](#) and attached the result. Make sure you at least include answers to the following questions in your attachment:

- Are you obtaining data from third parties, and if so, are they GDPR compliant?
- What data do you need?
- Why do you need these data?
- Are you processing any sensitive personal data?
- Have you obtained explicit consent from individuals for processing their personal data? If you are processing any sensitive personal data, what measures are in place to ensure their protection?
- Are you only collecting data that is relevant, and limited to your purpose(s)?
- How are you storing and securing the data?
- How long do you intend to retain the data for?

GO FURTHER

Explore [resources](#) provided by the UK Information Commissioner's Office to get started with data protection.

Adhering to GDPR's data minimization principle means only processing necessary, relevant, and limited data for specific purposes. Beyond compliance, it brings many benefits to your organization:

- Mitigates the harmful effects of data misuse, unauthorized access, or breaches.

- Lowers expenses associated with managing and securing extensive datasets.
- Emphasizes data quality over quantity, resulting in precise insights, informed decisions, and improved outcomes.

WORKBOOK | Enter your input

Name your Data Protection Officer that has been designated to complete the GDPR compliance assessment.

Conduct GDPR Compliance assessment and provide the necessary information and documentation.

5.3 DATA QUALITY ANALYSIS

TASK: Conduct and record a statistical analysis of your training, testing and validation data.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 10: Data and Data Governance.](#)

RATIONALE

Data quality analysis allows you to identify issues with the data and take appropriate measures to address them, fulfilling the AI Act data governance requirements.

DESCRIPTION

You should conduct and record a statistical analysis of training, testing and validation data to demonstrate that they are **relevant, sufficiently representative and to the best extent possible, free from errors and complete in view of the intended purpose.**

It is also your job to ensure the accuracy and quality of data even if you use data from external sources to train your AI system.

A recent case involving TransUnion Rental Screening Solutions Inc., [settling for \\$15 million over allegations of faulty AI producing flawed tenant screening reports](#), serves as a reminder. TURSS obtained eviction records from a third-party provider but failed to ensure their accuracy. As part of the settlement, Trans Union and TURSS must implement strict procedures to ensure eviction data precision and disclose its origin upon consumer request.

HOW TO APPROACH

Provide Descriptive Statistics of your dataset:

- Utilize descriptive statistics to understand the distribution, central tendency and variability of your dataset.
- Assess the descriptive statistics for each subgroup within your dataset, such as race, gender, industry, and sources.
- Evaluate the representativeness of your data across diverse population groups.
- Consider whether oversampling is needed for minority groups in your dataset.
- Verify if your statistical analysis confirms data quality across different groups or sources.

Try out [Facets](#) to generate visualizations that aid in understanding and analysing datasets.

Provide information about your Data Labelling practices (if applicable):

- **Criteria Specification:** provide detailed criteria for labelling data, including comprehensive descriptions for all potential labels, examples illustrating each label, and coverage of edge cases.
- **Accountability Definition:** clearly define lines of accountability for data labeling, delineating roles and responsibilities.
- **Procedure Description:** detail the data labelling procedure, including training provided to annotators, the review process, and any quality assurance measures implemented.
- **Agreement Level Documentation:** Document statistics on the level of agreement achieved by human labellers during the process. Systematic disagreements about ground truth labels between annotators of different sociodemographic groups can offer valuable insights into the representation of diverse perspectives.

Provide information about how Data Cleaning and Transformation.

Provide a description if you have applied the following transformations to your dataset:

- Anomaly Detection
- Rectifying Mismatched Values
- Addressing Missing Values
- Data Type Conversion
- Balancing classes in the data
- Data Aggregation
- Reducing Dimensionality
- Combining Input Sources
- Redaction or Anonymization
- Data normalization
- Any other methods (please specify).

EXPERTISE AND ENGAGEMENT

We advocate for participatory data stewardship, which opposes opaque or manipulative practices in data collection, storage, sharing and use. Instead, it promotes empowering individuals to inform, shape and sometimes govern their own data. For further insights, refer to the [Ada Lovelace Institute's report](#) on participatory data stewardship.

GO FURTHER

- Explore the [Know your data project](#)
- Read: [Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights](#)
- For a deeper understanding of the ethical implications of data annotation, read: [Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation](#). Reflect on the notion of invisible workers discussed in section 1.3. Consider how your data annotators might work as invisible workers for your AI product
- Data labeling is a process that is inherently influenced by a variety of factors including the cultural, social, and political contexts in which the data is collected, as well as the personal



beliefs, biases and experiences of the individuals involved. It should not be understood as purely technical and objective. For further information, read [‘Excavating AI: The Politics of Images in Machine Learning Training Sets’](#).

WORKBOOK

Provide the necessary information and documentation to statistical analysis of your training, testing and validation datasets, including any visualizations, reports, or summaries. Ensure that the attachment covers all the aspects mentioned in “How to approach” (Descriptive Statistics, Data Labeling practices, Data Cleaning and Transformation), demonstrating a comprehensive analysis of your data quality.

5.4 DATASHEETS

TASK: Complete the datasheet documentation.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 10: Data and Data Governance.](#)

RATIONALE

Completing this documentation for your datasets serves not only as a compliance requirement but also as an opportunity for thoughtful analysis.

DESCRIPTION

Filling out the table below should encourage you to reflect on how collecting and processing your data affects various metrics of your AI system, as well as the individuals affected by it.

GO FURTHER


For additional insights on factors to consider during the documentation process, read the study [Datasheets for Datasets](#).

WORKBOOK | Enter your input

Complete the following table, you can skip if the question is not applicable to you.

Motivation

- *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
- *Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?*
- *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*



Composition

- *What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
- *How many instances are there in total (of each type, if appropriate)?*
- *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*
- *What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.*
- *Is there a label or target associated with each instance? If so, please provide a description.*
- *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*
- *Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
- *Are there recommended data splits (training, development/validation, testing)?*

If so, please provide a description of these splits, explaining the rationale behind them.

- *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
- *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
- *Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals’ non-public communications)? If so, please provide a description.*
- *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
- *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*
- *Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
- *Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.*
- *Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
- *Any other comments?*

Collection Process

- *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
- *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?*
- *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?*
- *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?*
- *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
- *Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
- *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*
- *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
- *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
- *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
- *Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please*

provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

- *Any other comments?*

Preprocessing/Cleaning/Labelling

- *Was any preprocessing/cleaning/labelling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.*
- *Was the “raw” data saved in addition to the preprocessed/cleaned/labelled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*
- *Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.*
- *Any other comments?*

Uses

- *Has the dataset been used for any tasks already? If so, please provide a description.*
- *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
- *What (other) tasks could the dataset be used for? Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labelled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*
- *Are there tasks for which the dataset should not be used? If so, please provide a description.*
- *Any other comments?*

Distribution

- *Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
- *How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
- *When will the dataset be distributed?*
- *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
- *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
- *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
- *Any other comments?*

Maintenance

- *Who will be supporting/hosting/maintaining the dataset? How can the owner/curator/manager of the dataset be contacted (e.g., email address)?*
- *Is there an erratum? If so, please provide a link or other access point. Will the dataset be updated (e.g., to correct labelling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*
- *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
- *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*
- *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*
- *Any other comments?*

SPACE 6

Model governance space



OVERVIEW

This Space offers guidance on implementing model governance with an emphasis on addressing social inequality alongside technical standards.

- [Task 6.1](#) underscores the need to prioritize ‘fairness’ and ‘explainability’ throughout the model development process. To effectively prioritize these aspects, it’s essential to draw upon the insights obtained from previous tasks, particularly those that focus on how existing social inequality relates to your AI system.
- [Task 6.2](#) involves creating and maintaining a traceable record of all significant design decisions. Each choice should be justified in a way that considers both technical requirements and societal implications.
- [Task 6.3](#) helps you create a Model Card Documentation. This document serves as a comprehensive record of the model’s characteristics, including its training, evaluation, and validation processes.
- [Task 6.4](#) guides you to conduct a fairness evaluation for your model. However, it’s vital to recognize that fairness is inherently contextual and domain-specific, which means that even when fairness techniques are employed and fairness objectives are met, it doesn’t necessarily imply that your model is universally fair. This exercise should prompt you to think beyond surface-level assessments to truly understand the model’s impact on different stakeholders.

6.1 MODEL DEVELOPMENT BEYOND TECHNICAL CONSIDERATIONS

TASK: Think about strategies to prioritize ‘fairness’ and ‘explainability’ throughout the model development process.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Recital 27;](#)
[Ethics guidelines for trustworthy AI.](#)

RATIONALE

Failing to consider ethical considerations alongside technical performance may lead to unethical product decisions that negatively affect users and society.

DESCRIPTION

This task helps you recognize that the selection of your model involves more than just technical considerations. It should be completed before you begin building and training your model.

HOW TO APPROACH

- If the problem you aim to address with your AI system has historical inequalities that make the tool risky or harmful for particular groups of people (refer to [1.3](#) and [2.1](#)), ensuring fairness must be a primary consideration in your model development.
- For instance, hiring processes in some sectors have historically favored one gender over another. When AI products attempt to solve these problems, they might inadvertently perpetuate or amplify these biases if not designed thoughtfully.
- [Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.](#) If your system will be used to make high-stake decisions (which can directly or indirectly affect people’s fundamental rights, e.g., health, safety, access to education and work), it is essential to prioritize interpretable models over black-boxed ones.

EXPERTISE AND ENGAGEMENT

Provide resources and training on equitable and ethical technology design to your team members, especially developers. Ensure they can incorporate ethical considerations into work.

GO FURTHER

- **Please Note:** Even when fairness techniques are employed, and certain fairness objectives are met, it doesn't necessarily imply that the model is universally fair. There are dozens of mathematical definitions of fairness. Fairness is both contextual and domain specific. Its meaning can vary among individuals, groups and cultures and can shift over time.
- To learn more about fairness for AI, explore [Hands-on tutorial on ML Fairness](#).
- To learn more about Interpretable [Machine Learning](#), explore [Interpretable Machine Learning: A Guide for Making Black Box Models Explainable](#).
- To learn more about how to make decisions made by your AI understandable, explore [Explaining decisions made with AI](#).

WORKBOOK | Enter your input

Can your AI system impact users or people with different demographic characteristics in varied ways (refer to [2.1](#))? If yes, how are you addressing fairness issues in your model design?

If your system is used for high-stake decisions, how are you ensuring your model and its outputs are interpretable for people using it and affected by it?

6.2 DESIGN DECISION LOG

TASK: Create a traceable record of all design decisions made throughout the process and model development.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 17: Quality Management System.](#)

RATIONALE

The design decision log not only enhances transparency but also contributes to developing a systematic approach with which to understand your design choices and make good decisions about model type, training process and evaluation metrics.

DESCRIPTION

Create a traceable record of all major design decisions made throughout the model development, capturing the options considered, trade-offs evaluated, selected choices, and the reasoning behind each decision.

HOW TO APPROACH

When making and recording design decisions, take all the exercises you've done regarding social context and the impact of your product on various stakeholders in [Spaces 1](#), [Space 2](#), [Space 3](#) and [Space 4](#) into account.

Make sure your Design Decision Log includes the following information.

1. Model Type Selection:

- **Options considered:** list all potential model types you considered.
- **Trade-offs evaluated:** discuss the pros and cons of each model type, especially in light of the specific problem you're solving.
- **Selected choice:** detail the model type you opted for.
- **Reasoning:** justify your choice of model types from the following aspects:
 - **Interpretability:** elaborate on how your choice aids in making the model's results understandable and explainable.
 - **Privacy:** explain how the chosen model type addresses any privacy concerns.
 - **Robustness:** explain how the chosen model type ensures consistent and predictions

- **Fairness:** detail how the model type promotes unbiased results and mitigates potential prejudices.

2. Training Process:

- **Options considered:** list the training techniques and processes you considered.
- **Trade-offs evaluated:** analyze the implications of each technique in the context of your data and problem.
- **Selected choice:** highlight the training technique(s) you employed.
- **Reasoning:** justify why the chosen training step(s) is(are) optimal for your model.

3. Evaluation Metrics:

- **Options considered:** enumerate the different metrics you considered for evaluating your model's performance.
- **Trade-offs evaluated:** assess the potential benefits and drawbacks of each metric.
- **Selected choice:** specify the metric(s) you adopted for your final evaluation.
- **Reasoning:** rationalize why the selected metric(s) provides the most comprehensive and relevant assessment of your model's capabilities considering both performance and the statistical fairness matrix.

EXPERTISE AND ENGAGEMENT

Use [Space 3](#) to support communication among diverse stakeholders or departments within your organization, including data science, legal, ethics and product management. This collaborative effort facilitates a comprehensive approach to decision-making in AI development, taking technical, ethical, and legal considerations fully into account.

WORKBOOK | Enter your input

Complete the Design Decision Log

Model Type Selection

What different model types did you consider?

What are the pros and cons of each model type?

Which model did you choose in the end? Why did you choose this model type?

Training Process

What training techniques and processes did you consider?

What are the implications of each technique?

What processing technique(s) did you employ in the end? Why are the chosen technique(s) optimal?

Evaluation Metrics

What different metrics did you consider for evaluation?

What are the potential benefits and drawbacks of each metric?

Which metric(s) did you adopt for final evaluation?

Why do the selected metric(s) provide the best assessment?

6.3 YOUR MODEL CARD

TASK: Complete the Model Card documentation.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Annex IV: Technical Documentation](#)
[Referred to in Article 11\(1\).](#)

RATIONALE

Completing the Model Card documentation will help you fulfil the AI Act technical documentation requirements and ensure transparency.

DESCRIPTION

When completing the Model Card documentation, we encourage you to continue the practice of the design decision log in [Task 6.2](#). This involves providing information not only about the final decisions made for each section in the Model Card, but also about the decision-making process and the reasoning behind each decision.

WORKBOOK | Enter your input

Complete the [Hugging Face Model Card](#) or similar model documentation and attach your documentation as a link

6.4 WHAT DOES 'FAIRNESS' MEAN FOR YOUR AI SYSTEM?

TASK: Understand what 'fairness' means for your AI system and conduct bias detection and fairness evaluation.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Recital 27;](#)
[Ethics guidelines for trustworthy AI.](#)

RATIONALE

Evaluating your model against different fairness metrics will provide you with a measurable fairness goal that can be operationalized into the AI system.

DESCRIPTION

Important: even when fairness techniques are employed and fairness objectives are met, it doesn't necessarily mean that the model is universally fair. Fairness is both contextual and domain specific. Its meaning can vary among individuals, groups, and cultures, and it can shift over time.

HOW TO APPROACH

Consider both group and individual fairness.

- **Group fairness** relates to demographic/statistical parity, equalised odds, true positive rate parity and positive predictive value parity. It demands equitable treatment for protected groups comparable to the advantaged group or the overall population. You can test group fairness for the following protected attributes: age, class (income or socioeconomic status), gender, race and ethnicity, religion, sexual orientation, disability. This list is not exhaustive and protected attributes vary across the world.
- **Individual fairness** stipulates consistent treatment of individuals.

Choose your fairness metrics by considering which groups of people may be disproportionately negatively impacted by your system and in what ways.

- Familiarize yourself with [common fairness metrics](#).
- Refer to the Fairness Decision Tree for guidance on appropriate fairness metrics.
- The fairness of your system cannot be fully assessed without understanding the lived experiences and perceptions of those who bear its unfair outcomes.

Evaluate your model performance against chosen fairness metrics.

You can use the following tools to facilitate your evaluation:

- [Fairness Assessment.](#)
- [Bias and Fairness Audit Toolkit.](#)

Mitigate fairness-related issues and harms.

You can consider appropriate [mitigation techniques](#). However, be mindful of the potential false sense of security provided by technically optimized fairness metrics, as this does not guarantee fair outcomes.

EXPERTISE AND ENGAGEMENT

- Engage with stakeholders (Space 3) and experts to assess your fairness metrics in relation to the bigger picture of the application context.
- Discuss the advantages and disadvantages of different fairness metrics with your team and stakeholders.

GO FURTHER

- Explore the existing technical standard: [ISO/IEC TR 24027:2021 Information technology - Artificial intelligence \(AI\) - Bias in AI systems and AI aided decision making.](#)
- Read: [Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI.](#)

WORKBOOK | Enter your input

What is your definition of fairness for your AI product? How did you reach the definition?

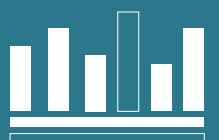
What metric(s) do you use to assess fairness? How did you reach the decision?

What are the outcomes of your fairness assessment?

What steps or strategies did you take to mitigate fairness-related issues and harms?

SPACE 7

Monitoring, evaluation and care space



OVERVIEW

Ethics is an ongoing process, requiring constant maintenance, updates, rethinks and tweaks to the system. This section provides a task list and guide to the Evaluation and Care of your AI system. The tasks here should be performed at least once pre-deployment, then regularly post-deployment.

What the AI Act says about monitoring, evaluation, quality management: [Article 17: Quality Management System](#).

Key principles to uphold in this space include:

- 1. Deploy with care.**
- 2. Be transparent and communicate limitations and harms.** Openly communicate the capabilities, limitations and potential risks associated with your AI system. Provide clear and accessible documentation to deployers, users and people affected, enabling informed decision-making and mitigating the likelihood of unintended consequences.
- 3. Enable contestability and refusal.** Facilitate mechanisms for people to

contest and refuse decisions made by your system. Incorporate feedback loops and avenues for recourse to address harms.

- 4. Ensure timely fixes and fair compensation.** Implement protocols for timely resolution of issues and fair compensation for individuals adversely affected by the AI system. Establish clear channels for reporting and addressing harms, with mechanisms in place to rectify errors, provide redress, and prevent recurrence.
- 5. Iteratively improve your system.** Embrace a culture of continuous improvement, wherein feedback and insights gathered from real-world usage inform iterative refinements to the AI system.

Pre-deployment refers to all activities that occur before your AI system is released or integrated into real-world operations.

Post-deployment refers to all activities undertaken after the AI system is released or goes live. Schedule a date to repeat the tasks in this section.



7.1 ROBUSTNESS TESTING

TASK: Perform robustness testing.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 15: Accuracy, Robustness and Cybersecurity.](#)

RATIONALE

Ensures the reliability and security of your AI product, catching bugs and vulnerabilities before they become problems at scale.

DESCRIPTION

Robustness testing assesses whether your system performs as expected and checks its ability to resist attacks, challenges, and variations. It helps identify and address any limitations or vulnerabilities in your system before they escalate into serious problems. You should complete this task bearing in mind the intended use cases you identified in Space 1.4.

HOW TO APPROACH

Explore various robustness testing methods, including:

- **Adversarial attacks:** assess the model's resilience against data poisoning attacks and model evasion attacks.
- **Data perturbation:** introduce variations in input data by adding noise, flipping pixels, and adjusting data intensity to evaluate the model's response.
- **Hyperparameter tuning:** evaluate the model's performance under different hyperparameter settings, such as learning rate, batch size, and regularization strength.
- **Distributional shifts:** test the model's performance under diverse data distributions, including variations in classes, domains, and environments.

Revisit the risk mitigation section (see [Task 2.4](#)) and revise user instructions (proceed to [7.2](#)) as needed, considering the insights obtained from robustness testing. This includes addressing any identified issues related to the impact, limitations or vulnerabilities of your model.

EXPERTISE AND ENGAGEMENT

Community audits or collective red teaming as part of robustness testing provides an invaluable opportunity to harness collective insights and experiences to assess and enhance the

robustness of AI systems. Limitations and harmful behaviours of your systems are challenging to detect outside of situated contexts of use. By enabling stakeholders to test the AI system in the ways they prefer, developers can benefit from their situated knowledge, obtaining insights into how the system performs under varying conditions. This method not only broadens the range of testing environments but also integrates practical, real-world relevance into the assessment process.

Examples of Community Audits: in 2020, Colin Madland, a PhD researcher at the University of Victoria, posted an image on Twitter of himself and a Black colleague who had been erased from a Zoom call after Zoom's algorithm failed to recognise his face. Twitter automatically cropped the image to only show Madland. After Madland posted this on Twitter, users joined him in testing the [Twitter cropping algorithm](#). You can explore more examples of Community Audits [in this paper](#).

Integrating Bug Bounty programs into your robustness testing is a proactive approach to identifying issues with your system. These programs offer monetary rewards to ethical hackers for identifying and reporting vulnerabilities or bugs to application developers. By extending these incentives to individuals who uncover algorithmic bugs, vulnerabilities and potential harms, you engage a wider community in the early detection and remediation of such issues, enhancing the robustness of your AI systems.

GO FURTHER

Explore additional testing tools and resources for robustness testing at: [Adversarial Robustness Toolbox](#).

WORKBOOK | Enter your input

Reflecting on the risks regarding Security, Reliability and Robustness you identified in [Task 2.1](#), what measures do you take to manage, mitigate or address these risks?

7.2 RECORD KEEPING

TASK: Develop a record keeping mechanism for product updates, testing, and events.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 12: Record Keeping.](#)

RATIONALE

Failure to keep a record of your system's operations can result in serious issues, including regulatory non-compliance, difficulty in identifying the errors, and challenges in providing information.

DESCRIPTION

Implement a built-in functionality that automatically records operations and events throughout the lifecycle of your system. Develop a system for logging product updates. Document and define a testing protocol that can be used regularly.

HOW TO APPROACH

Implement automatic event recording:

- Develop a mechanism within the AI system to automatically record events (logs) throughout its entire operational lifetime. For example, if you are using an AI model to make decisions about a person's credit application, every time your model makes a credit decision there needs to be a record of it.

Key considerations for the event record:

- Period recording: include functionality to record the period of each use of the high-risk AI system, capturing start date and time, end date and time for each use.
- Record reference database information: implement logging to record the reference database against which input data has been checked by the system.
- Capture input data for matches: develop logging capabilities to capture the input data for which the search has led to a match.
- Identify personnel involved in verification: redocument model information and contractual requirements at every system update. (for example, update the model information when re-training the system or using datasets with new contractual requirements).

WORKBOOK | Enter your input

Summarize: how you’re planning to implement a record keeping system?

7.3 INCIDENT RESPONSE AND REPORTING

TASK: Report serious incidents to the surveillance authorities of the Member States where the incident occurred, meeting regulatory obligations.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 73: Reporting of Serious Incidents.](#)

RATIONALE

It is mandatory to promptly inform the relevant market surveillance authorities whenever a serious incident related to your AI system occurs.

DESCRIPTION

When reporting a serious incident related to your AI system, it is crucial to follow a well-defined process to ensure compliance with regulatory requirements and maintain transparency.

HOW TO APPROACH

When fulfilling your reporting obligation, you need to consider:

Nature of incidents: a 'serious incident' refers to an occurrence or malfunction of an AI system leading directly or indirectly to:

- (a) Death of a person or serious harm to health
- (b) Serious and irreversible disruption of critical infrastructure management or operation
- (c) Breach of Union law obligations protecting fundamental rights
- (d) Serious harm to property or the environment.

Notification procedure: notify the relevant authorities immediately upon identifying a serious incident. Once a causal link between the AI system and the incident is established or reasonably likely, submit a detailed report.

Investigation and remediation: upon notification, conduct comprehensive investigations, including risk assessments and implementation of corrective measures.

WORKBOOK | Enter your input

You should maintain detailed records of incidents and responses as part of them regulatory compliance. The documentation should include:

- Incident Details: Description of the incident, its impact, and initial actions taken.
- Investigation Findings: Results of investigations, risk assessments, and root cause analysis.
- Corrective Measures: Steps taken to mitigate risks and prevent recurrence.
- Communication Records: Documentation of notifications to market surveillance authorities and relevant stakeholders.

Describe your approach to reporting serious incidents.

7.4 CONTINUOUS MONITORING AND EVALUATION

TASK: Continuously monitor and evaluate input data and model performance.

WHAT THE AI ACT SAYS ABOUT THIS TASK

[Article 72: Post-Market Monitoring by Providers and Post-Market Monitoring Plan for High-Risk AI Systems.](#)

RATIONALE

Without ongoing monitoring post-deployment, undetected issues like data drift, model drift or biases can lead to reduced customer satisfaction, higher operational inefficiencies, and potential financial losses from inaccurate AI output.

DESCRIPTION

Continuously monitor and evaluate the AI model's input data and performance to detect and address any potential issues that arise after deployment and to ensure the model remains accurate and reliable over time.

HOW TO APPROACH

Best Practices to Monitor Data:

- Data validation: implement validation checks to ensure that the input data is consistent with the expected format and range.
- Data drift detection: AI models can degrade over time due to changes in input data (known as data drift). The drift can be evaluated both with univariate methods regarding each feature and with a multivariate approach. Check [frouros](#) and [nannyml](#) libraries for a Python implementation of the common data drift detection methods.

Best Practices to Monitor Model:

- Model drift detection: monitor the model's behaviour over time to identify any changes in its predictions or outputs (known as model drift).
- Feedback loop mitigation: develop strategies to mitigate the feedback loop, where the model's previous outputs influence a retrain.
- Continuous training and updating: AI models may require retraining or fine-tuning to maintain their accuracy over time. Implement strategies for periodic updates and retraining of the model, especially in response to significant changes in input data or user

requirements. In case of an automated retrain pipeline, an expert must be involved in ensuring that the new model resolves the problem without underperforming or losing other expected characteristics.

Decommissioning an AI system

You may need to decommission or replace the system if it:

- Fails to meet its intended use
- Ceases to fulfil its initial purpose or if that is no longer relevant.
- Causes severe material or non-material harms

Stakeholder engagement

Communicate with stakeholders promptly if you detect many variances in input data and model performance to:

- Inform them about the changes in data and model performance.
- Discuss potential implications and necessary actions to address the issues.
- Collaborate on updating the model or adjusting the monitoring plan as needed.

Notify stakeholders if you retrain or update your model to:

- Provide transparency about the changes made to the model.
- Share the updated model's performance metrics and any improvements.
- Discuss any potential impact on the system or user experience.

GO FURTHER

- To know more about types of feedback loops that can happen in Machine Learning, check the following paper: [A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems](#).
- For simple implementation of data validation in Python, check [Pandera](#) or [Pydantic](#) libraries.

WORKBOOK | Enter your input

Complete Monitoring Plan for Input Data:

System used for monitoring

Date of evaluation

Monitoring Frequency [Daily/Weekly/Monthly/6 Months/Yearly]

Evaluation Type [Data Validation/Data Drift Detection/other]

Results [link to file or summarize the result]

WORKBOOK | Enter your input

Complete Monitoring Plan for Model Performance:

System used for monitoring

Date of evaluation

Monitoring Frequency [Daily/Weekly/Monthly/6 Months/Yearly]

Evaluation Type [Model Drift Detection/Feedback Loop Mitigation/Other]

Results [link to file or summarize the result]



High-risk EU AIAct Toolkit



UNIVERSITY OF
CAMBRIDGE



LEVERHULME CENTRE FOR THE
FUTURE OF INTELLIGENCE



Part of Accenture