Funded by
the European Union

Operational ethics guidelines on use cases related to human behaviour and cognition

AIOLIA DELIVERABLE 3.1

Horizon Europe Grant Agreement N° 101187937
AIOLIA PUBLIC

Funded by the European Union

Project Name	AIOLIA
Deliverable Title/Number	D3.1
Description	Operational ethics guidelines on use cases related to human behaviour and cognition
Lead beneficiary	CENTRIC
Lead Authors	Petra Saskia Bayerl, Erin Lawlor, Babak Akhgar
Contractual delivery date:	28 February 2026
Actual delivery date:	28 February 2026
Sensitivity	PUBLIC

Document History

Name	Organisation	Role	Action	Date
V1	CENTRIC	Lead	Draft sent to T3.1 partners for review and validation	20 November 2025
V1 for review	AUMC, AFLIANT, CEA, ETICAS, NIT, OXIPIT, THWS	Contributors	Review and validation	21.11 – 15.12. 2025
V2	CENTRIC	Lead	Version sent for quality review	03 February 2026
V2 for review	CEA, RISE	Reviewers	Review feedback sent	04-09 February 2026
V3	CENTRIC	Lead	Updated version sent for validation and review by use case partners	18 February 2026
V3 for review	AUMC, AFLIANT, CEA, ETICAS, NIT, OXIPIT, THWS	Contributors	Review and validation	18-24 February 2026
V4	CENTRIC	Lead	Updated version sent for final check	24 February 2026
V4 for review	RISE	Reviewer	Final review	26 February 2026
V4 for check	CEA	Coordinator	Final check	25-27 February 2026
Final version	CENTRIC	Lead	Ready for submission	28 February 2026

Funded by
the European Union

Configuration Management

Nature of Deliverable	
R	Document, report (excluding the periodic and final reports)

Dissemination level	
PU	Public, fully open

Acronym/abbreviations	
Artificial Intelligence	AI
Use Case	UC
European Union	EU

How to cite
Bayerl, P.S., Lawlor, E., Maris, M.T., Miorandi, D., Pekšys, G., Bjelica, M., Stojšin, K., Anastasova, M., Smith, O., Henestrosa, A., Yamshchikov, I., Bak, M.A.R., & Akhgar, B. (2026). Operational ethics guidelines on use cases related to human behaviour and cognition. AIOLIA Deliverable D3.1.

Acknowledgements

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein. Reproduction is authorised provided the source is acknowledged.

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Use of AI

AI systems have been used during proofreading to locate spelling, formatting, or grammar issues. No content of this deliverable has been generated by an AI system.

EXECUTIVE SUMMARY

Deliverable D3.1 is the result of task *T3.1: Co-create operational guidelines on industrial use cases*. The aim of T3.1 was to "co-create concrete guidelines to implement the ethics principles and values in three selected use cases" (cf. AIOLIA DoA). This was done "via matching one academic partner with one or more industrial partners". D3.1 is the result of this process, whereby we chose to conduct the co-creation process on all six European use cases in AIOLIA (instead of the required minimum of three) ensuring a broad and varied basis for the operational guidance.

D3.1 presents the findings of the operationalisation process. Firstly, it offers important insights into how ethics principles appear and are defined in practical settings, specifically the interlinkages of high-level ethics principles and their practical components. These observations suggest that, for AI researchers and developers, the status and meaning of ethics principles and the exact nature of components may be much more fluid than in the abstract AI ethics frameworks.

The core of D3.1 is the synthesis of the technical and organisational measures identified for the six AIOLIA European use cases. The synthesis provides a concise overview of the technical and organisational measures to implement the ethics principles across use cases. This is complemented by information about the potential challenges organisations may encounter when implementing the practical measures, as well as resources required for successful implementation. The co-creation process further yielded guidance (eight recommendations) on how to ensure a meaningful operationalisation of ethics principles for AI in human cognition and behaviour. Lastly, D3.1 describes an updated operationalisation pathway based on lessons learned during the co-creation process, which streamlines and simplifies the original methodology proposed in D2.3. The appendices within D3.1 present the full information provided by use cases in terms of definitions of ethics principles and components, as well as the technical and organisational measures.

Results at a glance

Purpose of D3.1

The aim of D3.1 is to offer concrete guidance on deploying AI in research and industrial settings, by presenting a collection of diverse practical measures.

Core findings

The benefits of successful ethics operationalisations: risk management and organisational gains

In the co-creation process with the six European use cases in AIOLIA, it became clear that industrial partners view the **operationalisation of ethics as a vital strategy to manage risks**. Risk management is a strong motivating focus in applied contexts as it helps in assessing and managing known risks or potential future risks that may lead to inefficiencies or even legal consequences. The framing around risk management seemed more 'natural' and easier than focusing discussions exclusively on ethics. In applied contexts, conversations about ethics should thus make use of the more familiar framing around risks to establish common ground and motivations to address ethics. However, use cases also put forward a positive perspective on the operationalisation of ethics, focusing on the **concrete organisational and societal benefits from the successful implementation of AI ethics**, such as the recognition as an ethics-aware organisation, legitimacy and even business opportunities. The systematic and early operationalisation of ethics in the form of practical technical and organisational measures may thus support organisations' risk management to prevent negative outcomes, as well as the achievement of positive outcomes.

Status of ethics principles and components

Across the six European use cases, the co-creation resulted in 12 contextualized ethics principles, broken into 37 components which covered 30 unique aspects. Generally, the **ethics principles in use cases align well with the overarching ALTAI framework but also show important patterns and deviations** (see Table below). The direct comparison of ALTAI with the UC-specific ethics principles demonstrates that the bottom-up process conducted in AIOLIA foregrounds very similar ethics concerns as established frameworks. The high-level principled frameworks such as ALTAI thus seem well-reflected in practitioner discussions about AI ethics. However, we also observed that across the diverse contexts, different focus and emphasis was given to either overall ethics concerns (e.g., human oversight) or specific sub-aspects (e.g., auditability, deskilling, safety). Moreover, we found that the same ethics consideration could be seen as either an overarching principle or as an aspect of other ethics principles (e.g., transparency as part of non-bias), and that the same component could be linked to different ethics principles (e.g., accuracy appeared as component in human oversight and non-maleficence, while auditability emerged as a component in robustness/reliability, non-maleficence, and accountability). Together, these observations illustrate that in AI research and design perspectives the status of principles and the exact relations between the components may be much more fluid than in more abstract AI ethics frameworks.

Comparison of ethics principles in the use cases and ALTAI (replication of Tables 5 and 8 in the report)

UC	Ethics principles in use cases	ALTAI Principles (Requirements)	
UC-principles covered in ALTAI			
UC2	Robustness and reliability	Req #2	Technical robustness and safety
UC5	Privacy and data protection	Req #3	Privacy and data governance
UC1, UC3	Transparency and explainability	Req #4	Transparency
UC3, UC4	Non-bias, fairness and non-discrimination	Req #5	Diversity, non-discrimination and fairness
UC1, UC4, UC6	Accountability and responsibility	Req #7	Accountability
UC-principles address sub-parts of ALTAI principles: *UCs each consider disparate aspects of Req #1 or sub-aspects as distinct ethics principle*			
UC2	Human oversight	Req #1	Human agency and oversight
UC5, UC6	Autonomy/User agency	Req #1	Human agency and oversight
UC2, UC3	Over-reliance and deskilling	Req #1	Human agency and oversight
UC-principles named in different ways but addressing aspects similar to ALTAI			
UC1, UC6	Non-maleficence ***Focus:*** *Covers important aspects within General Safety*	Req #2	Technical robustness and safety
UC4	Freedom of expression and non-censorship ***Focus:*** *Covers important aspects of Oversight*	Req #1	Human agency and oversight
UC-principles named in similar ways but addressing aspects different from ALTAI			
UC5	Safety/Human safety ***Difference:*** *Addresses primarily safety of users rather than safety of AI system*	Req #2	Technical robustness and safety
UC5	Human well-being ***Difference:*** *Addresses individual well-being, rather than broader societal or environmental issues*	Req #6	Environmental and societal well-being

Ethical challenges and tensions

The bottom-up process revealed core challenges and tensions for the practical implementation of ethics. One example includes possible tensions between responsibility, transparency and accountability. Responsibilities for activities, quality assurance, etc., are often distributed across different functions and roles within an organisation. As formulated in UC1 in the context of the professional behaviour in medical settings, this "creates a tension between the demand for traceability and the risk that accountability becomes either so widely shared that no one is clearly responsible or concentrated on individual clinicians, who work with systems they did not design and cannot fully control or understand."

Deskilling as part of professional behaviour was named as another challenge, in that AI tools may lead individuals to become "passive validators" (UC2) of AI decisions, removing professional scrutiny. This was seen as an issue specifically for less experienced individuals. Longer term, the usage of AI may lead organisations to remove more qualified personnel in favour of less experienced or "non-specialized personnel who may lack the domain expertise to detect subtle AI errors" (UC2).

UC3 reflected on a preference for technical versus organisational measures, amongst other tensions. Framed as a "stakeholder challenge", it underscores that different stakeholders "will lean towards those measures with which they are most familiar. For example, data scientists will likely reach for technical solutions and lawyers governance solutions." This creates a need to carefully consider who is involved and shapes the operationalisation process to avoid potential biases/gaps in consideration of measures.

The above are selected examples of the various observations during the operationalisation and validation discussions. While technically collected in T3.1, the observations themselves will be reported on and further developed at a non-technical level in D3.3. They will also feed into the training development within WP4.

Practical measures

The bottom-up exercise elicited 175 practical measures across the six use cases, 101 of them technical measures and 74 organisational measures. Most of them were linked to 'Non-bias, fairness and non-discrimination', followed by 'Autonomy/user agency' and 'Accountability/responsibility'. All ethics principles were represented through both technical and organisational measures, with the only exception of Human oversight which was only represented through technical measures.

The synthesis of practical measures for each ethics principle allowed us to identify measures that may be applicable across use case contexts, and those that are specific to an AI deployment setting or AI capability (e.g., AI use in a medical vs HR context). Measures identified across several use cases were, for instance, *oversight mechanisms* such as the creation of logs; documentation with defined oversight roles to keep humans in the loop; human validation for high-impact decisions or continuous performance validation for traceability; *access controls* that regulate which data is accessible or which individuals may have access; and *knowledge* pertaining to the AI system or area of AI application, as well as training to ensure knowledge is available in the organisation.

Use case specific measures addressed requirements or challenges within a domain, such as the need for competitor analyses of AI systems in case of commercial applications (UC5), technical requirements for classified data in AI training or deployment in security contexts (UC4), or specific knowledge of regulations (e.g., regulations for patient data, UC1).

Guidance for a successful operationalisation of AI ethics

Based on the reflections and insights provided by use case partners, eight recommendations were formulated to support the successful operationalisation of AI ethics principles in the context of human cognition and behaviour.

- ***Recommendation 1: Plan for a collaborative multi-disciplinary and durable process:*** Successful operationalisations need to be based on strong, long-term collaboration, using the respective strengths of partners (practitioner and academic) in the process. The design process further relies on continuous communication and feedback loops between research and practice-oriented participants ensuring that the theoretical and practical perspectives aligned. This also ensures that perspectives can be reviewed and validated or refined over time.
- ***Recommendation 2: Right people in the process:*** A common challenge in the operationalisation process is that it only includes a limited number of experts, such as legal or

technical experts in the organisation. A broad inclusion of actual or potential end-users with varying levels of expertise throughout the process ensures end-user concerns such as usability, interpretability or trustworthiness of the system can be taken on board.

- ***Recommendation 3: Practicality and accessibility as basis for formulation of measures:*** Practical measures are only useful if they are feasible within the specific industrial and operational context in which the AI is deployed and if the individuals tasked with their implementation understand them and understand their rationale. Therefore, practicality and accessibility should be at the forefront in the formulation of measures, keeping the core audience(s) in mind that need to implement, assess and/or audit them.
- ***Recommendation 4: Operationalisations for early-stage AI systems:*** Operationalisations may be conducted when AI capabilities or AI systems still are in an early, evolving stage. In this case, some aspects may only be assessed as 'not yet in place' or 'planned' rather than directly observable in practice (e.g., dashboards, KPIs, user behaviours) and thus based on anticipated risks and safeguards, not yet on real deployment experience. This does not invalidate the operationalisation process, but where operationalisations address early features, they should clearly be marked as 'preliminary' or 'forward-looking' and will need to be revisited once the system is fully in place.
- ***Recommendation 5: Considerations for business sensitive and classified contexts:*** Successful operationalisation of AI ethics tends to require a wide range of fine details about the AI capabilities, data, deployment, and outcomes, etc., as well as the business processes and users involved in their deployment. In industrial settings, such information is often sensitive; in security settings it may also be classified and bound by legal restrictions. It is therefore important to understand whether any special requirements apply for an AI deployment and from there consider the right choice of people, process, data handling, etc.
- ***Recommendation 6: Keep an open communication stance:*** Conversations about ethics can raise concerns in industrial partners that the process is primarily about finding gaps and faults. This presents the risk that ethics conversations might become defensive, especially from end-users. An open, practice- and solution-based communication stance will help support constructive engagements.
- ***Recommendation 7: Safeguard transparency and auditability of the process:*** To ensure that decisions along the process can be revisited, good documentation is important. This also includes potential challenges and disagreements during discussions. Overall, documentation not only of the operationalisation outcomes but also of the process supports transparency, auditability and accountability.
- ***Recommendation 8: Ensure sustainability:*** AI systems or their usage may change over time, e.g., through modifications/updates. Equally, conditions under which AI systems operate may change due to updates to business practices or new regulation. To ensure practical measures remain relevant, reviews either at regular intervals or at specific timepoints should be planned in for the duration of the AI deployment.

Simplified operationalisation pathway

Based on the lessons learned in the co-creation process, we updated the operationalisation pathway originally developed by D2.3 into a streamlined and simplified version. We propose a 4-step process to facilitate the operationalisation of AI-related ethics in applied settings. This updated process is context-independent, in the same way as the original operationalisation pathway, and as such applicable across contexts, AI systems, features and AI application areas.

Overview of the proposed 4-step process for ethics operationalisation (replication of Table 36 in the report)

Steps
1. Identification of the relevant ethics principles
Which ethics principles are relevant for your context and AI deployment?
EXAMPLE: Identified principle: Transparency (source considered for identification: ALTAI)
2. Identification of components for each ethics principle
What are the components that together describe the ethics principles?
EXAMPLE: Components identified to describe Transparency: Traceability, Explainability, Communication
3. Creation of practical measures per component
How to achieve the successful implementation of each component?
EXAMPLE: Practical measures identified to implement Traceability: Logging mechanisms, data labelling, documentation
4. Validation
Are the ethics principles, components and practical measures complete and described correctly?
EXAMPLE: A review identifies that further measures are required that focus on organisational practices to fully ensure Traceability, e.g., establishment of a governance and oversight mechanism that will guide the procurement and deployment of the AI system.

CONTENTS

List of Tables

List of Figures

1. Introduction

Deliverable D3.1 is the result of task *T3.1: Co-create operational guidelines on industrial use cases*. The aim of T3.1 was to "co-create concrete guidelines to implement the ethics principles and values in three selected use cases […] via matching one academic partner with one or more industrial partners" (cf. AIOLIA DoA). D3.1 is the outcome of this co-creation process. Yet, in contrast to focusing on three use cases as indicated in the DoA, we were able to conduct the co-creation process with all six European use cases in AIOLIA. This allowed us to create the operational guidance presented in this report on a broader and more varied set of information and contexts.

In the following sections, we provide the rationale and background that directed our work towards the co-creation of the operational guidance (Section 1.1), as well as the purpose and ambitions of this deliverable (Section 1.2).

1.1. THE RATIONALE FOR OPERATIONAL GUIDANCE

The overarching mission of AIOLIA is to address emergent ethical concerns linked to the changes that artificial intelligence (AI) introduces into both human cognition and behaviour. The emphasis is on short- or medium-term tangible shifts as opposed to hypothetical long-term scenarios.

The impact of AI on human cognition and behaviour is well established, both with positive effects supporting human cognition and behaviour and with more problematic impacts (Table 1 for examples). These varied consequences of AI require careful balancing and thus guardrails how to design and deploy AI to maximise benefits and at the same time minimise risks and negative impacts.

On the problematic side, AI mechanisms can shape human cognition and behaviours without committing explicit coercion, which raises concerns about user autonomy, manipulation, and consent; especially as users are often unaware that or in which direction such behavioural steering occurs. Experimental evidence, for instance, highlights how the convenience of AI may suppress critical thinking (Gerlich, 2025) and, combined with a tendency of humans to trust and follow AI advice (automation bias; Araujo et al., 2020) may shape individuals' decision-making through biases, fallacies or simple errors. Decision-making in the presence of AI may further be affected by the offloading of responsibility to the AI system (e.g., Bleher & Braun, 2022; Chan et al., 2020) potentially moving individuals towards more unethical, selfish behaviours (e.g., Krügel et al., 2023). AI has also been linked to shifts in social behaviours and societal norms, for instance, when recommendation algorithms create 'bubbles' of homogeneous content and communities, which over time may encourage social fragmentation, polarisation, and radicalisation (Mondal et al., 2025).

At the same time, AI has obvious benefits for many application areas, communities and industries. It can automate mundane, repetitive tasks in planning, maintenance or quality control which can increase efficiencies while reducing mental or physical burdens and fatigue. Walmart, for instance, uses AI logistics for optimising their driver routes and Ocado for automated picking and packing of parcels.[1] AI is moreover able to identify patterns in complex datasets that humans may find challenging to discover, leading to meaningful innovations and faster discoveries in various fields such as medicine

[1] https://intellias.com/ai-in-supply-chain/

and engineering. Beneficial are also AI applications that can shield humans from the effects of negative content, such as AI that pre-filter imagery and may thus help protect journalists against stress and trauma from continued exposure to violent materials (Sarridis et al., 2025). Even deepfake capabilities can be harnessed for positive uses, e.g., when AI is deployed to create believable fake content to support deaf communities with sign language deepfakes where interpreters may be unavailable or too expensive (Naeem et al., 2025).

Table 1: Examples of beneficial and detrimental AI impacts on human cognition and behaviour (cf. D2.1[2])

Domain	Examples of beneficial impacts	Examples of detrimental impacts
Decision-making	Decision support in highly complex situations or highly complex data	Automation bias, anchoring, over-reliance, reduced scrutiny
Attention	Removing tedious, repetitive tasks allow more time for meaningful/concentrated attention	Information narrowing, passive consumption, addiction, emotional manipulation
Responsibility	Increased accuracy in specialised fields; increased safety for humans in dangerous environments	Moral distancing, accountability diffusion
Trust	AI reviewing massive datasets reducing potential of missed information increasing trust in subsequent decisions	Mis-calibrated confidence or mistrust/insecurity, projections of competence

In response to the varied – beneficial as well as detrimental – effects of AI, a wide range of AI ethics frameworks have been developed with the aim to provide structured guidelines to ensure AI is developed and used responsibly (e.g., AI HLEG, 2020; OECD, UNESCO, 2023). While no single universal standard exists, most frameworks rely on a similar set of core ethics principles such as transparency/explainability, reliability and trustworthiness, fairness/non-discrimination, autonomy, inclusion and pluralism, non-maleficence, accountability and privacy (Corrêa et al., 2023).

Principle-based approaches have the advantage that they are more flexible and adaptive to context than prescriptive rules. High-level principles can thus provide a useful foundation for shared approaches to ethics, while remaining adaptive across a wide variety of contexts.

At the same time, high-level principles are too abstract to guide the concrete implementation of AI ethics in a specific context (health, education, energy, etc.) or in the light of specific AI capabilities (e.g., facial recognition vs deepfake creation). Another limitation of high-level principles is that they are difficult to evaluate. An assessment of adherence to principles such as 'fairness' or 'transparency' is meaningful only when they are translated into measurable indicators. Therefore, while AI ethics principles are a useful basis, guidance is needed on how to accomplish them in practice.

1.2. PURPOSE AND AMBITION OF D3.1

The aim of D3.1 is to offer concrete guidance to organisations that aim to deploy AI or are already deploying it in industrial settings, by presenting a collection of diverse practical measures. Practical measures refer here "to measurable features, dimensions or attributes related to the chosen ethical

[2] Grinbaum, A., Adomaitis, L., and AIOLIA consortium (May 2025). AIOLIA D2.1: Report on selected AI research areas and use cases. https://cea.hal.science/cea-05091448/

principle relevant in the design or deployment of an AI model or capability" (see D2.3[3], p. 25) and may be either focused on technical aspects or organisational aspects.

The portfolio of practical measures reported in D3.1 stems from the bottom-up co-creation process followed in AIOLIA within its ten use cases (see D2.3 for details on the process, D2.2[4] on the ten use cases). The co-creation process or 'Operationalisation Pathway' started with the high-level abstract AI ethics principles identified in D2.2, which were then translated into components and from there into concrete technical and organisational measures ('low-level requirements'; Mittelstadt, 2019) for the implementation of these ethics principles (cf. **Section 3** for methodology). This pathway ensured that the practical measures are grounded in the specific realities of AI deployments. D3.1 reports on the results from the six European use cases in AIOLIA.[5]

The co-creation approach in producing this Guideline led to important insights into how ethics principles appear in practical settings. The most important observation is how interlinked ethics principles and their components often are, when they come into contact with the reality of concrete AI instantiations and deployments (e.g., radiologists making judgements about a cancer patient supported by AI imaging in UC1 or quality engineers making judgements about vehicle safety in UC2). **Section 4** describes our findings on the status of higher-level principles versus their components and their frequent overlaps and intersections, which builds the foundation for the presentation of the practical measures themselves.

The core of D3.1 is the synthesis of the technical and organisational measures identified by the six AIOLIA use cases in the EU, presented in **Section 5.1**. This synthesis of the technical and organisational measures showcases concrete ways to implement an ethics principle, either across use case contexts or for a specific context. As part of the synthesis, we also present the very practical steps engineers and organisations need to take to ensure ethics principles are implemented effectively and efficiently. Next to the synthesis, the full set of technical and organisational measures has been included in **Appendix D**, offering a full portfolio of the practical measures to ensure that the rich and detailed information provided by use case partners remains available.

The extensive work done by the use cases going through the operationalisation pathway also led to insights about the factors they deem relevant for the successful implementation of these measures, specifically resources required, how to assess their successful execution, possible challenges for the implementation and risks if not implemented. These are presented in **Sections 5.2 and 5.3** to allow reflections on the practical issues organisations may encounter or have to consider when aiming to implement AI ethically and responsibly.

The co-creation further offered lessons about the operationalisation process itself, allowing us to update and simplify the way operationalisation of ethics principles can be conducted, compared to the original approach proposed in D2.3. The updated operationalisation pathway is presented in **Section 7**, which also synthesises lessons from use cases on what to look out for and consider at each step. We include this updated approach in D3.1 as we consider it an important result of the work done for this report, reflecting learnings stemming directly from our bottom-up approach. The materials in D3.1 are a solid foundation for the development of subsequent AIOLIA guidance (D3.3: Context-enriched

[3] Shiji, A.N., Bayerl, P.S., & Akhgar, B. (2025). AIOLIA D2.3: Practical Handbook for the Co-Creation Process.
[4] Teo, S.A., Kyosovska, N. and Armengol, A. (June 2025). AIOLIA D2.2: Report on the selection of ethical principles and values.
[5] Findings from the remaining use cases outside Europe are reported in AIOLIA Deliverable D3.2.

operational guidelines for AI research areas) and the development of training materials in AIOLIA (WP4).

However, in itself the portfolio of practical measures also aims to support organisations and end-users that design or deploy AI to speed up the identification of possible measures to ensure their responsible design or usage, as well as to inform them about possible broader considerations when planning to implement these practical measures such as resources required, possible challenges or even relevant regulations.

In the co-creation process with the six use cases, it became clear that industrial partners view ethics strongly as an opportunity to manage risks. Thus, our Operational Guideline aims to support industrial partners in assessing and managing known risks that tend to arise in the context of ethics considerations and requirements such as transparency, accountability or oversight. In our discussions, use cases further put forward a positive perspective on the operationalisation of ethics, focusing on the concrete organisational and societal benefits that the successful implementation of AI ethics can bring, such as the recognition as an ethics-aware organisation, legitimacy and even business opportunities. Our operationalisation of ethics in AIOLIA, and the practical technical and organisational measures, may thus support organisation's risk-management, as well as the identification of benefits.

1.3. WHAT THIS GUIDANCE IS AND IS NOT

What the Operational Ethics Guidelines provide is a detailed view of the potential practical technical and organisational measures that support the implementation of specific ethics principles. This is accompanied by reflections on the concerns, resource requirements and overarching considerations, which were developed in a bottom-up process with our academic and industry partners in six use cases.

In the same vein, our Operational Ethics Guidelines are **not a self-assessment tool, nor a 'tick-list' or 'checklist'** against which to assess whether all possible steps to implement ethics have been taken; especially, as AI technology develops and the concrete practical measures mentioned in our Guidance may become or have to be supplanted by others.

The Operational Ethics Guidelines are thus exactly that: context-bound considerations to implement AI ethically and responsibly in practical settings. **While it is comprehensive, the Guidelines do not claim to be complete or cover all possible or relevant components of the AI ethics principles or all possible practical measures**. Implementing all measures and recommendations in the report does thus not imply that 'ethics is solved.'

Therefore, the Guidance should be considered rather as a portfolio of possible practical measures, which can be used to lead organisations, AI developers and AI users to formulate their own set of technical and organisational measures. Each AI deployment is different; hence, each situation needs to be assessed on its own merit and in its own context, and the current guidance offers inspiration as well as a co-creation approach to conduct this work.

2. Background and approach

2.1. AI ETHICS PRINCIPLES GUIDING D3.1 WORK

The work in D3.1 was guided by the ethics principles identified in *D2.2: Report on the selection of ethical principles and values*, which presented the "selection of ethical principles and values in relation to the use cases and research areas in the project" (D2.2, p. 3).

The ethics principles identified in D2.2 were guided by a review of relevant literature and ethics frameworks such as the EU AI Act, UNESCO's Recommendations on the Ethics of AI, the OECD AI principles and the Assessment List for Trustworthy AI (ALTAI; High Level Expert Group on AI, 2020), and supplemented by an empirical process with AIOLIA use case partners and external experts to identify the three most important ethics principles and values per use case (for details about the process, see D2.2). Table 2 presents the use cases and their linked ethics principles and values as identified in D2.2 for the six European use cases.

Table 2: Use cases and ethics principles as identified in D2.2 (based on Table 1, D2.2, p. 24)

Link to human behaviour and cognition	Use case description	AI research areas	Selected ethical principles
Change in human expertise and professional behaviour	UC1: Medical doctors using AI tools in diagnostics and treatment	Decision-support systems; Image recognition	1. Non-maleficence 2. Accountability and responsibility 3. Transparency and explainability
	UC2: Safety engineers using AI tools to speed up software release approvals	Decision-support systems; General-purpose AI	1. Robustness, safety and reliability 2. Oversight and autonomy 3. Risk of over-reliance and deskilling
	UC3: Recruiters using AI tools in hiring processes	Decision-support systems	1. Non-bias, fairness and non-discrimination 2. Transparency and explainability 3. Over-reliance and deskilling
	UC4: Security professionals using AI tools to detect hate speech	Decision-support systems; General-purpose AI	1. Freedom of expression and non-censorship 2. Non-bias, fairness and non-discrimination 3. Accountability and responsibility

Table 2: Use cases and ethics principles as identified in D2.2 (based on Table 1, D2.2, p. 24) (continued)

Link to human behaviour and cognition	Use case description	AI research areas	Selected ethical principles
Change in human cognition and private behaviour	UC5: AI systems as individual and family-level virtual assistants	Conversational general-purpose AI; Emotional AI	1. Non-manipulation and non-maleficence 2. Privacy, consent and data protection
	UC6: Deepfake therapy for processing trauma and grief	Multi-modal general-purpose AI (GPAI); Emotional AI	1. Non-maleficence 2. Autonomy and non-manipulation 3. Risk of over-reliance

While ethics principles are useful in focusing efforts for responsible AI deployments, they are difficult to implement. An important effort in the EU to support the implementation of ethics principles is ALTAI (Assessment List for Trustworthy Artificial Intelligence; AI HLEG, 2020). ALTAI offers a compliance checklist, based on a set of seven broadly agreed AI ethics principles in the EU, which are referred to as 'key requirements.' The set of seven ethics principles is:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Environmental and societal well-being
7. Accountability

ALTAI aims to translate these high-level ethics principles into an actionable, technical, and organisational checklist. For this aim, each key requirement is broken down into smaller 'elements' or 'issues' that should be addressed; for instance, *Transparency* is broken down into three elements, "1) traceability, 2) explainability and 3) open communication about the limitations of the AI system" (AI HLEG, 2020, p. 14). Our co-creation process follows a similar logic, where we refer to the constituting 'elements' of an ethics principle as 'components' (see D2.3 and Section 3 below).

ALTAI then puts forward a detailed list of questions that together constitute a detailed compliance checklist against which to assess a specific AI deployment (e.g., for the Traceability aspect of Transparency: "*Did you put in place measures to continuously assess the quality of the input data to the AI system?*"; ibid.).

Questions of this kind are helpful for guiding organisations and end-users in thinking about AI ethics in a comprehensive manner. However, a challenge arises for practitioners: namely, what these 'measures' can or should look like, which measures are sufficient or acceptable, and how to judge whether these are successful in safeguarding the key requirements of AI ethics? All these are issues of operationalisation, i.e., the ability to translate principles into concrete actions in specific contexts.

2.2. OPERATIONALISATION AS CORE APPROACH

The approach underlying this guidance is one of applied ethics with the ambition to guide the day-to-day execution of the AI ethics work in an organisation. In contrast, to high-level principles or requirements such as formulated in ALTAI, the Operational Guidance in AIOLIA aims to support the practical application and implementation of AI ethics requirements in granular detail.

The operationalisation of AI ethics in the six European AIOLIA use cases, on which the current guidance is built, thus moves from the higher-level principles identified in D2.2 to the concrete technical and organisational measures stakeholders should take to achieve AI ethics in practice.

> Operationalisation in this context "refers to the process of translating high level ethical principles into practical actions, tools, processes and governance structures that can guide and be applied throughout the lifecycle of AI systems to ensure ethical design, development, deployment and use." (see D2.3, p. 11)

The practical focus includes both technical and organisational measures to ensure that the operationalisation covers the technical/design features of AI, as well as context-specific human aspects that need to be addressed by decision-makers in the organisations that design, procure and deploy the AI systems.

As defined in Handbook for the Operationalisation of Ethics (D2.3), technical and organisational measures are understood as the following (cf. D2.3, p. 25):

Technical Measures: Technical methods focus on the design and technical aspects of AI systems and refer to specific tools, methodologies, technologies and processes that are implemented in AI systems to ensure it operates in an ethical manner. Technical measures to foster ethical AI practices include the addition of privacy-by-design approaches, Explainable AI (XAI) measures, use of benchmarks and key performance indicators, adversarial testing, federated learning, data anonymization techniques and security audits.

Core audience: engineers, AI developers

Organisational Measures: Organisational measures focus on how an organisation incorporates and manages ethical AI practices by referring to the structures, policies and governance framework in place. Organisational measures for ethical AI governance include the development of AI ethics boards, ethical AI policies, promoting community stakeholder engagement, fostering interdisciplinary collaboration, regulatory and legal compliance to existing regulations, ethics readiness indicators and the development of AI risk frameworks.

Core audience: management, training and HR departments

Both types of practical measures are required to implement AI ethics comprehensively and effectively. As indicated above, the core audiences responsible for their implementation differ, indicating that AI ethics requires close, multi-disciplinary collaboration to succeed.

3. Methodology for the Development of the Guidelines

3.1. GENERAL APPROACH AND UPDATES TO USE CASES

The Operational Ethics Guidance is the result of a structured, bottom-up co-creation process that engaged with cross-disciplinary stakeholders. Using the Operationalisation Pathway (developed and described in D2.3), the process followed principles of co-creation to ensure a collaborative and participatory process with stakeholders from academia and industry within each use case.

The operationalisation process was conducted with six use cases (UCs), all addressing disparate aspects of AI in human cognition and behaviour, allowing a broad view of possible ethical AI requirements and challenges.

During work in T3.1, three use cases changed their focus:

- UC3 changed its focus from HR recruitment to assessing phishing vulnerability at the workplace;
- UC4 broadened its focus from the detection of only hate speech to the detection of generally harmful or illegal content;
- UC5 removed the topic of family-level assistants by focusing on individual virtual assistants.

Further, UC5 and UC6 made changes to the ethics principles originally identified in D2.2.

Table 3 provides longer descriptions of the final use cases, while Table 4 provides an updated view of the use cases and their respective ethics principles, as they were included in D3.1 (changes compared to D2.2 marked in blue).

Table 3: Use case descriptions (shortened from the text provided by UC partners)

Use case descriptions (shortened)
UC1: Medical doctors using AI tools in diagnostics and treatment (AUMC, Oxipit, Afliant)
UC1 comprises two complementary healthcare scenarios. Both solutions apply medical imaging AI to support clinician decision-making: the first in chest X-ray interpretation and reporting workflows, the second in CT-based planning, intra-operative guidance, and follow-up for abdominal aortic aneurysm repair. **The chest x-ray suite looks for 75 most common radiological findings and identifies high confidence normal chest x-rays**. This allows for three applications: (1) automation of 40% of chest x-ray reports; (2) speeding up the radiological workflow through computer-assisted diagnosis (CAD) functionality; (3) screening the final radiological reports for potential misalignment with AI image interpretations. **The second scenario, on the other hand, covers the treatment of abdominal aortic aneurysms (AAAs) at three key decision points: pre-operative planning, intra-operative execution, and post- operative follow-up**. This integrated decision-support system is designed to improve professional behaviour of medical doctors by reducing planning time, enhancing procedural accuracy, and rationalising patient monitoring. Rather than replacing surgeons, the AI tool offers real-time, context-aware assistance that can help reduce cognitive load and improve treatment quality. The technology is also integrated with a surgical simulator (ANGIO Mentor) which provides professional training for trainee surgeons.
UC2: Safety engineers using AI tools to speed up software release approvals (NIT, CEA)
NIT is conducting a series of **projects for the automotive industry, in which they are tasked to use AI tools in conducting safety analyses for the automotive products**. Many automotive products include hardware and algorithms for autonomous and assisted driving. Cutting corners to release fast and frequently, with software patches and upgrades, is becoming a growing practice. Automation of safety analyses would help improve quality of the releases, since the stage gates related to safety in the current DevOps (CI/CD) cycle include manual checks which may last for weeks. Automated tools include the addition of templates, databases of scenarios and failure modes, and more recently, AI suggestions. **NIT is currently developing an AI decision-support tool to speed up and facilitate the System Theoretic Process Analysis (STPA) analysis which assesses potential hazards and safety risks**. This tool would help the safety engineer by giving initial suggestions on how to conduct the analysis; kick-off the analysis based on the general description of the system architecture, by providing the initial list of control actions and failure modes and help assess the analysis which is nearly completed, to identify the remaining gaps and suggest fixes; modifying the usual professional behaviour significantly. NIT is beginning to use this tool in preliminary studies with an independent parallel manual process. As the tool should become a part of each STPA analysis, it is crucial that the end quality and safety at the output are preserved.
UC3: HR professionals using AI tools to assess and reduce employee vulnerability to cyberattacks (Eticas)
UC3 focuses on an AI-based cybersecurity platform developed by a party outside the consortium. The **goal of the AI system is to assess and reduce human cognitive and behavioural vulnerability to phishing attacks**. These are attacks in which an attacker attempts to masquerade as someone else, for example a colleague or vendor of the company that the potential victim works for. Such attacks rely on psychological biases, for example: authority bias; familiarity bias; or urgency bias. To measure vulnerability to such attacks the AI system combines behavioural analytics, psychometric profiling (Behavioural Inhibition System (BIS) / Behavioural Activation System (BAS) / Need for Cognition (NC) scales), contextual risk factors (CR) such as role-based exposure to external email, typical device usage (e.g. mobile vs desktop), and historical reporting of suspicious messages, and predictive modelling. Only contextual variables that are available before the simulated attack are used for training and scoring, while post-attack debrief responses (e.g. “I realised it was suspicious”) are used solely for feedback and awareness to avoid information leakage and inflated performance estimates. The system operates across multiple European organisations, simulating phishing scenarios to measure individual and organisational exposure, cognitive and behavioural impact, and recommending preventive actions.

UC4: Security professionals using AI tools to detect harmful or illegal content (CENTRIC)
This use case examines the **ethical challenges and principles associated with the development and deployment of AI tools used in security practices, specifically those designed for image, video, audio or text processing**. Such tools are increasingly employed to detect and analyse potentially harmful or illegal content, including hate speech, extremist narratives, and terrorist propaganda, across digital and media platforms. While these systems contribute to public safety and counter-extremism efforts, their operation raises significant ethical concerns related to freedom of expression, fairness, accountability and privacy. The use case explores how AI-based detection systems assist and support security professionals in making complex judgements about context, intent and harm, and how these judgements can be affected by bias in training data, lack of transparency and limited human oversight. It addresses questions such as how to ensure that content moderation decisions are proportionate, explainable and non-discriminatory, and how responsibility is distributed between automated systems and human operators. The data in this use case came from security AI developers, as well as an NGO who are not AIOLIA partners. It must be noted that a prevailing feature of UC4 discussions was the inability of security practitioners to share sensitive information.
UC5: AI systems as personalised characters and individual virtual assistants (THWS)
THWS and their partners focused on two distinct personal assistant scenarios studied in contact with an industrial company outside AIOLIA consortium. The first involves a **one-to-one virtual character geared toward human-AI role-play**, where the primary user interacts privately with the system first describing the character they want to interact with and then communicating with this digital character emulated by a generative language model. This application is already mature and has an established user base. The second scenario addresses a **single adult user model,** in which the AI's characteristics are pre-defined and fixed to optimise for safety and context comprehension at an individual level. The assistant helps the user with daily organisation, habit tracking, and personal development through conversational AI support. Thus, it assists in maintaining healthy habits including hydration, exercise, smoking cessation, sleep hygiene, focus techniques, and time management. Although both use cases necessitate the handling and storage of sensitive personal information, the application contexts differ and pose different requirements for safety and user protection. While the personalized character requires security measures to protect users from psychological and physical harm, for instance, the daily assistant focuses on staying within the boundaries of permissible behaviour modification. Nevertheless, both prioritise the ethical principles of privacy, safety, and human autonomy.
UC6: Deepfake therapy for processing trauma and grief (AUMC)
In the healthcare context, **deepfake technology can be applied within psychotherapy. AI-generated hyper-realistic video footage can be used to simulate conversations** with a realistic but fabricated representation of a deceased loved one (in grief counselling) or a perpetrator of trauma (e.g. in treatment of PTSD from sexual violence). The **aim is to unlock closure through emotional processing or confrontation** that would otherwise be impossible in traditional talk therapy, such as saying goodbye or confronting trauma. Other potential applications of deepfakes in healthcare are found in projects studying 'deepfaked' clinicians giving medical advice, e.g. to improve medication adherence. Early reports suggest potential therapeutic value and positive experiences in specific cases, but deepfake technology can potentially also reconfigure how a therapeutic interaction is staged and can pose numerous ethical challenges with regard to human cognition and behaviour.

Table 4: Overview of final use cases in D3.1, with changes from D2.2 marked in blue

Link to human behaviour and cognition	Use case description	Ethics principles
Change in human expertise and professional behaviour	UC1: Medical doctors using AI tools in diagnostics and treatment	1. Accountability and responsibility 2. Non-maleficence 3. Transparency and explainability
	UC2: Safety engineers using AI tools to speed up software release approvals	1. Robustness and reliability 2. Human oversight 3. Over-reliance and deskilling
	UC3: HR professionals using AI tools to assess and reduce employee vulnerability to cyberattacks	1. Transparency and explainability 2. Non-bias, fairness and non-discrimination 3. Over-reliance and deskilling
	UC4: Security professionals using AI tools to detect harmful or illegal content	1. Accountability and responsibility 2. Non-bias, fairness and non-discrimination 3. Freedom of expression and non-censorship
Change in human cognition and private behaviour	UC5: AI systems as personalised characters and individual virtual assistants	1. Safety/human safety 2. Privacy, consent and data protection 3. Autonomy (added as new principle) 4. Human well-being (added as new principle)
	UC6: Deepfake therapy for processing trauma and grief	1. Non-maleficence 2. Autonomy (non-manipulation removed) 3. Accountability and responsibility (replacing risk of over-reliance)

An important observation already at this stage is how the ethics principles across the six use cases relate to disparate aspects and levels of ethics principles and values in frameworks such as ALTAI. Table 5 provides a direct comparison of the ethics principles in the AIOLIA use cases and the principles in ALTAI. A more detailed discussion on ethics principles and components is given in Section 4.

Table 5: Comparison of bottom-up ethics principles in use cases versus ALTAI principles

UC	Ethics principles in use cases	ALTAI Principles (Requirements)	
UC-principles covered in ALTAI			
UC2	Robustness and reliability	Req #2	Technical robustness and safety
UC5	Privacy and data protection	Req #3	Privacy and data governance
UC1, UC3	Transparency and explainability	Req #4	Transparency
UC3, UC4	Non-bias, fairness and non-discrimination	Req #5	Diversity, non-discrimination and fairness
UC1, UC4, UC6	Accountability and responsibility	Req #7	Accountability
UC-principles address sub-parts of ALTAI principles: *UCs each consider disparate aspects of Req #1 or sub-aspects as distinct ethics principle*			
UC2	Human oversight	Req #1	Human agency and oversight
UC5, UC6	Autonomy/User agency	Req #1	Human agency and oversight
UC2, UC3	Over-reliance and deskilling	Req #1	Human agency and oversight
UC-principles named in different ways but addressing aspects similar to ALTAI			
UC1, UC6	Non-maleficence ***Focus:*** *Covers important aspects within General Safety*	Req #2	Technical robustness and safety
UC4	Freedom of expression and non-censorship ***Focus:*** *Covers important aspects of Oversight*	Req #1	Human agency and oversight
UC-principles named in similar ways but addressing aspects different from ALTAI			
UC5	Safety/Human safety ***Difference:*** *Addresses primarily safety of users rather than safety of AI system*	Req #2	Technical robustness and safety
UC5	Human well-being ***Difference:*** *Addresses individual well-being, rather than broader societal or environmental issues*	Req #6	Environmental and societal well-being

3.2. DATA COLLECTION PROCESS

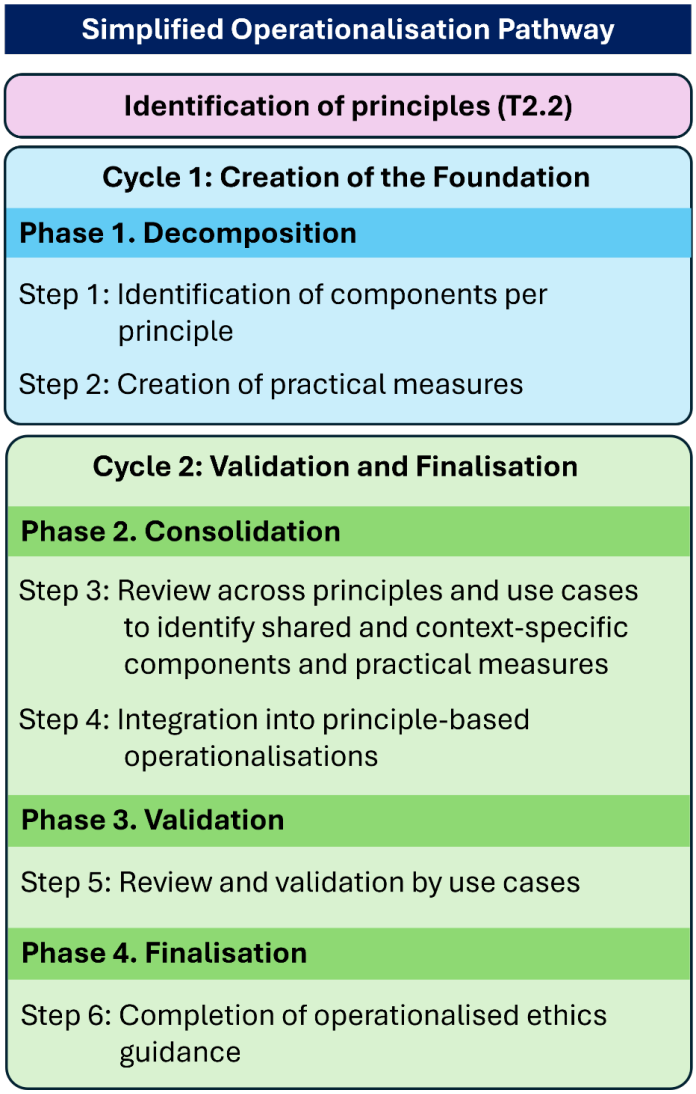
The data collection for the co-creation of the Guidance followed the Operationalisation Pathway defined in D2.3: Practical Handbook for the Co-creation Process. D2.3 outlined the concise steps required to accomplish the co-creation at the basis of this Operational Ethics Guidance, as well as templates to guide UC partners through the activities.

The operationalisation was conducted in two cycles:

- **Cycle 1 – Creation of Foundations** (July-October 2025): aimed at the identification of the components for each ethical principle in a use case and the practical measures (technical and organisational) to ensure they are achieved. The identification of components "ensures that all relevant aspects of a principle are captured and can subsequently be translated into practical measures" (D2.3, p. 17)
- **Cycle 2 – Validation and Finalisation** (November 2025 – January 2026): for review, cross-case analyses, validation, and finalisation of the guidance

Figure 1 provides an overview of the process followed (simplified from D2.3, p. 12).

Figure 1. Overview of operationalisation pathway (simplified from D2.3, p. 12)



3.2.1. Cycle 1 activities

Cycle 1 comprised two phases: decomposition (phase 1) and consolidation (phase 2). In this cycle UC partners identified the components of the ethics principles collected in D2.2 and translated them into practical measures.

Phase 1 - Decomposition of ethics principles: Decomposition refers to the translation of the chosen high-level ethical principles into context specific components and actionable practical measures. The decomposition process itself was broken down into two stages:

- **Stage 1: Identification of components:** The importance of the identification of components per principles lies in the ability to ensure all relevant aspects of the overarching principles are being considered, which will ensure effective translation into practical measures. The process relied on analyses of relevant academic literature, organisational policies, frameworks and experiences by industry partners. These activities aimed to build a shared understanding of how each industry/use case defines, assesses and applies the over-arching ethical principles.
- **Stage 2: Identification of the practical measures to implement ethics principles and components:** After establishing the components of the ethical principles, UC partners were asked to map out the practical technical and organisational measures to implement their ethical principles in practice.

Information collected in Phase 1

Information collected in Cycle 1 used standardised data collection templates (can be found in D2.3), to achieve a comprehensive description of components and technical and organisational measures.

For decomposition of the ethics principles (phase 1), we requested the following information for each ethics principles:

- **Components** – name of three or more component that comprise a specific ethic principle
- **Definition** – definition of the component
- **Relevance** – rationale for choosing this component in the given use case
- **Source(s)** – list sources that suggest the inclusion of this component
- **Disagreements** – describe any disagreements on the inclusion or meaning of this component

In the subsequent step ('Identification of the practical measures'), we collected technical and practical measures, in addition, to broader considerations of their implementations, challenges, risks, and requirements.

Per component, technical as well as organisational measures were requested. For each measure the following aspects were requested:

- **Description of the measure** – outlining its purpose and practical application within the use case context
- **Relevance** – explaining why the measure is critical to ensuring ethical, safe and reliable AI
- **Implementation** – describing how the measure can be achieved in practice, including supporting tools, standards and workflows
- **Assessment** – identifying how fulfilment of the measure can be verified or evaluated
- **Challenges** – describing potential barriers to the implementation of the measure

- **Risks** – the potential consequences of non-fulfilment
- **Specific requirements** – summarising the technical, procedural or regulatory conditions necessary for implementation

Taken together, these elements provided a solid foundation for the development of overarching, actionable guidance that can be adapted across sectors to strengthen accountability, safety and trust in AI-enabled systems. Table 6 lists the components each of the six use cases identified for their ethics principles.

Table 6: Components identified by use cases for each of their ethics principles

UC 1: Medical doctors		
Accountability and responsibility	**Non-maleficence**	**Transparency and explainability**
1. Auditability 2. Human oversight 3. Liability (responsibility)	1. Validity / accuracy (from technical perspective) 2. Bias 3. Privacy	1. Accessibility 2. Explainability 3. Justifiability

UC 2: Safety engineers		
Robustness, safety and reliability, with special focus on accountability and traceability of safety decisions [Robustness and reliability]	**Human oversight and autonomy, with additional focus controllability [Human oversight]**	**Risk of over-reliance and deskilling**
1. Technical robustness and resilience 2. Reliability through lifecycle testing and monitoring 3. Fairness in safety assurance	1. Human oversight and controllability 2. Accountability and traceability of safety decisions 3. Transparency of safety-critical performance and limits	1. Preservation of human skill and expertise 2. Feedback and learning loops for human adaptation 3. Organisational policies for shared responsibility

UC 3: HR professionals using AI to reduce cyberattack vulnerability		
Non-bias, fairness and non-discrimination	**Over-reliance and deskilling**	**Transparency and explainability**
1. Diversity 2. Representativeness / inclusivity 3. Objectivity 4. Non-stigmatising use / proportionality	1. Dependence 2. Contestability and human oversight	1. Openness 2. Accessibility / access to information 3. Documentation, traceability, and auditability

UC 4: Security professionals		
Freedom of expression and non-censorship	**Non-bias, fairness and non-discrimination**	**Accountability and responsibility**
1. Autonomy and agency 2. Proportionality 3. Non-discrimination	1. Equality and impartiality 2. Representation and inclusivity 3. Transparency of criteria	1. Human oversight 2. Auditability and evaluation 3. Responsiveness

UC 5: AI systems as personalised characters and individual virtual assistants			
Privacy and Data Protection	**Safety/Human Safety**	**Autonomy/User Agency**	**Human Well-being**
1. User consent and transparency 2. Data Minimisation, data use and storage 3. Third-party sharing and compliance	1. User protection 2. Security Measures 3. Human Oversight	1. Informed Consent 2. System Customisation 3. Transparency and User understanding	1. Promotion of User's health 2. Scope boundaries 3. Crisis Recognition

UC 6: Deepfake therapy for processing trauma and grief		
Non-maleficence	**Autonomy and non-manipulation [autonomy]**	**Accountability**
1. Subsidiarity and proportionality 2. Effectiveness 3. Societal well-being	1. Transparency 2. Privacy 3. Risk of over-attachment and dependency	1. Human agency and responsibility 2. Professional competence 3. Oversight

Considering practical measures, the bottom-up exercise elicited 175 practical measures across the six use cases, 101 of them technical measures and 74 organisational measures. Most of them were linked to Non-bias, fairness and non-discrimination, followed by Autonomy/user agency and Accountability/responsibility (see Table 7). As Table 7 shows, principles are represented through both technical and organisational measures, with the only exception of Human oversight, which is only represented through technical measures.

Table 7: Number of practical measures given by use cases within each ethics principle, ordered by number of total measures

	Technical measures	*Relevant UCs*	Organisational Measures	*Relevant UCs*	*Sum*
Non-bias, fairness and non-discrimination^	14	*UC3, UC4*	10	*UC3*	*24*
Autonomy	11	*UC5, UC6*	13	*UC5, UC6*	*24*
Accountability and responsibility	12	*UC1, UC4*	10	*UC4, UC6*	*22*
Robustness[6]/reliability	11	*UC2*	3	*UC2*	*14*
Transparency and explainability	11	*UC1, UC3*	6	*UC3*	*17*
Over-reliance and deskilling	9	*UC2, UC3*	9	*UC2, UC3*	*17*
Non-maleficence	8	*UC1*	6	*UC6*	*14*
Safety/human safety	6	*UC5*	6	*UC5*	*12*
Privacy, consent and data protection	5	*UC5*	5	*UC5*	*10*
Human oversight+	8	*UC2*	[no entries]		*8*
Human well-being	3	*UC5*	4	*UC5*	*7*
Freedom of expression, non-censorship	3	*UC4*	2	*UC4*	*6*
Sum	**101**		**74**		***175***

[6] While robustness is not treated as ethics principle in Horizon Europe, AIOLIA follows ALTAI in naming it as ethics principle to consider for use cases. Robustness also emerged as ethics concern in initial use case considerations (D2.2).

3.2.2 Cycle 2 activities

The objective of Cycle 2 was to review, validate, refine and finalise the guidelines through feedback loops with the academic and industry partners. It consisted of the three phases: Consolidation, Validation and Finalisation.

Phase 2 – Consolidation: In this phase, all information collected in Phase 1 was integrated into the draft guidance. The draft guidance focused on the review and integration of materials from the six industrial use cases and contained two parts: (1) presentation of the information from each use case 'as is' to allow review and sense-checking by UC partners and fill in missing information/details, (2) integration into a first synthesis of findings across use cases. The objective of the draft guidance was to ensure correct and complete information was collected before the finalisation of the Operational Ethics Guidance.

Phase 3 – Validation: Using the draft guidance as basis, validation comprised two interlinked activities:

- **Review of the draft guidance by UC partners** and, where needed, add corrections, refinements and details directly into the document
- Validation meetings with each UC led by CENTRIC to clarify opacities, discuss challenges in the implementation of measures, identify unique UC-specific versus transferable aspects (components, measures, risks, challenges, etc.), and obtain perspectives on the formulation of the final guidance. Part of the validation meetings was further used to include tensions in the implementation of ethics and training requirements, to support the further work in T3.3 (operational guidance for AI research areas) and WP4 (training), respectively.

For four of the six use cases, we were able to conduct validation meetings, for UC2 we conducted two validation sessions with the second a follow-up discussion on ethics challenges. (For UC4 and UC6, we were unable to conduct validations, as key personnel were on leave.) Dates for the validation sessions were as follows: UC1: 18. December 2025; UC2: 11. December 2025 + 12. January 2026; UC3: 05 January 2026; UC5: 17. December 2025.

Phase 4 – Finalisation: The feedback from both the validation efforts were incorporated into the final guidance, i.e., the current document.

3.3. METHODOLOGICAL REFLECTIONS

As part of the methodology, use case partners were also asked to provide a reflection on the co-creation process. The review of the methodological reflections provided by partners shows that use cases mostly followed a similar iterative, multi-step and evidence-informed processes to identify and validate their components. Despite differences in context and focus all methodological approaches shared the following key features:

- **Collaborative multi-discipline design process:** The development of components has been conducted collaboratively between the academic and industry (or technical partners). The design process relied on continuous communication and feedback loops between the research and practice-oriented participants ensuring that the theoretical and practical perspectives aligned
- **Iterative refinement and validation:** Each use case employed an iterative workflow: initial findings or proposals were refined through successive stages of review, discussion and

validation. Use cases then utilised multiple rounds of feedback leading to a consensus on the final components.

- **Structured documentation and transparency:** Use cases systematically documented their results which served as a basis for discussion, validation and traceability of decisions made throughout the process

Whilst following a similar process there were methodological differences in that most use cases were interview driven (UC1, UC3, UC4 and UC5), whereas others utilised workshops and hands-on orientation techniques to facilitate discussion and had a system and process level focus (AI-based safety analysis) and were validated through a broader operationally grounded approach (UC2) or based on review of previous empirical and published work (UC6).

With respect to the identification of practical measures, the analysis of the method reflections shows that identification was led by industry and academically validated. The reflections on this process share the following similarities:

- **Industry scaffolding drafting:** Industry partners generally produced the first concrete list of measures grounded in their workflow and experiences.
- **Iterative co-development:** The partners then used back-and-forth cycles to allow academics to review/refine what the industry partner produced, which ensured the measures fit AIOLIAs ethical/technical framework and current good practice.
- **Documented outputs for validation:** Finally, structured documents were produced which were shared, discussed and agreed upon.

4. Reflection on Status of Ethics Principles and Components

4.1. INTERDEPENDENCE OF ETHICS PRINCIPLES

The bottom-up process used in the operationalisation pathway resulted in 12 contextualized ethics principles (for definitions provided by use cases see Appendix B).

Generally, these bottom-up principles align well with the overarching ALTAI framework but also show important patterns and deviations (cf. Table 5; reproduced in Table 8 for ease of reading). Firstly, most ALTAI principles are represented. *Human Agency and Oversight* (requirement #1) emerged most prominently across four of the six UCs; however, phrased as part of disparate and separate UC principles. For instance, UC2 drew out 'human oversight' and 'over-reliance and deskilling', while UC5 and UC6 focused on 'autonomy/user agency'. This suggests, on the one side, the high importance of human agency and oversight for AI ethics in the context of human cognition and behaviour in our use cases. On the other side, use cases focused on specific sub-aspects of requirement #1 rather than the whole principle (e.g., UC2 on oversight and UC5 on autonomy). Additionally, requirement #7 *Accountability* was identified as core ethics principle in three of the six use cases, referred to as 'accountability and responsibility': UC1, UC4, and UC6. At the same time, UC2 referenced 'robustness and reliability', again indicating overlaps in the way use cases conceptualised ethics principles, here in terms of 'reliability' either as element of accountability or robustness.

Moreover, different aspects with ALTAI principles emerged as distinct principles. UC2, for instance, split requirement #1 into two separate principles: Oversight and Over-reliance. Similarly, requirement #2 *Technical Robustness and Safety* in ALTAI emerged as separate concerns amongst use cases: UC2 focusing on *Robustness/reliability* and UC5 on *Safety/human safety*. This also explains why the number of use case principles is higher than the number of key requirements listed in ALTAI (12 versus 7).

These observations suggest that use contexts influence which aspect are seen as most relevant for the practical concerns of AI deployments and implementations. Also, even if the same ethics principle is seen as relevant, they may focus on disparate aspects.

In the same vein, ethics principles may be named the same way by use cases and ALTAI but actually address different issues. UC5 specifically named 'Safety', which seems to align with ALTAI requirement #2 *Technical robustness and safety*. However, UC5 focuses on safety of individuals, whereas ALTAI focuses on the safety of the AI system. Similarly, well-being in UC5 refers to individual well-being, in contrast to the broader societal or environmental issues addressed in ALTAI in requirement #6 *Environmental and societal well-being*.

Two ethics principles identified by UCs are not part of ALTAI: *Freedom of* Expression *and Non-Censorship* referred to in UC4 is mentioned only as part of the Human Rights discussion, while *Non-maleficence* referred to in the health-related use cases UC1 and UC6 is a classic bioethics principle which is not (directly) mentioned in ALTAI. For specific AI deployment areas, therefore additional considerations may be significant that reach beyond ALTAI. The fact that human rights were not explicitly mentioned in all UCs does not imply that they are unimportant, but rather that they take a central role in a specific UC among the ones studied in AIOLIA.

Table 8: Comparison of bottom-up ethics principles in use cases versus ALTAI principles (repeat of Table 5)

UC	Ethics principles in use cases	ALTAI Principles (Requirements)	
UC-principles covered in ALTAI			
UC2	Robustness and reliability	Req #2	Technical robustness and safety
UC5	Privacy and data protection	Req #3	Privacy and data governance
UC1, UC3	Transparency and explainability	Req #4	Transparency
UC3, UC4	Non-bias, fairness and non-discrimination	Req #5	Diversity, non-discrimination and fairness
UC1, UC4, UC6	Accountability and responsibility	Req #7	Accountability
UC-principles address sub-parts of ALTAI principles: *UCs each consider disparate aspects of Req #1 or sub-aspects as distinct ethics principle*			
UC2	Human oversight	Req #1	Human agency and oversight
UC5, UC6	Autonomy/User agency	Req #1	Human agency and oversight
UC2, UC3	Over-reliance and deskilling	Req #1	Human agency and oversight
UC-principles named in different ways but addressing aspects similar to ALTAI			
UC1, UC6	Non-maleficence ***Focus:*** *Covers important aspects within General Safety*	Req #2	Technical robustness and safety
UC4	Freedom of expression and non-censorship ***Focus:*** *Covers important aspects of Oversight*	Req #1	Human agency and oversight
UC-principles named in similar ways but addressing aspects different from ALTAI			
UC5	Safety/Human safety ***Difference:*** *Addresses primarily safety of users rather than safety of AI system*	Req #2	Technical robustness and safety
UC5	Human well-being ***Difference:*** *Addresses individual well-being, rather than broader societal or environmental issues*	Req #6	Environmental and societal well-being

4.2. INTERDEPENDENCE OF COMPONENTS

Across the six use cases, 37 components emerged, covering 30 unique aspects (definitions of the component as provided by use cases, see Appendix C). Table 9 lists the components use cases identified for each of the ethics principles.

The colour-coding in Table 9 illustrates the degree of inconsistency in positioning components within ethics principles. In the table, overlaps are indicated by identical colour, e.g., mentions of 'Privacy, consent and data protection' are highlighted in light blue, mentions of 'Transparency and explainability' are marked in light green.

Overlaps can be found for seven of the twelve ethics principles: (1) Non-bias, fairness and non-discrimination; (2) Accountability and responsibility; (3) Privacy, consent and data protection; (4) Autonomy; (5) Human oversight; (6) Transparency and explainability; (7) Over-reliance and deskilling.

The most noticeable inconsistencies emerged for ethics principles that were also listed as components:

- **Human oversight**, while being a principle itself, is being listed as component in four other principles: Accountability and responsibility; Over-reliance and deskilling; Robustness/reliability; Robustness/reliability
- **Privacy** is listed as component in the three principles: Autonomy; Non-maleficence; Safety/Human Safety
- **Transparency** is listed as a component in the three principles: Non-bias, fairness and non-discrimination; Privacy, consent and data protection; Autonomy

This clearly illustrates that the same ethics consideration (e.g., transparency) can be seen as either overarching principle or as aspect of other ethics considerations (e.g., transparency as part of non-bias).

Moreover, across use cases the same component could be linked to different ethics principles. For instance, accuracy appeared as component in human oversight and non-maleficence, while auditability emerged as a component in robustness/reliability, non-maleficence, and accountability.

Together, these observations illustrate again that in end-user perspectives the status of principles versus components and the exact relations of components may be much more fluid than in more abstract representations within AI ethics frameworks.

Table 9: Components listed for each of the ethics principles (colours mark identical focus between ethics principles and components; black background marks principles without overlaps)

Ethics principle from the use cases	Components
Non-bias, fairness and non-discrimination	Diversity Representativeness/inclusivity Objectivity Non-stigmatising use / proportionality Equality and impartiality Transparency of criteria
Accountability and responsibility	**Auditability** Human oversight Liability (responsibility) Human agency and responsibility Professional competence **Auditability** and evaluation Responsiveness
Privacy, consent and data protection	User consent and transparency Data Minimisation, data use and storage Third-party sharing and compliance
Autonomy	Transparency and User understanding Privacy Risk of over-attachment and dependency Informed Consent System Customisation
Human oversight	**Validity / accuracy** Bias Privacy
Transparency and explainability	Accessibility / access to information Explainability Justifiability Openness Documentation, traceability, and **auditability**
Over-reliance and deskilling	Dependence Contestability and human oversight Preservation of human skill and expertise Feedback and learning loops for human adaptation Organisational policies for shared responsibility
Freedom of expression, non-censorship	Autonomy and agency Proportionality Non-discrimination
Robustness/reliability	**Auditability** Human oversight Liability (responsibility)
Non-maleficence	Subsidiarity and proportionality Effectiveness Societal well-being **Validity / accuracy (from technical perspective)** Bias Privacy
Safety/human safety	User protection Security Measures Human Oversight
Human well-being	Promotion of User's health Scope boundaries Crisis Recognition

4.3. A TIDY PRINCIPLE-BASED FRAMEWORK IS AN ILLUSION

The initial, straightforward comparison of ALTAI with the UC-specific ethics principles demonstrates that the bottom-up process conducted in AIOLIA, which started from very concrete AI use cases, foregrounds very similar ethics concerns as established frameworks. The high-level principles of frameworks such as ALTAI thus seem well-reflected in practitioner discussions about the practical implementation of AI ethics. However, we also observed that across the diverse use cases contexts, different focus and emphasis was given to either overall ethics concerns (e.g., Human oversight) or specific sub-aspects (e.g., auditability, deskilling, safety).

More pertinent, however, is the view into the components identified as basis for the operationalisation of AI ethics. Across the very different use case contexts, we found how strongly ethics principles and components are interlinked and interdependent. The high degree of overlaps clearly illustrates that:

- ethics principles are not exclusive, separate, or independent from each other,
- ethics principles do not form a division of AI ethics into 'branches' or static 'subfields'; rather their meaning is constantly reconfigured in practical contexts,
- similar components can be relevant for multiple principles.
- ethics principles can even be seen as components of other principles, and
- other considerations (human rights, bioethics principles, environmental ethics principles) may enter in specific contexts and be lifted to the same level of importance as AI ethics principles

The bottom-up data collection process thus elicited important insights into the challenges of determining how exactly to define ethics principles, what they entail, how they differ from other principles, and whether it is ever possible to address one principle without also touching on others.

This means that a tidy principle-based framework is an illusion. In operational practice it is largely dissolved, because implementation is considerably more complex than a managerial checklist. Other components may arise, and practical technical and operational measures may address several principles or components at the same time. Assessing, evidencing and auditing these complex implementation processes is therefore a task that requires ongoing dialogue between ethics experts and AI engineers or researchers, a process that can be achieved through operationalisation in co-creation.

4.4. RELEVANCE FOR PRESENTATION OF PRACTICAL MEASURES

The observations above also have implications for the presentation of practical measures in this report. The core approach of work in AIOLIA and therefore also in D3.1 is acknowledging the importance of the contexts in which AI shapes human cognition and behaviour, and in extension the importance of end-user and practitioner perspectives when shaping applied ethics instruments. We therefore decided to retain the empirically developed ethics principles and present practical measures along the 12 UC-defined ethics principles, rather than merging them into ALTAI key requirements. However, the presentation of principles is ordered along the logic of ALTAI as listed below:

- *UC-principles linked to 'Human agency and oversight' (ALTAI requirement #1)*
    - Human oversight
    - Autonomy/User Agency
    - Over-reliance and deskilling
    - Freedom of expression and non-censorship
- *UC-principles linked to 'Technical robustness and safety' (ALTAI requirement #2)*
    - Robustness and reliability
    - Safety/Human Safety
    - Non-maleficence
- *UC-principles linked to 'Privacy and data governance' (ALTAI requirement #3)*
    - Privacy and Data Protection
- *UC-principles linked to 'Transparency' (ALTAI requirement #4)*
    - Transparency and explainability
- *UC-principles linked to 'Diversity, non-discrimination and fairness' (ALTAI requirement #5)*
    - Non-bias, fairness and non-discrimination
- UC-principles linked to 'Well-being' (ALTAI requirement #6)
    - Human Well-being
- UC-principles linked to 'Accountability' (ALTAI requirement #7)
    - Accountability and responsibility

5. Concrete operationalisation guidance for ethics principles

This section synthesises the practical technical and organisational measures which were identified across the six use cases. (The full set of practical measures as provided by the use cases can be found in Appendix D.)

The synthesis aims to provide a concise overview of the practical measures for each of the twelve ethics principles. Where an ethics principle was addressed by more than one UC, the synthesis also aims to indicate measures that were shared across use cases and those that were unique for each use base. Our synthesis also reviewed insights about challenges and resources required for putting in place the practical technical and organisational measures to directly support their implementation.

In summary, each principle section provides the following information:

- Focus of the ethics principle
- Technical and organisational measures (indicating shared versus case-specific measures, where applicable)
- Challenges in operationalising the ethics principle
- Resources required to implement the ethics principle

5.1. UC-PRINCIPLES LINKED TO 'HUMAN AGENCY AND OVERSIGHT' (ALTAI REQUIREMENT #1)

5.1.1 Human oversight

Focus of the principle: A continuous focus throughout Use Cases was the need to preserve meaningful human control, agency and responsibility over decisions and actions. AI must function as a decision-support tool rather than an autonomous authority, particularly in contexts where outputs may affect individuals' rights, safety or opportunities.

Across all Use Cases, human oversight is recognised as a core safeguard against inappropriate reliance on AI and as a prerequisite for ethical and legally compliant deployment of AI tools. Human oversight is not only relevant when physical harm is at stake but also where AI shapes professional judgement, behaviour or opportunities. Autonomy is protected not by removing automation, but by ensuring that humans remain empowered to intervene, override and contextualise AI outputs. Effective oversight therefore depends on both system design and organisational culture.

Table 10: Human oversight: Overview of practical measures across use cases

Component type	Shared themes	Use-case specific operationalisations
Technical measures	**Logs/Documentation:** Human-in-the-loop workflows **Guidance:** clear labelling of AI-generated outputs **Direct intervention:** manual override and pause functions confidence flags triggering review	**Logs/Documentation:** Clinician sign-off interfaces in healthcare **Direct Intervention**: safe-fail and override protocols in automotive, moderator escalation tools in virtual assistants; review checkpoints in workplace and security systems
Organisational measures	**Logs/Documentation:** Defined oversight roles **Guidance:** training for reviewers **Direct Intervention:** approval procedures, escalation and accountability mechanisms	**Logs/Documentation:** HR and worker-representation review processes **Guidance:** Clinical responsibility frameworks; safety engineer validation roles; trained moderators and narrative developers in security

Case Study example: An AI system flags certain outputs as high-risk or uncertain, requiring mandatory human review before any action is taken. Reviewers are trained to critically assess recommendations, can override them with documented justification, and are supported by clear procedures defining when and how intervention is required.

Challenges in operationalising Human oversight

There were several recurring challenges noted:

- **Automation bias:** which can undermine oversight, as users may defer to AI outputs even when review mechanisms exist
- **Workload and time pressures:** If overstretched, humans may be tempted to not rely on depth and consistency in their work
- **Unclear or insufficiently supported decisions:** This comes into play when there are more than one team involved interacting with the system. Roles and responsibilities must be clearly defined in guidance

Resources required to implement Human oversight

Whilst it may seem obvious, it is crucial to enact this principle that organisations invest in human resources. Reviewers must be trained, given sufficient time, and supported to exercise independent judgement as opposed to entirely relying on AI systems. Technically, systems must support intervention through clear interfaces, override capabilities and confidence signalling. The organisation must foster a culture that values critical engagement with AI rather than unquestioning acceptance of its outputs, recognising that autonomy and oversight are a dynamic ever-changing process.

5.1.2 Autonomy/User agency

Focus of the principle: Autonomy encompasses disparate aspects of user autonomy: individuals should be able to understand, question and influence how AI systems interact with them, and should not be compelled to accept automated outputs without the possibility of human review or intervention. Across all Use Cases, autonomy is protected not by removing automation, but by ensuring that humans remain empowered to intervene, override and contextualise AI outputs. Effective oversight therefore depends on both system design and organisational culture.

Table 11: Autonomy/User agency: Overview of practical measures across use cases

Component type	Shared themes	Use-case specific operationalisations
Technical measures	**User consent:** consent mechanisms **Access controls:** restrict to people that understand purpose and limitations of the AI	**Access controls:** Minimise number of people with access
Organisational measures	**Guidance:** internal policy on permissible behaviour and expectations of outcomes **Training:** clear community guidelines; user education resources about core features (e.g., origin of material) **User feedback:** feedback process; appeals process	**Impact assessment for possible over-attachment:** qualitative research with patients and therapists to explore the risk **Guidance:** edge case escalation guidance for human moderators necessary to make it as independent as possible from subjective judgement **Business advantage:** Competitor analysis

Challenges in implementing Autonomy/User agency

There were several challenges noted:

- Grey areas exist where consent may need to be sought during AI usage rather than before
- Excessive data collection for access controls can conflict with privacy
- Not every possible borderline case is predictable, therefore depending on post-hoc assessment and intervention
- Competitors are not transparent about own strategies
- Tagging disrupts user satisfaction and providers could therefore decide against it
- Users' engagement with guidelines can vary greatly
- Conflict are possible with the autonomy and decision-making authority of the provider

Resources required to implement Autonomy/User agency

Autonomy on the user side requires the creation of resources that allow users to make informed decisions independently. This includes educational resources for users and published documentation of permitted and prohibited content. Moreover, mechanisms for users to make content moderation decisions as well as automated explanations when content is restricted of flagged. On the business side, to keep competitive advantage, regular monitoring of competitor policies can help assess market positioning.

5.1.3 Avoidance of over-reliance and deskilling

Focus of the principle: Over-reliance and deskilling are concerned with the risk that AI systems substitute, outweigh and will eventually erode human judgement, expertise and responsibility. When AI outputs are treated as authoritative as opposed to advisory, users may defer uncritically to automated recommendations, leading to automation bias, reduced vigilance and loss of context specific reasoning. This principle reinforces the need for AI systems to be designed and governed to support, not replace human expertise, and that organisations actively preserve human competence, critical thinking and accountability.

Across all use cases, over-reliance is recognised as a gradual and often invisible risk, emerging through routine use rather than system malfunction or wilful negligence. While automation can improve efficiency and consistency, it also changes how humans engage with tasks, potentially shifting from active decision makers to passive validators. The synthesis has shown that deskilling is not limited to safety critical domains; it can also affect professional judgement, learning and institutional knowledge in everyday settings. Effective mitigation requires ongoing attention to how humans and AI collaborate in practice.

Table 12: Avoidance of over-reliance and deskilling: Overview of practical measures across use cases

Component type	Shared themes	Use-case specific operationalisations
Technical measures	**Knowledge:** Clear labelling of AI outputs, requirements for justification when accepting AI recommendations **Evaluation:** confidence or uncertainty indicators, comparison views between AI and human assessments,	**Evaluation:** Comparative safety analyses in automotive, clinician–AI comparison panels in healthcare; explanation prompts for moderators; bounded automation in workplace and security tools clinician–AI comparison panels in healthcare; explanation prompts for moderators; bounded automation in workplace and security tools
Organisational measures	**Knowledge- users:** mentorship and knowledge-sharing practices **Knowledge – organisations:** Training and continuous professional development **Oversight:** human validation for high-impact decisions	**Knowledge – organisations:** clinical training; engineering mentorship and certification; HR training to contextualise AI-derived scores **Oversight:** Peer review in healthcare

Case Study example: An organisation requires users to document their reasoning when accepting or rejecting AI recommendations and periodically reviews cases where AI outputs were overridden. These practices ensure that human expertise remains active and that learning flows in both directions—from humans to systems and from systems to humans.

Challenges in operationalising Avoidance of over-reliance and deskilling

Themes that emerged in terms of challenges in this area were:

- **Efficiency pressures:** Users may welcome reduced cognitive effort, particularly when working in high-volume, unfamiliar or time-critical environments
- **Time and resource constraints:** Investment in training and up-skilling of humans requires monetary investment from organisations who may not be able to afford such processes
- **Measurement indicators:** Deskilling is difficult to quantify as loss of expertise may only become visible when systems fail or face out of the normal situations

Resources required to implement Avoidance of over-reliance and deskilling

Alongside human investment, this principle requires a willingness to invest in organisational learning. Technically, systems must be designed to encourage reflection rather than passive acceptance, for example, through explanation prompts or comparison views. Organisationally, training programmes, mentorship structures and protected time for review and learning are essential. Experiences professionals must be supported to share vital knowledge and to supervise less experienced users. Finally, organisations must commit to valuing human expertise as a long-term asset, recognising that preserving skills and judgment is essential for resilience, accountability and ethical AI use.

5.1.4 Freedom of expression and non-censorship

Focus of the principle: Freedom of expression and non-censorship require that AI systems do not unjustifiably supress, distort or restrict lawful expression, and that any limitations on speech are proportionate, transparent and grounded in legitimate aims such as safety, legality and harm prevention. This principle recognises that AI systems used for content analysis, moderation or behavioural assessment can significant influence which voices are heard, and what topics are constrained. It therefore demands careful balancing between protecting individuals and society from harm and preserving the fundamental right to express opinions, ideas and perspectives without undue interference.

Freedom of expression is treated in these Use Cases as a qualified right, as opposed to an absolute one, requiring proportional and context-sensitive implementation. This is mainly due to the need to protect the physical and emotional safety of others which could be at risk with total unregulated freedom of expression. Safety-critical and security contexts may be able to justify certain restrictions, but all sectors emphasise the need to avoid over-blocking, viewpoint discrimination and opaque decision-making. It is also important to note that censorship risks are not confined to public facing platforms. Internal AI systems can also shape what individuals feel able to do or say. Effective protection of expression relies heavily on transparency, contestability and human judgement

Table 13: Freedom of expression and non-censorship: Overview of practical measures across use cases

Component type	Shared themes	Use-case specific operationalisations
Technical measures	**System:** Threshold-based classification; tiered response systems **Explanation:** explanation of restrictions **Logs/documentation:** logging of moderation actions	**Oversight:** Context-aware moderation in virtual assistants **System:** alternative-narrative systems in security **Limitations:** conservative flagging thresholds in workplace tools
Organisational measures	**Oversight:** Clear moderation policies; appeal and review mechanisms; human review of borderline cases; legal and policy oversight teams	**Oversight:** cultural and linguistic reviewers **Stakeholder engagement:** coordination with external platforms or regulators

Case Study example: An AI-supported moderation system classifies content into severity tiers and applies the least restrictive response compatible with safety goals. When content is limited or removed, users receive a clear explanation and can request human review, ensuring that restrictions remain proportionate and accountable.

Challenges in implementing Freedom of Expression and Non-censorship

Organisations face several challenges in applying this principle:

- **Contextual interpretation of speech is difficult to automate:** which could lead to false positives or inconsistent decisions
- **Legal and cultural standards of acceptability differ greatly:** These not only vary across jurisdictions but also across cultural borders, which complicates the creation of a uniform design
- **Pressures to prevent harm:** This may incentivise overly cautious moderation, increasing the risk of unnecessary censorship
- **Security, legal and operational constraints**

Resources required to implement Freedom of Expression and Non-censorship

Organisations must commit to ongoing review and dialogue which recognises that protecting freedom of expression is shaped by social, legal and technological change. Technically, systems need configurable thresholds, logging and explanation features. Organisationally, clear policies, appeal processes and governance structures are essential to ensure proportionality and accountability. Finally, as with the other principles, human investment is paramount. Trained moderators, cultural experts and legal advisors must be available to review complex cases and update policies as norms change and evolve.

5.2. UC-PRINCIPLES LINKED TO 'TECHNICAL ROBUSTNESS AND SAFETY' (ALTAI REQUIREMENT #2)

5.2.1 Robustness/reliability

Focus of the principle: Robustness and reliability is firstly a technical feature of the AI system describing its ability to operate reliably, securely, and predictably under both normal and adverse conditions. The practical relevance lies in the resilience of the AI system against errors or malicious attacks allowing it to withstand or recover from them. It further ensures the continuous assurance that an AI-based or automated system performs its intended safety and functional tasks consistently, accurately, and predictably across all phases of its lifecycle — from design and validation to deployment, operation, and maintenance.

Robustness and reliability has a second focus on users of the AI system that expects that all functions and decisions are applied consistently, transparently, and without unjust bias toward any group of users, contexts of operation, or system components. All users therefore should receive an equitable level of protection, consideration, and accountability throughout the system's lifecycle, regardless of demographic, geographic, or technological differences.

Table 14: Robustness/reliability: Overview of practical measures across use cases (no shared themes, as only UC2 addressed this principle)

Component type	Operationalisations
Technical measures	**Data:** diverse and representative training data; data quality, integrity, and validation **Testing:** bias testing and model auditing; multi-stakeholder validation; boundary testing **System:** resilient architecture **Logs/documentation:** versioned lifecycle traceability **Quality – checks:** continuous monitoring of fairness drift; continuous performance validation
Organisational measures	**Culture:** safety culture in the organisation **Quality – review:** cross-disciplinary reviews; AI quality management system

Challenges in implementing Robustness/Reliability

Challenges are present for technological as well as human/organisational aspects

- The data exchange between AI engines, report generators, and human-validation modules must remain deterministic and auditable.
- Generating clear explanations in the UI without cognitive overload is a design challenge.
- Designing multi-layer, redundant, and self-monitoring systems introduces high complexity in both hardware and software
- Realistic resilience testing (e.g. sensor loss, corrupted data, cyberattack) is expensive and time-consuming
- Heterogeneous and fragmented data sources; incomplete or noisy data; lack of ground truth for validation

- Non-linear and opaque model logic makes it hard to predict and isolate failure triggers; impossibility to test every edge case.
- On the human and organisational side, the following issues may arise:
    - *Culture*: integrating AI processes into existing quality management systems, resistance to change, inconsistent adherence to processes, lack of awareness of AI-specific risks
    - *Quality review*: scheduling conflicts, insufficient expertise across disciplines, potential for groupthink or overlooked biases

Resources required to implement Robustness/Reliability

The implementation of review (safety reasoning, hazard traceability) and feedback mechanisms emerged as a core requirement, which needs to cover every phase of AI development and deployment. This also includes regular validation cycles using benchmark datasets and expert-labelled cases.
Technically, version control tools are required to record dataset versions, model weights, code commits, and validation reports. Organisational resources are required for regular training sessions to ensure knowledge is kept up-to-date. In the longer term, promoting a culture of safety-first decision-making requires sustained management engagement, establishment of clear (and potentially new) roles and responsibility, as well as updates to quality and reward procedures. Further, time and resources are needed to allow for regular review workshops, including structured disagreement analysis and under involvement of multiple stakeholders for each safety-critical deliverable.

5.2.2 Safety/human safety

Focus of the principle: Safety as an ethical principle ensures that AI systems protect users from harm and operate within secure, controlled, and ethically guided boundaries. It requires robust user protection measures to prevent the generation or dissemination of harmful, violent, or psychologically distressing content. Security mechanisms–including age verification, multi-tier content classification, and active monitoring of harmful usage patterns–must be in place to detect and prevent unsafe or violent outputs, issuing warnings or bans when necessary. Equally, human oversight remains essential: trained moderators must be able to intervene when automated systems encounter complex or borderline scenarios. This combination of technical safeguards and human judgment ensures that AI systems act responsibly, uphold user well-being, and prevent the normalization or spread of harmful material.

Table 15: Safety/human safety: Overview of practical measures across use cases (no shared themes, as only UC5 addressed this principle)

Component type	Operationalisations
Technical measures	**System:** tiered severity system; multi-tier classification models **Features:** real-time content filtering; pattern detection algorithms; jailbreak detection **Data:** labelled training data sets
Organisational measures	**Focus:** cross-functional tier definition **Oversight:** regular policy review; legal compliance review; human moderation oversight; progressive intervention protocol

Challenges in implementing Safety/human safety

A number of challenges emerged that may hinder implementation of Safety/human safety:

- *Data management:*
  - managing cross-border data flows
  - possibility of re-identification is a risk for data privacy
  - identifying what data is truly necessary and how long it has to be stored to fulfil safety obligations
  - clear classification of cases that require a great deal of context and individual case decisions
  - complete removal of identifying elements is in conflict with monitoring system
  - balancing security and usability
- *Legislation:*
  - different regulations/policies regarding data encryption dependent on region
  - keeping up with latest regulations
  - liability for decisions
- *Training*: Training of moderators
- *Users protection*:
  - extensive monitoring is required to fulfil safety requirements, but users do not need to know about any detail of their data processing
  - being as transparent as possible while handling the risk that users feel their privacy violated
  - exposing moderators to potentially traumatic content only possible for a limited period of time

Resources required to implement Safety/human safety

Organisations should set up clear and concise guidance and privacy policies on data collection, use, and sharing practices that users can easily access before they access the AI system. These policies should also ensure implementation of transparent data policies and consent mechanisms. Users should further be given the right to access, delete, and transfer their data. At the same time, in drafting the guidance and policies, it is important to consider differences in legal requirements across jurisdictions. Advised is also the establishment of a comprehensive compliance program that includes data protection policies, procedures, staff training, and regular audits. This includes regular evaluation of privacy risks in monitoring systems.

Resources are also required for the management of vast amounts of logging data, compliance with logging standards and regulatory requirements. Safety also stretches to the protection of personal data with the de-identification of personal data for model training where possible.

5.2.3 Non-maleficence

Focus of the principle: Non-maleficence requires that AI systems are designed, deployed and used in ways that prioritise preventing harm and that minimise risk, both to individuals and society as a whole. Harm is not only deemed as physical, but as psychological, social or institutional and can arise from both incorrect system outputs and from how those outputs are interpreted or acted upon. This principle therefore demands continued vigilance, recognising that risk may emerge over time as

systems are updated, used in new contexts or interact with human logic and decision making. Non-maleficence also implies a duty to anticipate reasonably foreseeable harms, detect failures at the earliest possibility and intervene before harm occurs.

Across all use cases, non-maleficence is treated as a continuous responsibility rather than a one-time assurance. Technical measures alone are insufficient: even well-validated accredited systems can cause harm if used outside their intended scope or if early warning signs are ignored. Harm is not limited to overt safety failures; it can also arise through misclassification, inappropriate escalation, reputational damage, or erosion of trust. Effective non-maleficence therefore depends on combining monitoring, human judgement and responsive governance.

Table 16: Non-maleficence: Overview of practical measures across use cases

Component type	Shared themes	Use-case specific operationalisations
Technical measures	**Alignment – systems:** robustness and resilience testing **Bias prevention:** anomaly detection **Oversight:** Performance monitoring **Fairness:** safeguards against misuse	**Bias prevention:** Automated content filtering and behaviour pattern detection in virtual assistants **Oversight:** Evaluative frameworks for narrative effectiveness in security Hazard and failure-mode analysis in automotive Monitoring of risk scores Alerts in workplace/HR systems **Fairness:** Clinical validation and subgroup monitoring in healthcare
Organisational measures	**Oversight:** Risk assessment processes; Escalation and incident-response workflows **Bias prevention:** Periodic reviews; feedback loops	**Fairness:** Medical ethics committees and post-market surveillance External reviews **Oversight:** Functional safety reviews and lifecycle validation **Alignment – systems:** Medical Practitioner checkpoints external reviews; moderation escalation protocols and user protection policies

Case Study example: An AI system that flags potential risks is continuously monitored for abnormal patterns, such as sudden increases in false positives. When thresholds are exceeded, outputs are paused or downgraded pending human review, and corrective actions are documented. This prevents unsafe or misleading results from propagating while enabling learning and improvement.

Challenges in operationalising Non-maleficence

The commonalities of challenges across use cases when operationalising non-maleficence were

- **Harm is often delayed or indirect:** Which then creates difficulties in establishing thresholds for intervention or clear indicators for professionals to be mindful of

- **Data and Context shift:** Time between development, deployment and delivery can mean that models that were validated at deployment may degrade or behave unpredictable in new conditions
- **Resource Constrictions:** Can limit the frequency and depth of monitoring, particularly when human review is required
- **Tension between priorities:** If usability and efficiency are organisational priorities there can be difficulties when balancing that preventing harm and safeguards, which can lead to professionals disengaging from using the tools or the necessary precautions to prevent harm

Resources required to implement Non-maleficence

Operationalising non-maleficence requires sustained investment in both technology and staff. Technically, organisations need monitoring infrastructure, validation pipelines and mechanisms to pause, rollback or constrain system outputs. Organisationally, there must be clear escalation pathways, incident-response procedures and defined accountability for harm mitigation. Human resources are critical: domain experts must be available to interpret signals, assess risk and decide when intervention is necessary. Finally, organisations must allocate time and governance systems for regular review, recognising that preventing harm is an ongoing dynamic process, rather than a compliance milestone.

5.3. UC-PRINCIPLES LINKED TO 'PRIVACY AND DATA GOVERNANCE' (ALTAI REQUIREMENT #3)

5.3.1 Privacy, consent and data protection

Focus of the principle: Privacy, consent and data protection require that AI systems are designed and used in ways that respect an individual's rights to control their personal data, safeguard confidentiality and ensure lawful, fair and transparent processing. AI systems often rely on large volumes of personal or behavioural data, including data inferred from user's actions rather than directly provided. Ethical AI usage therefore demands that individuals are informed about how their data is collected, used and that consent is meaningful and freely given. Data processing must be limited to what is necessary for a clearly defined purpose. Protecting privacy is not only a legal obligation but a precondition for trust and legitimacy.

All Use Cases acknowledged that privacy risks are not limited to consumer-facing technologies. Internal AI systems can significantly affect individuals' dignity, sense of surveillance and willingness to engage openly within an organisation. In internal contexts, individuals may feel heightened pressure to consent or may lack realistic alternatives. Which amplifies the ethical importance of data minimisation, purpose limitation and clear safeguards.

While the regulatory landscape varies greatly by sector and jurisdiction, all use cases converge on the need for proportionality, transparency and continuous oversight. Across sectors extensive logging and monitoring are introduced to support accountability and safety, yet these same mechanisms increase privacy risks if not carefully controlled. Regardless of sector, organisations rely on the processing of personal or behavioural data to enable AI functionality, which makes privacy protection inseparable from questions of trust, legitimacy and user autonomy.

Table 17: Privacy, consent and data protection: Overview of practical measures across use cases

Component type	Shared themes	Use-case specific operationalisations
Technical measures	**Access controls – people:** encryption, access controls **Security – data:** Data minimisation, anonymisation or pseudonymisation where possible **Logs/documentation:** audit logging,	**Access controls – people:** secure logging and access control in safety systems in medical settings, restricted data access in security contexts **Security – data:** Strong safeguards for sensitive clinical data; **Logs/documentation:** large-scale consent and data management in virtual assistants and HR systems
Organisational measures	**Access controls – data:** Clear privacy policies, consent management processes **Compliance:** Compliance monitoring **Evaluation – impact:** data protection impact assessments	**Access controls – data:** Hospital data governance and retention policies **Transparency:** coordination with external partners such as payment processors

Case Study example: An organisation clearly informs users about what data are collected, how long they are retained, and for what purposes they are used. Users can access, correct, or delete their data, and sensitive information is protected through encryption and role-based access, ensuring both legal compliance and user trust.

Challenges in the implementing of Privacy, consent and data protection

As explored above there are tensions within the nature of this principle which presents challenges to the operationalising of privacy, consent and data protection, such as:

- **Balancing data minimisation with system performance and safety requirements**
- **Consent:** Which can be complex to manage in environments with power asymmetries, such as workplaces or essential services
- **Cross-border data flow:** Legalities and safeguards change regionally and can offer complexity when designing tools that are utilised across different legal structures
- **Extensive logging and monitoring incidentally creating privacy risks**

Resources required to implement Privacy, consent and data protection

To enact this principle ethically and soundly in practice, organisations must be willing to develop secure technical infrastructures, have relevant legal expertise and ensure continuous organisational governance. Technically, organisations need robust data security measures, access controls and lifecycle data management tools. Organisationally, clear policies, compliance programs and defined responsibilities (such as named data protection officers) are essential. Finally, organisations must allocate time and authority for regular review and adaptation.

5.4. UC-PRINCIPLES LINKED TO 'TRANSPARENCY' (ALTAI REQUIREMENT #4)

5.4.1 Transparency and explainability

Focus of the principle: Transparency and explainability require that AI systems are developed and used in ways that make their existence, purpose, functioning and limitations understandable to those who use them and those that are impacted by their usage. It cannot be communicated simply through technical explainability but needs to be included in the organisational accountability ethos and its openness to scrutiny. Transparency ensures that individuals are aware when and why AI is being used, then in turn, how its usage may impact them. Explainability requires that organisations can meaningfully explain how and why specific outputs or decisions have been produced. Together, these elements enable informed participation, trust, oversight and contestability.

Across all use cases, transparency and explainability are noted as key foundational enablers of the ethical implementation of AI. The inclusion of organisational and workplace systems demonstrates that explainability is not only required when there is the potential for high-risk or safety-critical decision making, but also wherever AI affects people's opportunities, treatment or behaviour. Whilst the level of technical detail varies by audience and domain all sectors recognise that explanations must be contextual, accessible and actionable, rather than overtly technical. Transparency is therefore best understood as a layered practice, combining high-level openness with the ability to provide deeper explanations when needed.

Table 18: Transparency and explainability: Overview of practical measures across use cases

Component type	Shared themes	Use-case specific operationalisations
Technical measures	**Explanation:** User-facing explanations **Knowledge**: confidence or uncertainty indicators **Bias prevention:** confidence or uncertainty indicators documentation of model behaviour, logging of decisions and inputs	**Knowledge**: Patient-facing summaries in healthcare **Explanation:** visualisation of safety boundaries and confidence levels in automotive sector; explanation notices in moderation systems; simplified explanations for employees in workplace tools
Organisational measures	**Explanation:** Information policies, documentation standards, audit and review procedures, clear allocation of responsibility for explanations	**Explanation:** Clinical communication protocols: safety documentation aligned with engineering standards **Knowledge:** transparency statements constrained by security considerations; platform policies and user education resources

Case Study example: An AI-supported assessment includes a clear notice that AI was used, a plain-language explanation of the main factors influencing the result, and information on how the output can be reviewed or challenged. More detailed documentation is available for auditors or experts when required.

Challenges in operationalising Transparency and explainability

Whilst noted as a cornerstone of ethical AI usage, Transparency does present a series of challenges across Use Cases that are also important to detail:

- **Tension between clarity and cognitive overload:** Providing too little information undermines trust, whilst overly technical explanations risk alienating non-technical audiences.
- **Conflict with security, confidentiality or intellectual property:** One of the largest challenges this task faced was the gathering of information on AI tools due to the above concerns, which is also mirrored in certain sectors' (such as policing, defence, etc.) ability to be transparent at decision making levels.
- **Adaptability:** Organisations must ensure that transparency practices remain up to date as their systems evolve, rather than disclosures becoming static and disconnected from system behaviour.

Resources required to implement Transparency (and explainability)

Technically, organisations need explainability features, logging mechanisms and interfaces tailored to different audiences. Organisationally, they require clear communication strategies, documentation standards and assigned responsibilities for the maintenance of transparency. Human resources are critical: staff must be trained to interpret explanations, respond to questions and engage with users or regulators. Finally, organisations must allocate time for review and iteration, recognising that effective transparency is an ongoing practice which will evolve along AI systems and the needs of the users.

5.5. UC-PRINCIPLES LINKED TO 'DIVERSITY, NON-DISCRIMINATION AND FAIRNESS' (ALTAI REQUIREMENT #5)

5.5.1 Non-bias, fairness and non-discrimination

Focus of the principle: The aim of this principle is to ensure that AI tools are designed, deployed and delivered in a way that avoids potential disparities in how individuals or groups are assessed, treated or affected. This principle acknowledges that bias arises at multiple stages in the AI lifecycle – through data selection, model design, deployment context or human interpretation – and that discriminatory outcomes may occur without explicit intent. Fairness therefore demands proactive measures to ensure that AI systems reflect the diversity of the populations they are utilised within, apply objective and consistent criteria and are used in ways that respect dignity, equality and proportionality. Table 19 provides an integrated view of practical measures across use cases.

Fairness is understood as both a technical and social challenge. Technical bias mitigation alone is insufficient if systems are deployed or interpreted in ways that systematically disadvantage characteristic groups. Discriminatory effects may be subtle, cumulative and embedded in everyday practices such as risk scoring, prioritising or access to opportunities. Effective fairness therefore requires continuous monitoring, contextual awareness and organisational willingness to question whether system outputs align with ethical and legal commitments to equality.

Table 19: Non-bias, fairness and non-discrimination: Overview of practical measures across use cases

Component type	Shared themes	Use-case specific operationalisations
Technical measures	**Bias prevention:** Diverse and representative datasets, bias testing and monitoring **Human-in-the-loop:** subgroup performance analysis, fairness drift detection	**Bias prevention:** Subgroup validation in healthcare, sensitivity analysis in automotive safety models **Human-in-the-loop**: cultural and linguistic checks in security narratives; monitoring of moderation outcomes across user groups; fairness reviews of risk scores in workplace systems
Organisational measures	**Bias prevention:** Inclusive design processes, proportional use policies **Human-in-the-loop:** multi-stakeholder review, documented criteria for decision-making,	**Bias prevention:** Involvement of clinicians and patient representatives in healthcare **Human-in-the-loop:** cultural experts and trusted messengers in security, HR, legal, and worker-representation input in organisational AI

Case Study example: An organisation regularly analyses AI outputs across different user groups to identify disparities. Where differences are detected, thresholds and decision rules are reviewed, and outputs are used for supportive interventions rather than punitive actions, ensuring proportionality and avoiding stigmatisation.

Challenges in operationalising Non-bias, fairness and non-discrimination

Organisations face persistent challenges in identifying and addressing bias:

- **Relevant diversity dimensions may be difficult to measure or legally sensitive:** This will have an impact on available data for organisations to rely on when designing their own models
- **Bias may emerge from system use rather than model design:** Which makes early detection harder and increases likelihood of discrimination occurring
- **Fairness trade-offs:** "Fair" is a complex, context-dependent concept; improving outcomes for one group may unintentionally affect others, etc.

Resources required to implement Non-bias, fairness and non-discrimination

Operationalising this principle requires sustained access to data expertise, domain knowledge and stakeholder input. In terms of technical measures, organisations need tools for bias testing, subgroup analysis and continuous monitoring. Organisationally, they require inclusive governance structures, clear documentation of criteria and processes for revising system behaviour when disparities are

identified. As with other principles, significant investment in human oversight is also crucial. Diverse perspectives must be involved throughout design and deployment, and staff must be trained to recognise and respond to potential discrimination.

5.6. UC-PRINCIPLES LINKED TO 'WELL-BEING' (ALTAI REQUIREMENT #6)

5.6.1 Human well-being

Focus of the principle: Human Well-being encompasses the obligation to ensure AI systems promote users' physical and psychological health, support their flourishing as individuals, and avoid causing harm through inappropriate advice, reinforcement of unhealthy patterns, or failure to recognize crisis situations. This includes providing genuinely helpful habit formation support while preventing the AI from encouraging harmful behaviours or failing to identify when users need professional intervention beyond AI capabilities.

Table 20: Human well-being: Overview of practical measures across use cases (no shared themes, as only UC5 addressed this principle)

Component type	Operationalisations
Technical measures	**System:** Automated classifiers for harmful advice prevention **Control:** Multi-layered scope control **Focus:** Having an external solution specifically trained for crisis detection
Organisational measures	**Quality checks:** advisory consultation; ethics review for new features **Feedback:** User feedback collection **Limitations:** Explicit scope boundaries

Challenges in implementing Human well-being

Challenges to human well-being are

- **Users:** Users trying to jailbreak the system
- **User experience:** Tagging disrupts user satisfaction as applications operate in a critical area where the boundaries of harmful behaviour must be carefully balanced
- **Identifying harmful impacts:**
  - Minor changes that do not constitute a new feature that have consequences on AI behaviour in other contexts are overlooked
  - Defining harmful advice in borderline cases

Resources required to implement Human well-being

Protections to ensure human well-being require a combination of technical and organisational means. Technical features need to ensure the system does not allow or suggest harmful uses. Access to appropriate training data and procedures that allow output monitoring to prevent harmful responses. Further, automated classifiers may be used to detect search for topics that suggest psychological or physical harms or risks (e.g., weight loss, medication-related requests). Clear policies and terms of conditions need to be put in place about the purpose, focus and deployment of the AI systems. To capture possible negative impacts, a reporting system and customer team should be available where users can give feedback, concerns or complaints.

5.7. UC-PRINCIPLES LINKED TO 'ACCOUNTABILITY' (ALTAI REQUIREMENT #7)

5.7.1 Accountability and Responsibility

Focus of the principle: Accountability and responsibility require that AI systems are developed, deployed and governed in ways that make it clear who is responsible for decisions, outcomes and harms, and that these responsibilities can be demonstrated, reviewed and enforced over time. Across the Use Cases, accountability can be operationalised through a combination of technical mechanisms that make system behaviour traceable and organisational measures that allocate, exercise and sustain human responsibility. Whilst there are components and themes that resonate across sectors, the implementation varies according to domain-specific risks, regulatory environments and institutional roles.

Across all sectors, accountability is not treated as an abstract ethical idea but as an operational property of socio-technical systems. Technical traceability enables audits and reconstruction of decisions, whilst organisational structures ensure that responsibility remains anchored to identifiable human agents. Accountability must be distributed across departments and hierarchical responsibility structures, rather than centralised, and it must combine engineering controls, procedural safeguards and institutional governance. At the same time the principle is shaped by context: where risks involve, for example, patient safety, road safety, societal harm, national security or large-scale user exposure, accountability mechanisms are adapted to reflect the nature of the potential impact.

Table 21: Accountability and responsibility: Overview of practical measures across use cases

Component type	Shared themes	Use-case specific operationalisations
Technical measures	**Logs/documentation:** Audit logs and traceability of AI outputs; logging of human validation **Quality – data:** version control for models and data **Overrides and human-in-the-loop:** access control and role-based permissions; overrides, and changes	**Logs/documentation:** Case-level evidence chains and appeal records in healthcare; standards-driven traceability aligned with safety artefacts in automotive systems; large-scale logging and moderation records integrated with platform and payment systems in virtual assistants **Security – systems:** restricted or undisclosed tooling in security contexts due to confidentiality
Organisational measures	**Incident handling:** defined escalation and incident-response processes; **Oversight and governance:** Clear allocation of roles and responsibilities (e.g. RACI); mandatory human-in-the-loop review for high-impact decisions; governance and oversight structures	**Oversight:** Clinical governance boards and clinician sign-off in healthcare; safety managers and formal approval workflows under ISO/SOTIF in automotive; shared responsibility across technical, legal, and moderation teams, including external partners, in virtual assistants and worker representatives/works council in HR professionals; **Knowledge – field expertise:** locally grounded narrative developers and trusted messengers in security;

Case Study example: In high-risk contexts, AI-generated outputs are logged together with the identity and rationale of the human reviewer who approved, modified, or rejected them, ensuring that responsibility remains transparent and auditable even when automation is used at scale.

Challenges in operationalising Accountability and Responsibility

Across use cases, organisations face a common set of challenges when translating accountability and responsibility from principle to practice.

- **Fragmented ownership:** responsibility for data, models, decisions and impacts is often distributed across technical, legal, HR and operational teams, making it difficult to maintain clear end-to-end accountability. Without explicit role definitions and escalation pathways, accountability risks becoming symbolic rather than operational
- **Documentation and traceability adding to practitioner fatigue:** Maintaining up to date logs, version histories, decision records and audit trails requires continuous effort and discipline, particularly as systems are swiftly evolving. Use Cases report tensions between delivery speed and the time needed to document decisions in a way that is auditable and meaningful
- **Vulnerability of human oversight mechanisms:** Even when there are review processes and override mechanisms in place, operational pressures, automation bias or unclear incentives can lead to over-reliance on AI outputs and reduced scrutiny in day-to-day use
- **Regulatory complexity and uncertainty:** Use cases report particular difficulties when there are multiple legal regimes that all apply simultaneously (for example, AI Act, GDPR and sector specific safety/labour rules). Organisations must interpret and reconcile overlapping obligations without having clear precedents or expertise to navigate these.

Resources needed to implement Accountability and Responsibility

Effective operationalisation consistently depends on a combination of technical, organisational and human resources. From a technical perspective, organisations require a robust infrastructure for logging, version control, access management, and audit retrieval, integrated into normal development and deployment workflows rather than treated as optional additions

Organisationally, formal governance structures are essential. This includes clearly defined roles (e.g., systems owner, reviewer, approver), decision-making bodies (such as AI or ethics governance boards) and procedures that link accountability to change management, incident handling and escalation

Human resources are equally critical. Sustained training and capacity-building are needed so that clinicians, engineers, HR professionals, moderators or managers understand both their responsibilities and the limits of AI outputs. Accountability relies not only on tools and policies, but on people who are empowered, competent and supported

Finally, organisations require time and institutional commitment. Accountability cannot be implemented as one-off compliance exercises; it demands ongoing monitoring, review and adaption as systems, data and organisational contexts change.

6. Overarching considerations: Relevance, resources, assessment, challenges and risks

As part of the co-creation process for operationalising ethics principles, use cases were also asked to provide reflections on overarching considerations that are relevant for the effective application of the practical measures. The subsequent sections summaries those considerations.

6.1. RELEVANCE OF THE PRACTICAL MEASURES

According to use cases, implementing the practical measures is relevant to achieving seven disparate aspects linked to risk management, as well as achieving benefits.

Table 22: Relevance of practical measures supporting successful risk-management

Supporting successful risk-management
A) Achieving compliance with respect to
• Compliance with laws and regulations • Facilitating accountability • Successful passing of audits • Privacy requirements
B) Ensuring safety and security and reducing harms
• Preventing problems and harms, e.g., by detecting risks early, preventing misuse, protecting minors from accessing inappropriate materials, reducing biases • Improved safety culture • Better incident handling, including reduced response time to issues Protect dataset and reduce risk of data breaches, including secure/legal data exchanges
C) Creating shared understandings and ownership
• Relevant stakeholders have a shared understanding of AI's intended use, known limitations, and validation evidence. • Clear ownership of safety outcomes
D) Safeguarding sustainability
• Knowledge of AI users is kept up-to-date to ensure users can interpret AI recommendations, identify potential biases, and maintain responsibility for safety decisions. • AI systems and processes are kept in line with technological, regulatory and societal developments

Table 23: Relevance of practical measures supporting organisational benefits

Achieving organisational benefits
E) Gaining positive outcomes with respect to
• Successful achievement of AI purposes • Benefits are in line with expectations by AI users, clients, customers • Improved economic opportunities for AI developers or companies that use AI systems
F) Achieving positive perceptions and acceptance
• Improved trust by AI users, clients, customers • Satisfaction of users, clients, customers • Acceptance of AI as such and/or specific AI usage
G) Improving effectiveness of processes
• Improved and more efficient processes of AI use and/or AI integration into existing processes • Better consistency and robustness of procedures such as incident handling by responding to violations in an appropriate, consistent and systematic manner. • Improved governance

6.2. RESOURCES AND REQUIREMENTS TO SUCCESSFULLY IMPLEMENT PRACTICAL MEASURES

Across all use cases, five main requirements emerged to deploy AI ethically and responsibly into organisations.

Table 24: Summary of resources and requirements for implementation of practical measures

A) Setup of oversight mechanisms
• **Human-in-the-loop controls** should be installed as universal safeguards in high-risk contexts. Whether through clinician sign off in healthcare, safety engineer approval in automotive systems, or moderator review in virtual assistant/security applications, human oversight remains essential to ensure that automated outputs are contextually appropriate and ethically sound. • **Continuous monitoring and feedback mechanisms** is required to reinforce accountability, using dashboards, thresholds, and feedback loops to detect drift and feed lessons into training and process updates.
B) Ensure systems and processes can deliver transparency and traceability
• **User-facing transparency** is required, operationalised through confidence indicators, reasoning traces and explanations – whether via visual cues, lay summaries, or detailed justification panels – enabling users to understand and contest system outputs. • **End-to-end traceability** should be supported by versioning, immutable logs and changeable management processes. These measures allow for independent audits, facilitate incident reconstruction and ensure accountability across the AI lifecycle. • **Fairness safeguards,** including disagreement analysis, sensitivity testing, and bias detection, should be embedded to ensure equitable outcomes across user groups.

D) Establish full life-cycle governance
• **Commitment to lifecycle governance** is needed, with structured workflows, periodic reviews, and role clarity embedded through standard operating procedures (SOPs) and compliance with domain-specific standards.
E) Ensure people are trained and retain AI competencies
• **Training and competency development** should be prioritised through structured onboarding, certification, and mentorship.
F) Security and system integrity
• **Security and system integrity** should be a core focus, with encryption, access control, and cryptographic artefacts protecting data and infrastructure.

6.3. APPROACHES TO ASSESS SUCCESSFUL IMPLEMENTATION OF PRACTICAL MEASURES

User case partners provided a wide range of criteria to assess the successful implementation of practical measures. The below list gives a detailed overview of the seven core measures and actions that emerged from the data.

Table 25: Ways to assess successful implementation of practical measures

A) Logs and record keeping

• **Keeping logs and records** allows to capture and maintain evidence of actions, decisions, performance, adherence to laws, regulations, professional standards, etc.
• **Focus of the logs and records** may be
  ○ **Critical decisions** taken
  ○ **Corrective actions or changes** conducted and related change management
  ○ **Appeals**, appeals process and outcomes
• **Security test** such as pen-tests

B) Guidelines, protocols and contracts
• Drafting and implementation of clear guidelines and protocols • Contracts, e.g., with external AI or data providers • Set up of safeguards, e.g., 'no account generation possible without consent'
C) Clear, pre-defined and stable metrics and KPIs

• Example metrics/KPIs named in use cases to test fulfilment are:
  ○ incident response KPIs meet targets
  ○ time to triage incidents
  ○ mean time to repair (MTTR)
  ○ alert precision/recall
  ○ complaints rate
• **Stability of metrics** ensures that results remain comparable and meaningful over time and AI implementations

Table 25: Ways to assess successful implementation of practical measures (continued)

D) Evaluations and feedback

• **Set up of an evaluation process** that meaningfully assesses process, performance, outcomes, limitations, etc.; features suggested by use case partners are
  ○ **Feedback loops** to ensure evaluation leads to improvements
  ○ **Use 3rd party evaluators** to check for compliance, e.g., with EU data protection requirements
  ○ **Use of dashboards** to make review of information more effective and efficient
• **Evaluation against clear criteria** to assess whether effects are as expected
• **Focus of an evaluation** might be
  ○ AI model performance
  ○ usability, satisfaction and acceptance
  ○ human skills and performance, including AI-human interactions
  ○ validation of well-functioning oversight process
• learning continuity and expertise transmission

E) Audits

• **Implementation of audits**, regular inspections and audit retrievals
• **Auditors** should be a combination of internal and 3rd party audits
• **Audits are suggested for**:
  ○ internal audits of AI processes, including with legal teams
  ○ AI output traceability and QA checks
  ○ human oversight process, with verification that every final report includes at least one human reviewer's and one responsible engineer's digital signature
  ○ review of recorded knowledge assets (case analyses, "lessons learned" files) stored in internal repositories
  ○ review of change management logs, with time-stamped validation and review actions
• onboarding documentation for completeness and updates

F) Implementation of changes

• **Implementation of change requests** from audits, user feedback, etc
• **Set up of structured change management** to guide, including regular reviews whether required changes have been implemented and are successful
• **Focus of changes** may be, amongst other:
  ○ AI system, models and outcomes
  ○ human-oversight process
  ○ knowledge and skills

G) Adequate knowledge and understanding

• **Conduct (repeated/pre-post) surveys for structured assessment** about understanding of the AI system, functions, limitations, impacts, legal and ethics requirements, etc.
• **Conduct regular training** to ensure knowledge stays up to date
• Check training attendance and success through
  ○ attendance logs
  ○ completion rates of CPD modules
  ○ renewal certifications

6.4. CHALLENGES FOR THE IMPLEMENTATION OF PRACTICAL MEASURES

Across all use cases, consistent themes emerged regarding challenges to implementation of practical measures. These cross-cutting challenges highlight that embedding AI ethically demands not only technical competence but also organisational adaptability, human resilience, and sustained ethical reflection.

Table 26: Overview of core challenges to successfully implement practical measures

A) Data and system-related challenges
• **Data and validation** form an ongoing challenge in ensuring data quality, accessibility, and representativeness. Healthcare, automotive, and security use cases all report barriers related to fragmented datasets, privacy constraints, and inconsistent validation frameworks. • **Compliance and privacy management** present persistent tensions, especially in sectors handling personal or sensitive data such as healthcare and virtual assistants. Organisations must reconcile strict privacy requirements with the need for transparency and data-driven improvement. • **Balance between complexity and interpretability**: As AI systems become increasingly sophisticated, ensuring that outputs remain transparent and comprehensible to human users becomes more difficult. Designing user interfaces that explain reasoning without overloading users is a shared design challenge.
B) Integration-related challenges
- **Workflow and integration of AI in existing processes** are two of the most prominent challenges, as partners across sectors face challenges embedding AI tools without disrupting established professional roles, regulatory procedures, or decision-making workflows. Integrating new technologies requires balancing innovation with the need for continuity and control, often creating friction within existing systems.
C) Workforce-related challenges
- **Human workload and oversight pressures**: These data issues often intersect with human workload and oversight pressures, as human-in-the-loop mechanisms—essential for maintaining ethical and safety standards – are resource-intensive and prone to fatigue and inconsistency. Clinicians, safety engineers, and moderators must sustain attention and judgment across complex, repetitive, or emotionally demanding tasks.
D) Resource-related challenges
- **Resource and infrastructure strain** further compounds the difficulty of implementing most practical measures. Continuous validation, retraining, and monitoring require substantial financial, computational, and human resources.
E) Context-related challenges
- **Cultural and ethical sensitivity** is a recurring theme, particularly in healthcare and security, where social and emotional contexts shape the ethical evaluation of AI use.

6.5. RISKS IF PRACTICAL MEASURES ARE NOT FULFILLED

The use cases identified six core risks in case practical measures principles are not addressed successfully.

Table 27: Overview of main risks to successfully implement practical measures

A) Security and data breaches
• **Missing or insufficient security measures** in the AI system and its data can lead to misuse, tampering or general liabilities and legal consequences. This includes sharing with third parties. • **Data breaches equally** can results in privacy complaints and legal consequences. • **Inconsistent application of safety procedures** and unclear responsibilities can lead to delayed responses or inconsistent responses to incidents.
B) Problematic quality of decisions or outcomes
• **Errors in AI outputs or unsafe AI outputs** due to biases, inconsistent procedures or lacking controls leading AI outputs to be inaccurate, biased, or non-compliant. • **Wrong or unsafe decisions** which may be due human errors in reviewing AI outputs, subjective moderation decisions (biases) or persistent usability issues with the AI. • **Errors may be recurring and remain undetected** if no measures are put in place, increasing the risk of harms.
C) Lacking accountability
• **Accountability gaps** if no justification can be provided for decisions about AI systems or outcomes, which lead to liability exposure in case of accidents or regulatory inspections. • **Unclear accountability arrangements** and unclear responsibilities can lead to audit failures, compliance issues and delays in incident handling. • **Lack of traceability** can threaten privacy requirements and lack of understanding what causes safety incidents, which means issues cannot be resolved.
D) Harms to AI users, clients or patients
• **Harms** can occur and/or continue through biased/incorrect content which may be challenging to re-capture once it spread, or if AI hazard/biased are missed and remain uncorrected.
E) Negative outcomes organisations

• **Deskilling** if overreliance on AI leads to
  ○ acceptance of erroneous outputs,
  ○ discontinuity of organisational safety knowledge and experience,
  ○ progressive loss of tacit safety expertise among engineers, or
  ○ reduced human capacity to intervene effectively during unexpected system behaviours.
• **Loss of reputation** and **loss of user trust** in case of misuse, negative consequences of the AI use.
• **Economic damages** and **lost economics opportunities** can result in case of data leaks, if AI systems are misused or harms come to users or individuals affected by the AI deployment.
• **Legal consequences** and liabilities for AI users or companies.

6.6. SUGGESTIONS TO IMPROVE THE QUALITY OF ETHICS OPERATIONALISATIONS

This section provides recommendations on how to ensure the ethics operationalisations process yields high quality results. The recommendations are based on the reflections and insights provided by use case partners, after going through the operationalisation process in their own use case. These insights were collected as part of our methodology (see Section 3: Methodology) and are integrated into the list of practical recommendations below.

6.6.1 Recommendation 1: Plan for a collaborative multi-disciplinary and durable process

The successful operationalisations were based on strong, longer-term collaboration between the use case partners, using the respective strengths of partners (practitioner and academic) in the process. The design process further relied on continuous communication and feedback loops between research and practice-oriented participants ensuring that the theoretical and practical perspectives aligned. This also ensures that perspectives can be reviewed and validated or refined over time.

Table 28: Recommendation set 1: Collaborative multi-disciplinary and durable process

Iterative co-development
- An iterative workflow allows initial findings or proposals to be refined through successive stages of review, discussion and validation. - Plan for multiple rounds of feedback, covering all relevant stakeholder groups, to develop the consensus on the final components. - Document any potential disagreements, that may remain (cf. Recommendations 7)
Industry scaffolding drafting
- Industry partners are often best place to produce an first concrete list of measures grounded in their workflow and experiences.
Testing in real-case scenarios
- Where possible, the operationalisations should be tested in several realistic deployment scenarios with the AI system, diversity of scenarios (e.g., from common/every-day operation to critical incidents).

6.6.2 Recommendation 2: Right people in the process

A common challenge in the operationalisation process is that it only includes a limited number of experts, such as legal or technical experts in the organisation. As a result, perspectives from other relevant stakeholders are missing in the discussions and the operationalisation outcome. This can lead discussions, e.g., around principles such as explainability being framed primarily as a technical issue requiring a technical solution, which may not represent end-user concerns such as usability, interpretability or trustworthiness of the system.

Table 29: Recommendation set 2: Right people in the process

Ensure Inclusive stakeholder selection
- Stakeholder selection should be comprehensive to ensure an inclusive and balanced approach to ethics issues and ethics implementation. A comprehensive approach avoids narrow discussions, e.g., from primarily legal or technical perspectives. - Especially end-users of the AI system should be represented from the start of the process. Also consider groups that may not directly use the system but may still be affected by its impacts. - Keep stakeholder selection adaptive, as the process may identify the need to engage with additional stakeholder groups.
Ensure correct expertise in the team conducting operationalisation
- The team leading the operationalisation process needs to high expertise areas in the relevant industry, application context of the AI and the AI capabilities. - Moreover, a combination of technical and organisational expertise should be included to ensure technical and organisational measures can be addressed competently. - Adequate expertise levels not only ensure quality of the outcome but can also serve as a gateway to build trust in the team and the process.

6.6.3 Recommendation 3: Practicality and accessibility as basis for formulation of measures

Practical measures are only useful if they are feasible within the specific industrial and operational context in which the AI is deployed and if the individuals tasked with their implementation understand them and understand their rationale. Therefore, practicality and accessibility should be at the forefront in the formulation of practical measures, keeping the core audience(s) in mind that need to implement, assess and/or audit them.

Table 30: Recommendation set 3: Practicality and accessibility

Practicality and proportionality of measures
- Balance best practices of AI ethics implementations with feasibility (it is possible) and proportionality (is it proportionate). - Practical measures therefore should be: • implementable given current systems and resources, • proportionate to the scale and risk of the deployment, • clearly owned (who does what, in which team). - These considerations should be reflected in the distinction between technical vs organisational measures and the explicit assignment of stakeholders.
Measures need to be understandable and reachable for different employee groups
- Focus on “usable transparency” as a practical priority. - For this, treat transparency and accessibility together, with attention to: • plain language and realistic reading-age targets, • role-adapted explanations (employees vs managers), • multiple channels for non-desk/shift workers, not just intranet/email.

6.6.4 Recommendation 4: Operationalisations for early-stage AI systems

Operationalisations may be conducted when AI capabilities or AI systems still are in an early, evolving stage. In this case, some aspects may only be assessed as 'not yet in place' or 'planned' rather than directly observable in practice (e.g., dashboards, KPIs, user behaviours) and thus based on anticipated risks and safeguards, not yet on real deployment experience. This does not invalidate the operationalisation process, but where operationalisations address early features, they should clearly be marked as 'preliminary' or 'forward-looking' and will need to be revisited once the system is fully in place.

Table 31: Recommendation set 4: Operationalisations for early-stage AI systems

Clearly mark 'preliminary' operationalisations
- Clearly mark practical measures that address features in development as 'preliminary' and 'need to review'.
Include future users or audiences in the process
- Aim to include a full cross-section of future users or worker representatives when developing practical measure, to ensure that measures (especially on organisational culture and day-to-day use by managers).
Review as systems/features more into more mature stage
- Plan in review cycles once the AI system/feature moves into a more mature stage.

6.6.5 Recommendation 5: Considerations for business sensitive and classified contexts

Successful operationalisation tends to require a wide range of fine details about the AI capabilities, data, deployment, and outcomes, etc., as well as the business processes and users involved in their deployment. In industrial settings, such information is often sensitive, in security settings it may also be classified and bound by legal restrictions. This makes collecting the right amount of information challenging. This may also be accompanied by concerns that internal information could leak out to the detriment of market advantages. It is therefore important to understand whether any special requirements apply for an AI deployment and from there consider the right choice of people, process, data handling, etc.

Table 32: Recommendation set 5: Business sensitive and classified contexts

Identify special requirements and individuals/group with correct access
- Clarify whether there are specific requirements for people involved in the operationalisation process (e.g., rights to access certain types of information, clearance levels) or the process (e.g., data access/management, documentation). - Identify which internal personnel or external groups of people have the correct clearance levels.
Identify any special requirements for implementation of practical measures
- Clarify any special requirements with respect to practical measures, especially for technical versus organisational measures. - Clarify how potential disparities in requirements are to be managed along the process (definition of practical measures, implementation, assessment, auditing, etc).

6.6.6 Recommendation 6: Keep an open communication stance

Conversations about ethics can raise concerns in industrial partners that the process is primarily about finding gaps and faults. This presents the risk that ethics conversations might become defensive. For instance, asking too directly about ethics issues can reduce trust and interest in further conversations. Moreover, ethics discussions may be seen as stigmatising 'high-risk' employees or as introducing hidden performance or trust metrics. An open, practice- and solution-based communication stance helps support constructive engagements.

Table 33: Recommendation set 6: Open communication stance

Focus discussions core interests and practical benefits
- It is crucial to recognize that industry partners often do not think in a framework of ethics principles, but rather in terms of the advantages to the application and uses own terminology. - By concentrating on the industry partner's core interests (e.g., protection of users, usability) conversation can emphasise the practical value of implementing ethics.
Emphasise non-stigmatising, proportionate AI use
- To avoid stigmatising discussions, consider • using scores mainly for support, coaching and training; • favouring group/role-level insights over individual rankings; • requiring human review and context before any high-impact AI use.

6.6.7 Recommendation 7: Safeguard transparency and auditability of the process

The operationalisation process involves numerous decisions about who to involve (stakeholders), how to run the process (methodology), how and what to document, etc. To ensure that decisions along the process can be revisited, good documentation is important. This also includes potential challenges and disagreements during discussions. Disagreements are a normal part in the operationalisation process and can provide important information about differences in priorities or perspectives across stakeholders. Documenting them allows to review these differences at a later stage. Overall, documentation not only of the operationalisation outcomes but also of the process supports transparency, auditability and accountability.

Table 34: Recommendation set 7: Safeguard transparency and auditability

Document how the operationalisation was developed
- Keep a record of core decisions and reasons for those decisions, such as • People/groups involved • Approach and methodology followed (e.g., literature reviews, expert interviews, group discussions) • Instruments used to collect information or guide discussions (e.g., data collection templates, questions lists) • Steps taken during data analysis • Decisions on inclusion or removal of ethics principles, components and/or practical measures • Sources that suggested the inclusion or exclusion of ethics principles, components and/or practical measures

Document challenges and disagreements
- Keep a record of disagreements or challenges throughout the process to allow revisiting different perspectives or priorities that may not be represented in the final outcome.

6.6.8 Recommendation 8: Ensure sustainability

AI systems or their usage may change over time, e.g., through modifications/updates, changes in legislation, or changes in business practices or interests.

Table 35: Recommendation set 8: Ensure sustainability

Regular review of the operationalisation
- Plan in reviews, either in regular intervals or at important milestones (e.g., major feature changes/updates), to ensure the operational guidance is still relevant, and update if needed.
Regular review of relevant stakeholders and impacts
- Review whether new stakeholders or impacts need to be considered that require reflection in the operationalisations (e.g., new users with a new risk profile).

7. Updated process for the operationalisation of AI ethics

The operationalisation of AI ethics principles is a complex multi-disciplinary bottom-up process based on co-creation amongst stakeholders (Mittelstadt, 2019). Based on the learnings from the co-creation process, we updated the operationalisation pathway originally developed by D2.3 into a simplified version.

The simplified process proposes a 4-step process to facilitate the operationalisation of AI-related ethics in applied settings. This updated process is context-independent, in the same way as the original operationalisation pathway, and as such applicable across contexts, AI systems, features and AI application areas.

The subsequent sections describe the four steps, outlining the core considerations and information collected at each of the four steps.

Table 36 provides the outline of the simplified operationalisation process. As this table shows, the 4-step process progresses in a sequence from identifying the high-level ethics principles to the formulation of concrete, practical measures for their implementation and their validation.

Table 36: Overview of the proposed 4-step process for ethics operationalisation

Steps
1. Identification of the relevant ethics principles
Which ethics principles are relevant for your context and AI deployment?
EXAMPLE: Identified principle: Transparency (source considered for identification: ALTAI)
2. Identification of components for each ethics principle
What are the components that together describe the ethics principles?
EXAMPLE: Components identified to describe Transparency: Traceability, Explainability, Communication
3. Creation of practical measures per component
How to achieve the successful implementation of each component?
EXAMPLE: Practical measures identified to implement Traceability: Logging mechanisms, data labelling, documentation
4. Validation
Are the ethics principles, components and practical measures complete and described correctly?
EXAMPLE: A review identifies that further measures are required that focus on organisational practices to fully ensure Traceability, e.g., establishment of a governance and oversight mechanism that will guide the procurement and deployment of the AI system.

7.1. STEP 1: IDENTIFICATION OF ETHICS PRINCIPLES

The identification of ethics principles should ensure that AI deployments consider and implement ethics in a comprehensive manner.

Information required for each ethics principle:
1. A definition of the ethics principle 2. Rationale for selecting this principle (risks managed/benefits gained, see above) 3. Sources that suggest its inclusion (e.g., national laws and regulations, professional standards, …) 4. Any disagreements between people involved in the identification process (e.g., on the meaning or inclusion of the ethics principle)

Considerations
The identification of ethics principles should be guided by two perspectives:
1. **Risk management:** What are main risks when deploying AI that need to be avoided/managed? 2. **Benefits gathering:** What can your organisation gain when fulfilling these principles. Benefits can be wide-ranging such as increased legitimacy, reputation, quality of products or services, business opportunities, etc.
Completeness versus feasibility:

- The ambition in the identification of relevant ethics principles is certainly to be as comprehensive as possible, while remaining feasible to implement.
- To ensure the identification is comprehensive, the following approaches have shown as beneficial in the use cases:
  ○ Involve a broad set of stakeholders and experts in the process (AI, legal, ethics, (prospective) users, customers/consumers, …)
  ○ List core risks and benefits from (not) fulfilling AI ethics (see above) and ensure that the ethics principles address all detected risks and benefits
- To increase feasibility, consider removing ethics principles that show large overlaps in their practical measures

Inter-dependence:
- Ethics principles are often interlinked, which means there can be overlaps in their focus or in the practical measures needed to address them (cf. Section 4). - This interdependence is not in itself problematic and should not impact the identification of ethics principles. However, it can help inform decisions about completeness vs feasibility (see above).

7.2. STEP 2: IDENTIFICATION OF COMPONENTS FOR EACH ETHICS PRINCIPLE

The identification of components per ethics principle ensures that all relevant aspects of a principle are captured and subsequently translated into practical measures. This process is referred to as 'decomposition' in the Practical Guidance (D2.3) and refers to "the translation of the chosen high-level ethical principles into actionable and context specific components" (D2.3, p. 19).

Information required for each component:
1. A definition of the component 2. Its relevance for the given context (why select this component and not others?) 3. Sources that suggest the inclusion of this component 4. Any disagreements on meaning (definition) or relevance of the component

Considerations
Completeness versus feasibility:

- Similar to ethics principles, the ambition in the identification of relevant components is to be as comprehensive as possible, while remaining feasible to implement.
- To ensure the identification of components is comprehensive, the following approaches have shown as beneficial in the use cases:
  ○ Review existing frameworks and academic literature
  ○ Involve a broad set of stakeholders and experts in the process (AI, legal, ethics, (prospective) users, customers/consumers, …)
- To increase feasibility, consider removing components that show large overlaps in their focus or practical measures

Inter-dependence:
- Similar to ethics principles, components are often interlinked and/or may address more than one ethics principle. For instance, safety considerations can focus on technical challenges such as data and cybersecurity but can also be linked to well-being outcomes of users. - This interdependence is not in itself problematic and should not impact the identification of ethics principles. However, it can help inform decisions about completeness vs feasibility (see above).

7.3. STEP 3: CREATION OF PRACTICAL MEASURES PER COMPONENT

Practical measures ensure the successful implementation of the ethics principles. However, the practical measures are created for the components, rather than the ethics principles themselves, and should be concrete and actionable.

Information required for each practical measure:
1. Detailed description of the measure 2. Why is it relevant? 3. How can it be achieved? 4. How can be assessed whether this measure has been fulfilled? 5. What are (potential) challenges to fulfilment? 6. What are risks if not fulfilled? 7. What are benefits of fulfilling it? 8. Which are the functions/roles/stakeholders responsible for implementation and monitoring? 9. Are there specific requirements to consider (e.g., laws/regulations, industrial standards, organisational obligations, etc)?

Considerations
Practical measures should cover both:
- **Technical measures:** which focus on the design and features of AI capabilities, as well as methodologies and processes to ensure AI is operated in an ethical manner - **Organisational measures:** which focus on organisational resources, structures, procedures, policies, knowledge and culture that ensure AI is operated in an ethical manner
Detail required:
- Practical measures need to actionable *and* auditable. - Hence, they need to be described in sufficient detail – not only for people that implement them (e.g., AI developers) but also for people who need to monitor, review and audit their fulfilment (e.g., quality control, legal departments, oversight bodies, ….). - Descriptions should also reflect on responsibilities, resource requirements, criteria to judge successful implementation, likely implementation challenges, etc. A list of suggested information is provided below.
Contextualisation:
- **Stakeholders and audiences:** Are there stakeholders or audiences that require specific types of practical measures (e.g., *oversight bodies* that require *auditing logs* to ensure *accountability*)? Are there practical measures that need adaptation depending on the audience (e.g., the levels of *technical detail* in *documenting* decisions to ensure *traceability*)?

7.4. STEP 4: VALIDATION

Validation ensures that the operationalisation is as complete as possible and correctly identified and describes the ethics principles, components and practical measures. It aims to find and fill potential gaps and remove ambiguities and inconsistencies.

Considerations
Validation purposes:
- **Review** for any missing or unclear information and terminology and reduce overlaps - **Testing in actual use cases**: use case testing ensures that the operationalisation is fit-for-purpose and easy to use by the intended audience - **Finalisation** of the guidance
Completeness and understandability:
- Check with all relevant audiences for completeness, terminology, feasibility, etc. considering both organisational (internal) and external perspectives
Sustainability:
To ensure the guidance stays up-to-date and relevant, - feedback and experiences during the implementation of the guidance should be used to update where needed - the guidance should be reviewed on regular basis

8. Conclusion: AI ethics in human cognition and behaviour

The Operational Ethics Guidelines on Use Cases related to Human Behaviour and Cognition are concrete guidance co-created in six highly diverse European use cases. The insights gathered throughout the operationalisation pathway reveal the strength of a bottom-up co-creation process not only for creating a portfolio of likely technical and organisational measures to implement AI ethics, but also to appreciate how practitioners understand, conceptualise and apply ethics in their concrete AI deployment contexts.

The information in this report helps to understand how successful operationalisation of AI ethics principles in the context of human cognition and behaviour can be supported in practical terms, offering a large portfolio of practical measures, combined with concrete recommendations on designing the process itself.

Yet, the process also provides learnings about applying ethics where AI affects human cognition and behaviour. Firstly, the six use cases differed in that four addressed professional contexts (UC1-4 in health, automotive industry, HR and national security, respectively) and the remaining two private contexts (UC5+6 for personal virtual assistants and trauma/grief therapy, respectively). Practical measures in the two private contexts showed an emphasis on individual safety and well-being, as well as on user autonomy and agency. Professional contexts, in contrast, more often identified accountability, transparency and non-biases in their ethics requirements. While it is challenging to make clear statements with the limited number of use cases in each category, these observations suggest that professional contexts may focus more strongly on managing risks in the decision making from AI systems and their traceability and quality, whereas private contexts were more concerned with end-users' well-being and behaviours.

Reviewing the practical measures further reveals the extent to which they focus on disparate aspects of AI impacts: some aim to detect and monitor the impacts, some aim to prevent and safeguard against negative impacts or rectify them, while others aim to gain positive outcomes. For instance:

- ***Monitoring and detecting AI impacts on cognition and behaviours through*** logging and documentation; error analysis; bias detection dashboards; broad community/stakeholder engagement beyond the direct AI user group
- ***Preventing of and safeguarding against negative AI impacts through*** human oversight; user consent mechanisms; diversity-focused risk assessment; system designs choices such as non-stigmatising presentation of scores or automated consistency checks
- ***Rectifying negative AI impacts through*** creation of manual overrides; appeals processes; redress mechanisms; governance rules on model updates; regular policy reviews; organisational learning from human-AI Interaction outcomes
- ***Gaining positive outcomes from AI impacts through*** research on positive impacts of AI; training programs that increase knowledge about AI in users and broader society; competitor analysis to understand business advantages of using AI

All these aspects are relevant for a comprehensive ethics monitoring and implementation, and ethics operationalisations should thus ensure that technical and organisational measures cover all four areas.

Another important observation from the portfolio of practical measures is the extent to which they address different time scales: while many aim to manage risks at the time of AI usage (*cognition based*: e.g., avoiding wrong decisions or harmful advice, preventing disappointments or harms by managing expectations what an AI application can do; *behaviour based*: e.g., preventing wrong users such as minors from accessing the AI application), some also aim to forestall impacts with a longer term focus. Examples of longer-term risks are the potential deskilling of the work force over time, which is addressed by adopting continuous professional development and certification renewal procedures, or the risk that AI applications become ethically and legally unsound, which is answered by conducting ethics review for new features or regular policy reviews. Use cases are thus clearly aware that impacts on human cognition and behaviour are not only a short-term or immediate issues but also raise longer-term concerns that need to be solved systematically and proactively. It is especially the organisational measures, identified by use cases, that ensure the monitoring of and safeguarding against longer-term risks. Therefore, operationalisations should ensure not only that technical and organisational measures are in place but also that, next to short-term concerns, they identify and address long-term risks.

Lastly, the 'humans' necessary for applying ethics in the use cases emerged as a diverse set of groups. Beyond AI designers and end-users, they also include managers and other stakeholders with oversight responsibilities, HR and training providers, people surrounding AI users such as family members, as well as researchers to investigate actual and long-term benefits, law and policy makers, and society as a whole, which should accept AI uses (e.g., in medical or HR decisions).

Thus, while the practical technical and organisational measures formulated by use cases often seem to primarily focus on the immediate AI usage contexts (the systems, their users and – in professional contexts – organisations), reviewing the measures needed to achieve them makes clear that in fact it requires a broad network of actors to set the conditions for their successful implementation; namely: how AI systems and features are decided on and designed, how they are used, who should or should not be using them, what kind of outputs are produced, what is done with the outputs, how outputs and impacts are evaluated and fed back, and how learning and knowledge is created and spread. Where possible, operationalisations should include perspectives from all or as many of these stakeholders as feasible to capture the full picture of ethics requirements, principles, components and practical measures for their implementation.

The Operational Ethics Guidelines provided in this document certainly have their limitations: given the impact of the AI usage contexts on the specific ethical principle and the technical and organisational measures considered in the six use cases, our guidance cannot – and does not – claim to be complete or cover all possible or relevant components or measures. Considerably future work is needed to establish whether the differences between contexts (professional vs personal) and across application across (medical, HR, security, etc) are in fact systematic and can be replicated. Similarly, our Guidance cannot establish whether the practical measures collected are sufficient and effective in ensuring that the ethics principles can be successfully achieved.

Still, its portfolio of practical measures presents a helpful set of diverse means for engineers and organisations to implement, evidence and assess AI ethics in the context of human cognition and behaviour. The discussions during the operationalisation, as well as the validation process moreover highlighted important tensions, challenges and training needs. This material will provide crucial direction for the subsequent work in AIOLIA to develop its non-technical guidance (D3.3) and AIOLIA training materials (WP4).

9. References

Araujo, T., Helberger, N., Kruikemeier, S. & Vreese, C.H. (2020). In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence. AI & Society, 35(3), 611-623, https://doi.org/10.1007/s00146-019-00931-w

Bleher, H., & Braun, M. (2022). Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. AI Ethics 2, 747–761, https://doi.org/10.1007/s43681-022-00135-x

Chan, L., Doyle, K., McElfresh, D., Conitzer, V., Dickerson, J., Schaich Borg, J., & Sinnott-Armstrong, W. (2020). Artificial Intelligence: Measuring Influence of AI 'Assessments' on Moral Decision-Making. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 214–220. https://doi.org/10.1145/3375627.3375870

Corrêa, N.K., Galvao, C. Santos, J.W., ..., Galvao, L., Terem, T., de Oliveira, N. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. Patterns, 4(10), 100857, https://doi.org/10.1016/j.patter.2023.100857

Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, *15*(1), 6. https://doi.org/10.3390/soc15010006

High-Level Expert Group on Artificial Intelligence (AI HLEG). (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Krügel, S., Ostermaier, A. & Uhl, M. (2023). Algorithms as Partners in Crime: A Lesson in Ethics by Design. Computers in Human Behavior, 183, 107483, https://doi.org/10.1016/j.chb.202da07483

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence,1, 501-507, https://doi.org/10.1038/s42256-019-0114-4

Mondal, D. T., Kadyan, D. J. S., Jayaprada, N., Udayakumar, D. S., Upadhyaya, D. M. & S, D. M. (2025). Artificial Intelligence and Social Interactions: Understanding AI's Role in Shaping Human Psychology and Social Dynamics. *Journal of Marketing & Social Research, 2*(4), 208-215, https://doi.org/10.61336/jmsr/25-04-28

Naeem, S., Khan, M., Tariq, U., Dhall, A., Colon, I., & Al-Nashash, H. {2025}. Generation and Detection of Sign Language Deepfakes - A Linguistic and Visual Analysis. https://arxiv.org/abs/2404.01438

Sarridis, I., Spangenberg, J., Papadopoulou, O., & Papadopoulos, S. (2025). Mitigating Viewer Impact From Disturbing Imagery Using AI Filters: A User-Study. International Journal of Human–Computer Interaction, 41(2), 1234–1245. https://doi.org/10.1080/10447318.2024.2313890

UNESCO. (2023). Recommendations on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Appendix A. Summary of criteria to measure successful implementation per principle

Principle 1: Human Oversight

Purpose: To ensure that AI systems support rather than replace, human decision making.

A. Human control by design	
Organisational/Technical	**Measure**
TECH	The AI System is designed as a decision-support tool rather than an autonomous decision-maker
BOTH	Clear points of human intervention are defined for high-risk or sensitive uses of the system
ORG	The organisation has specific which decisions must involve human judgment
TECH	Automated actions can be paused, overridden or reversed by humans
ORG	Human oversight is not static but involves active review, validation or challenge of system outputs
ORG	Oversight responsibilities are clearly defined and embedded into every day work flows
B. User autonomy and informed interaction	
Organisational/Technical	**Measure**
ORG	Checks are in place to ensure individuals using the tool retain meaningful choices and are not coerced into accepting AI outputs
BOTH	Users show they understand when they are interacting with AI and when outputs are AI generated
ORG	Organisation offers users opportunities to question, challenge or deviate from AI recommendations

Principle 2: Autonomy/User agency

Purpose: To ensure that AI tools do not intentionally or unintentionally manipulate users' behaviour, emotions, choices or perceptions in ways that undermine autonomy, dignity or informed decision making.

A. Identification of manipulation risks	
Organisational/Technical	**Measure**
BOTH	Potential manipulation risks arising from the tool design, outputs or interactions have been identified and documented
ORG	Risk analysis considers psychological, emotional and behavioural influences – not only technical performance
BOTH	Attention has been given to asymmetries of power, knowledge or vulnerability between the system and its users
ORG	The organisation has defined what constituted unacceptable manipulation in its specific operational context

B. Design safeguards against undue influence	
Organisational/Technical	**Measure**
TECH	System design avoids techniques intended to covertly steer user behaviour (e.g., deceptive framing, emotional pressure etc)
BOTH	Outputs are framed to inform and support decision-making rather than pressure, persuade or exploit cognitive bias
TECH	The system does not personalise influence strategies in a way that exploit individual vulnerabilities without justification
ORG	Legitimate behavioural influence is clearly distinguished from manipulative practices and documented
C. Transparency and User Agency	
Organisational/Technical	**Measure**
BOTH	Users are informed when system outputs are intended to influence decisions or behaviours
ORG	Users maintain meaningful choice and are not penalised for rejecting, ignoring or questioning system suggestions
ORG	The organisation has processes to assess whether users experience system interactions as coercive or misleading
D. Oversight, monitoring and correction	
Organisational/Technical	**Measure**
BOTH	Human oversight exists to review system outputs for manipulative effects, especially in sensitive or high-impact contexts
ORG	User feedback and complaints related to perceived manipulation are systematically collected and reviewed
ORG	Identified manipulation risks trigger corrective actions, design changes or restrictions on system use

Principle 3: Over-reliance and Deskilling

Purpose: To prevent AI systems/tools replacing or eroding human expertise, judgement and responsibility, and ensuring that automation supports human capabilities rather than diminishing them over time.

A. Managing dependence on AI outputs	
Organisational/Technical	**Measure**
ORG	The organisation has defined, clear boundaries and guidance for appropriate usage of AI systems/tools
BOTH	Decisions with significant impact are not solely based on automated outputs
ORG	Guidance exists on when AI outputs should be questioned, verified or supplemented with human input
ORG	The organisation actively discourages treating AI outputs as definitive or infallible
B. Preserving human judgement in workflows	
Organisational/Technical	**Measure**
BOTH	Workflows require active human engagement with checkpoints
TECH	Tool interfaces avoid designs that encourage automatic acceptance (e.g. default approvals without review).
ORG	There are no penalties for challenging or overriding AI outputs

C. Skills, training and competence	
Organisational/Technical	**Measure**
ORG	Training supports users in understanding both the capabilities and limitations of AI outputs
ORG	Opportunities exist for users to practice independent judgment rather than relying exclusively on automation
ORG	The organisation periodically reassesses reliance patterns as systems evolve or scale

Principle 4: Freedom of Expression and Non-censorship

Purpose: To ensure that AI systems do not unjustifiably restrict, suppress or distort legitimate expression, and that any limitations on expression are proportionate, transparent and grounded in ethical and legal justification.

A. Defining legitimate boundaries	
Organisational/Technical	**Measure**
ORG	The organisation has clearly defined what types of content, behaviour or expression are restricted and why
ORG	Restrictions can be justified on a legal, ethical or safety basis rather than reputational or convenience of the organisation
BOTH	The scope of restrictions is documented and communicated to relevant users or stakeholders
B. Human review and contestability	
Organisational/Technical	**Measure**
BOTH	Humans can review, override or correct automated restrictions on expression
ORG	Clear mechanisms for affected individuals to report, question or contest moderation or restriction
ORG	Human review involves consideration of contextual, cultural or situational factors that automated systems may have missed
C. Transparency and monitoring	
Organisational/Technical	**Measure**
ORG	Users are informed when and why expression is restricted using clear and accessible explanations
BOTH	The organisation monitors patterns of restriction to detect systemic over-censorship or bias
ORG	Policies governing expression and moderation are periodically reviewed as contexts, risks or norms evolve

Principle 5: Robustness/reliability

Purpose: To ensure that an AI system operates reliably, securely, and predictably under both normal and adverse conditions, and to withstand, detect, and recover from errors, perturbations, or malicious attacks that could compromise safety.

A. Technical robustness and reliability	
Organisational/Technical	**Measure**
TECH	Display confidence scores or uncertainty metrics for each AI-generated safety statement
TECH	Use modular and layered design and introduce redundancy and diversity
TECH	Map potential component, data, and algorithm failures
TECH	Implement a fairness drift detector which triggers model retraining if imbalance occurs
TECH	Use sensitivity analysis to check fairness across different input distributions
ORG	Conduct 'disagreement analysis' sessions – if AI and humans diverge, identify bias sources
B. Human-in-the-loop	
Organisational/Technical	**Measure**
TECH	Allow users to expand reasoning traces
TECH	Provide comparison views between human-validated and AI-suggested reports
ORG	Clearly document roles and responsibilities
ORG	Schedule formal review workshops, document decisions, include structured disagreement analysis, and involve multiple stakeholders for each safety-critical deliverable
ORG	Include diverse expert groups in validating tool outputs
C. Traceability	
Organisational/Technical	**Measure**
TECH	Ensure datasets are fully traceable (source, preprocessing, labelling, and version control)
TECH	Use data audits to detect underrepresented cases (edge scenarios, rare hazards)
ORG	Establish SOPs for AI use, version control of datasets and models, validation of AI outputs, audit trails, and periodic reviews
BOTH	Implement version control tools to record dataset versions, model weights, code commits, and validation reports
B. Safety culture	
Organisational/Technical	**Measure**
ORG	Promote a culture of safety-first decision-making
ORG	Set up regular training on AI-assisted analysis and decision-making
ORG	Conduct regular validation cycles, using benchmark datasets, and expert-labelled cases
BOTH	Embed safety reasoning, hazard traceability, and control feedback into every phase of AI development and deployment

Principle 6: Safety/human safety

Purpose: To protect the user from harmful content or consequences and ensure that legal and other compliance obligations are in place and monitored.

A. Safeguards against misuse and prohibited usage	
Organisational/Technical	**Measure**
TECH	AI models are trained on human-labelled datasets to automatically classify content into safety tiers at scale
TECH	Prohibited content during conversations is automatically detected and filtered
BOTH	Systems are in place to identify concerning behavioural patterns over time
BOTH	Measures are in place to identify and prevent attempts to circumvent safety systems
ORG	An escalation process is implemented and communicated from monitoring, to highlighting, to warning to user ban
ORG	Regular legal team assessments of moderation decisions are taking place
B. Tiered security system	
Organisational/Technical	**Measure**
BOTH	Tiered severity system is periodically reviewed, inspected, and updated
BOTH	Tier definitions are continuously assessed against emerging patterns
ORG	Safety categories are developed collaboratively by manual moderators, data team, legal team, and payment platform representatives
C. Human oversight and moderation	
Organisational/Technical	**Measure**
ORG	Curate datasets of examples labelled by human moderators to train the automated systems
ORG	Employ a team of human moderators and ensure supervision and further training on a regular basis

Principle 7: Non-maleficence

Purpose: To ensure that AI tools are designed, deployed and operated in ways that actively reduce harm, minimise foreseeable risks, and avoid causing physical, psychological or institutional damage.

A. Harm identification and risk awareness	
Organisational/Technical	**Measure**
BOTH	Potential harms associated with the AI system (direct and indirect) have been identified and documented before deployment of tool
ORG	Harm assessments consider multiple dimensions of impact (e.g. safety, dignity, fairness etc)
BOTH	Risks are reviewed not only at design stage but also when contexts, data or usage patterns change
ORG	Organisation has a clear, accessible definition of “harm” in the system specific use context

B. Risk mitigation and safeguards	
Organisational/Technical	**Measure**
BOTH	Potential harms associated with the AI system (direct and indirect) have been identified and documented before deployment of tool
ORG	Harm assessments consider multiple dimensions of impact (e.g. safety, dignity, fairness etc)
BOTH	Risks are reviewed not only at design stage but also when contexts, data or usage patterns change
ORG	Organisation has a clear, accessible definition of "harm" in the system specific use context
C. Human responsibility and intervention	
Organisational/Technical	**Measure**
BOTH	Human staff are aware of duties to intervene when the tool produces outputs that may cause harm
ORG	Clear guidance exists on when and how human intervention should occur in response to harmful or unsafe outcomes
ORG	Responsibility for monitoring harm indicators and responding to them is explicitly assigned
D. Monitoring, learning and continuous improvement	
Organisational/Technical	**Measure**
TECH	Ongoing monitoring track indicators related to harm and unintended consequences over an extended time period
BOTH	Feedback from affected users or stakeholders is systematically collected and reviewed
ORG	Identified harms, near-misses or emerging risks lead to documented corrective actions and system improvements
ORG	Risk mitigation practices are periodically reassessed

Principle 8: Privacy, Consent and Data Protection

Purpose: To ensure that AI systems respect individuals rights to privacy and data protection, consent is appropriately managed and that personal data is processed lawfully and securely.

A. Lawful, transparent, and informed data use	
Organisational/Technical	**Measure**
ORG	The organisation clearly defines and documents the purposes for which personal data is processed
ORG	Individuals are informed in clear and accessible language about what data is collected and how it is used
ORG	Consent mechanisms are clearly separate to ensure appropriate meaningful engagement
ORG	Individuals can refuse or withdraw consent without unjustified negative consequences
B. Data minimisation and protection	
Organisational/Technical	**Measure**
BOTH	Only data necessary for the stated purpose is collected and processed
TECH	Technical measures are in place to protect data from unauthorised access, loss or misuse
ORG	Data retention periods are defined, documented and enforced

ORG	Sensitive data is subject to additional safeguards appropriate to the associated risk level
C. Rights, access and accountability	
Organisational/Technical	**Measure**
ORG	Clear processes exist for individuals to exercise their data protection rights (e.g., access, correction, deletion)
BOTH	Responsibility for handling data protection issues and incidents
ORG	Data processing activities are documented to support accountability and auditing review
ORG	Data protection practices are reviewed following incidents, complaints or regulatory changes

Principle 9: Transparency and Explainability

Purpose: To ensure that AI systems and their outputs are understandable, traceable and open to scrutiny but those who use them, are affected by them or are responsible for their governance.

A. Openness about tool use and purpose	
Organisational/Technical	**Measure**
ORG	There is clear guidance on the existence, purpose and intended use of the AI tool
ORG	Information about why the tool is used and how it may affect individuals is made available in accessible non-technical language
ORG	Responsibility for the tool/system governance and oversight is clearly communicated
ORG	Obligations are reviewed whenever the tools scope or use changes
B. Accessibility of information	
Organisational/Technical	**Measure**
BOTH	Information about the tools functioning, inputs, outputs and safeguards is accessible to all relevant users
ORG	Communication channels and formats are adapted to different roles, languages, digital literacy and accessibly needs
ORG	Users know where to find information and whom to contact with concerns about the tool
BOTH	Transparency is communicated clearly in non-technical language and without intent to overwhelm the reader
C. Explainability of outputs and decisions	
Organisational/Technical	**Measure**
TECH	The system provides explanation of outputs at a level appropriate to their use context
BOTH	Explanations enable users to understand key factors, limitations and uncertainty associated with outputs
ORG	Explanations support meaningful review, contestation or justification of AI-supported decisions

Principle 10: Non-bias, fairness and non-discrimination

Purpose: To ensure that AI systems do not create, reinforce or legitimise unjustified disparities in how individuals or groups are assessed, treated or affected. Leading to outcomes that are fair proportionate and respectful of human dignity.

A. Diversity and representativeness	
Organisational/Technical	**Measure**
BOTH	Data used to design, train or configure the tool is reviewed for representativeness across relevant groups and contexts
ORG	The organisation considered intersectional diversities across roles, locations, languages and protected characteristics
BOTH	Multiple stakeholder perspectives are included in the design and review of the tool
ORG	If there are limitations in data diversity or coverage, they are documented and communicated clearly
B. Objective and consistent assessment	
Organisational/Technical	**Measure**
BOTH	Data that is used by the tool to generate outputs is transparent, proportionate and consistent
ORG	Subjective or ad-hoc judgements are minimised in how tool outputs are interpreted or acted upon
TECH	Tool configurations and thresholds are reviewed for unintended disparity impact
ORG	Decisions influenced by AI outputs can be justified and the reasoning should be documented
ORG	The organisation avoids labelling, penalising or profiling individuals based solely on AI outputs
C. Objective and consistent assessment	
Organisational/Technical	**Measure**
TECH	The organisation monitors outputs for patterns of bias or unequal impact over time
ORG	Mechanisms exist for users to challenge, review/ correct outcomes perceived as unfair or discriminatory
BOTH	Identified bias triggers corrective action (e.g., adjustment to data or organisational practices)
D. Monitoring, review and correction	
Organisational/Technical	**Measure**
TECH	The organisation monitors outputs for patterns of bias or unequal impact over time
ORG	Mechanisms exist for users to challenge, review/ correct outcomes perceived as unfair or discriminatory
BOTH	Identified bias triggers corrective action (e.g., adjustment to data or organisational practices)

Principle 11: Accountability and Responsibility

Purpose: To ensure that responsibility for AI-supported decisions is clearly assigned, traceable and actionable throughout the tool/system lifecycle, and, that failures or harms can be investigated, addressed and remedied.

A. Responsibility allocation	
Organisational/Technical	**Measure**
ORG	Explicit responsibility for AI tool outcomes is defined and documented (e.g., ownership of decisions, approvals and incident response)
ORG	Clear escalation pathways exist for issues identified during development, or operation of the AI system
BOTH	Accountability is maintained across system updates, handover, or organisational changes (no gaps in responsibility)
ORG	Responsibilities are integrated into existing governance, quality, or risk-management processes rather than treated as ad-hoc ethical tasks
B. Traceability and auditability	
Organisational/Technical	**Measure**
TECH	All outputs, configurations and system versions are logged in a way that allows for outside reconstruction of decision making
TECH	Changes to models, data or system behaviours are documented and traceable
BOTH	Records are sufficient to support internal reviews, external audits or regulatory scrutiny
ORG	Processes created for reviewing audit findings and translating them into corrective action
C. Human Responsibility in practice	
Organisational/Technical	**Measure**
BOTH	AI-systems are positioned clearly to all as decision support tools with clear human judgement points noted
ORG	Responsibility for follow up on issues and anomalies is clear assigned
BOTH	Multiple layers of accountability have been assigned to provide effective oversight
D. Learning, remediation and improvement	
Organisational/Technical	**Measure**
ORG	Lessons from incidents, errors or near-misses are systematically captures and used to improve systems and processes
BOTH	Feedback loops exist to ensure that accountability mechanisms evolve as systems, contexts or risks change
ORG	Accountability expectations are periodically reviewed to remain aligned with organisational values, legal obligations and societal expectations

Principle 12: Human well-being

Purpose: To ensure that the system or deployment is beneficial to human well-being and avoids harms.

A. Integrate features that detect threats to well-being	
Organisational/Technical	**Measure**
TECH	Automated classifiers allow detecting topics or behaviours that indicate risks to well-being
TECH	Implement crisis detection AI that flags up possible critical situations in the health or well-being of users
B. Prevent uses that may jeopardize user well-being	
Organisational/Technical	**Measure**
TECH	Integrate keyword filtering to catch out of scope requests
TECH	Constraint training data to limit model knowledge to appropriate domains
TECH	Monitor outputs to prevent harmful responses even if filters missed request
C. Reporting of well-being impacts	
Organisational/Technical	**Measure**
BOTH	Set up reporting system and customer team interface to let users report on well-being impacts

Appendix B. Principle definitions as provided by use cases

B.1 NON-BIAS, FAIRNESS AND NON-DISCRIMINATION (UC3, UC4)

UC3: Non-bias, fairness and non-discrimination
Definition
Four components are considered within this principle: diversity, representativeness / inclusion, objectivity, and non-stigmatising use / proportionality. Diversity is the inclusion of data and behavioural patterns that are representative of the organisation's workforce (across roles, locations, languages and protected groups), and integration of multi-stakeholder perspectives (e.g. HR, IT/security, worker representatives) in the design, monitoring and deployment of phishing-risk models. Representativeness / inclusivity is the extent to which the system's risk signals, features, and interventions reflect the realities of different employee groups (roles, locations, contract types, digital literacy levels, accessibility needs), and are designed so that all users can understand, access and benefit from the system on equal terms. Objectivity is the use of transparent, evidence-based and standardised criteria to assess phishing vulnerability, minimising subjective judgments or ad-hoc decisions in how risk signals are generated, aggregated and interpreted across employee groups. Non-stigmatising use / proportionality is the use of vulnerability scores in ways that are proportionate to the security objective and avoid labelling or penalising individuals or groups, focusing on support and risk reduction rather than blame.
Relevance for this use case
Diversity is essential to avoid systematically over- or under-estimating risk for specific groups (e.g. by role, location, language, gender or contract type). Ensuring diverse data and perspectives helps prevent discriminatory patterns in alerts, coaching actions or HR-relevant decisions, and aligns the system with EU AI Act and GDPR principles on fairness, purpose limitation and non-discrimination. Representativeness is important for systems that measure vulnerability to cyberattacks such as phishing because phishing exposure and digital behaviours vary significantly across roles, seniority, locations and digital skills. Ensuring that risk scores, thresholds and training/coaching content are understandable and usable for all employees supports non-discrimination, mitigates disparate impact in HR-relevant decisions, and aligns such systems with the EU AI Act fairness requirements and GDPR principles of fairness and data minimisation. Vulnerability assessments can influence how employees are perceived, prioritised for training, or flagged as "high risk". Therefore, relying on objective criteria is essential to avoid arbitrary or biased treatment. Clear, documented and auditable methods for scoring and segmenting users reduce the risk of discrimination, support contestability of decisions, and align with the EU AI Act's requirements on robustness and governance, as well as GDPR principles of fairness, accountability and transparency in automated processing. If vulnerability scores are used to stigmatise certain employees or groups (e.g. "unreliable", "untrustworthy"), this can create unfair treatment, workplace tension and potential discrimination. Anchoring the system in proportional, non-punitive use ensures that outputs serve security and resilience goals while respecting dignity and equal treatment. This is consistent with the AI Act's risk-based and fundamental-rights-oriented approach, and with GDPR principles such as purpose limitation, fairness and data minimisation in processing employee data.

Source(s) that suggest inclusion of this principle
Kavvadias, A., & Kotsilieris, T. (2025). Understanding the Role of Demographic and Psychological Factors in Users' Susceptibility to Phishing Emails: A Review. Applied Sciences, 15(4), 2236. https://doi.org/10.3390/app15042236. Frank L. Greitzer, Wanru Li, Kathryn B. Laskey, James Lee, and Justin Purl. 2021. Experimental Investigation of Technical and Human Factors Related to Phishing Susceptibility. Trans. Soc. Comput. 4, 2, Article 8 (June 2021), 48 pages. https://doi.org/10.1145/3461672. Diaz, A., Sherman, A. T., & Joshi, A. (2020). Phishing in an academic community: A study of user susceptibility and behavior. Cryptologia, 44(1), 53-67. Monsoro, N., Martinie, C., Palanque, P., Saubanère, T. (2025). A Systematic Task and Knowledge-Based Process to Tune Cybersecurity Training to User Learning Groups: Application to Email Phishing Attacks. In: Clarke, N., Furnell, S. (eds) Human Aspects of Information Security and Assurance. HAISA 2024. IFIP Advances in Information and Communication Technology, vol 721. Springer, Cham. https://doi.org/10.1007/978-3-031-72559-3_12. Microsoft Design. (2025). Secure by Design: A UX Toolkit. Microsoft. Retrieved from https://microsoft.design/articles/secure-by-design-a-ux-toolkit/ Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. Big Data & Society, 8(1), 2053951720983865. Usman, Q., & Jackson, M. (2022). Ethical AI in Cybersecurity: Addressing Bias and Fairness in Automated Threat Detection Systems. Bahangulu, J. K., & Owusu-Berko, L. (2025). Algorithmic bias, data ethics, and governance: Ensuring fairness, transparency and compliance in AI-powered business analytics applications. World J Adv Res Rev, 25(2), 1746-63. Capasso M, Arora P, Sharma D, Tacconi C. On the Right to Work in the Age of Artificial Intelligence: Ethical Safeguards in Algorithmic Human Resource Management. Business and Human Rights Journal. 2024;9(3):346-360. doi:10.1017/bhj.2024.26.
Any disagreements between use partners on inclusion of this principle?
None

UC4: Non-bias, fairness and non-discrimination
Definition
Non-bias, fairness, and non-discrimination ensure that AI systems treat all individuals and groups with equal respect and dignity, free from prejudice or unequal impact. They require equality and impartiality in decision-making and outcomes, so no demographic group is unfairly advantaged or disadvantaged. Representation and inclusivity demand that diverse voices, languages, and cultural contexts are recognised and accurately reflected in training data and system design. Finally, transparency of criteria gives users the right to understand how and why decisions are made, fostering accountability and trust. Together, these principles promote equitable, just, and socially responsible AI systems.
Relevance for this use case
When viewing potentially sensitive and contentious data, it is vital that the tools are not analysing or creating content that could be biased or discriminatory, which could unintentionally lead to further marginalisation, polarisation and persecution for communities. In the execution of ensuring that non-bias, fairness and non-discrimination is happening in AI tools, training datasets should include varied linguistic and cultural expressions to prevent further marginalisation or discrimination of particularly underrepresented communities. There must be clear explanations of *why* content is flagged, ensuring that that the tools are not disproportionately flagging material from any demographic group and finally there must be the ability to appeal to decision to flag content.

Source(s) that suggest inclusion of this principle
Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516. Maronikolakis, A., Baader, P., & Schütze, H. (2022). Analyzing hate speech data along racial, gender and intersectional axes. arXiv preprint arXiv:2205.06621. Peterson-Salahuddin, C. (2024). Repairing the harm: Toward an algorithmic reparations approach to hate speech content moderation. Big Data & Society, 11(2), 20539517241245333. Gonçalves, J., Weber, I., Masullo, G. M., Torres da Silva, M., & Hofhuis, J. (2023). Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. new media & society, 25(10), 2595-2617. Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. Science and engineering ethics, 26(6), 3333-3361. Hayes, P., van de Poel, I., & Steen, M. (2023). Moral transparency of and concerning algorithmic tools. AI and Ethics, 3(2), 585-600.
Any disagreements between use partners on inclusion of this principle?
None

B.2 AUTONOMY (UC5, UC6[7])

UC5: Autonomy/User Agency
Definition
Autonomy and agency as ethical principles ensure that users retain meaningful control over their interactions with AI systems and the data these systems process. In the context of habit formation, human autonomy and agency refers to users' fundamental right to make their own decisions about behaviour change, maintain control, over their personal development goals, and avoid manipulation or coercion by AI systems. It requires informed consent, meaning users must be able to make voluntary, well-informed choices based on transparent terms and conditions—and must be free to withdraw or refuse consent without adverse consequences. System customization supports user autonomy by allowing individuals to personalise and adapt the AI application within clearly defined and safe boundaries, reflecting their preferences while maintaining ethical safeguards. Finally, transparency and user understanding demand that information about data use and interaction monitoring is communicated clearly and accessibly. Together, these elements empower users to engage with AI systems confidently and responsibly, preserving personal freedom and decision-making authority.
Relevance for this use case
Autonomy and agency are central to the ethical use of personalized AI systems, as these applications rely on users expressing personal preferences and giving consent to specific types of content. Users must clearly understand the system's boundaries and moderation rules while retaining the ability to make informed choices about their engagement and personal limits to avoid potentially harmful interactions. Those seeking customized AI experiences value the freedom to express themselves and shape their interactions according to individual needs and expectations. Those seeking habit modification want to remain self-determined and benefit from a genuinely supportive rather than directive assistant. The provider, therefore, faces the ongoing challenge of balancing user autonomy with safety, ensuring that legitimate self-expression is supported while preventing harmful behaviour or attempts to bypass safeguards, such as jailbreaking. Maintaining this balance upholds user trust, protects well-being, and preserves meaningful human control in AI-mediated environments.
Source(s) that suggest inclusion of this principle
EU AIA Art. 50; Definition under Art 3 (59). Xi Yang and Marco Aurisicchio. 2021. Designing Conversational Agents: A Self-Determination Theory Approach. In CHI Conference on Human Factors in Computing Systems (CHI '21), May 08–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 16

[7] No definition available for UC6.

pages. https://doi.org/10.1145/3411764.3445445. EU AIA Art 50; Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019, May). Guidelines for human-AI interaction. In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1-13).
Any disagreements between use partners on inclusion of this principle?
None

B.3 ACCOUNTABILITY AND RESPONSIBILITY (UC1, UC4, UC6)

UC1: Accountability and responsibility
Definition
Accountability and responsibility require AI systems to be developed, deployed and governed in ways that ensure human oversight, auditability, and clear allocation of liability. This means that high-risk AI systems (as defined in the EU AI Act) must be auditable, with accessible, consistent and up-to-date documentation, with logging and records that enable independent verification and attribution of decisions and outcomes. AI must function as a tool under human agency and control, serving people and respecting human dignity and autonomy through meaningful oversight. Liability must be clearly defined so that manufacturers and, where applicable, healthcare providers remain accountable for system safety and performance, ensuring that patients retain the right to recourse in the event of harm.
Relevance for this use case
Accountability and in turn who is responsible for ensuring accountability is central to the application of AI tools in healthcare. Auditability forms a vital component of the EU AI Act for high-risk medical AI, allowing automated reports to be traced and reviewed. Auditability supports accountability and fosters trust among clinicians and patients when AI is involved. Human oversight is also vital to ensure that AI assists in improving clinical outcomes as opposed to becoming a substitute for professional judgement. Clinicians remain the final decision maker and oversight ensures that AI tools remain subject to clinical review, therefore preventing over reliance and safeguarding patient safety. Finally, liability holds relevance due to its role in determining who is legally responsible when AI errors cause harm. In high-risk settings clear liability rules are imperative to safeguard patients, clarify accountability and maintain trust in the usage of AI tools.
Source(s) that suggest inclusion of this principle
Hartmann, D., De Pereira, J. R. L., Streitbörger, C., & Berendt, B. (2025). Addressing the regulatory gap: moving towards an EU AI audit ecosystem beyond the AI Act by including civil society. *AI and Ethics*, *5*(4), 3617-3638. >> Li, Y., & Goel, S. (2024). Making it possible for the auditing of ai: A systematic review of ai audits and ai auditability. Information Systems Frontiers, 1-31. Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024, June). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2495-2507). >>> Enqvist, L. (2023). 'Human oversight in the EU artificial intelligence act: what, when and by whom?. Law, Innovation and Technology, 15(2), 508-535. Bottomley, D., & Thaldar, D. (2023). Liability for harm caused by AI in healthcare: an overview of the core legal concepts. *Frontiers in Pharmacology*, *14*, 1297353.
Any disagreements between use partners on inclusion of this principle?
Liability in the context of AI in healthcare remains a grey area. Much will depend on how provisions of the EU AI Act are implemented and interpreted at the national level, and current enforcement roadmaps suggest a relatively gradual approach. This is especially challenging for high-risk systems: at present, there is no established standard in Europe for what a fully compliant high-risk medical AI device should look like. As a result, liability is largely outside our direct control. For now, the most

realistic path is to follow evolving codes of practice, while recognizing that clear technical measures cannot yet be mapped with confidence.

Accountability and responsibility
Definition
Accountability and responsibility ensure that AI systems operate under meaningful human oversight, with humans remaining "in the loop" for sensitive or ambiguous cases to ensure context-aware and ethical decision-making. They require auditability and evaluation, allowing independent regulators or external experts to assess system performance, fairness, and compliance through transparent records and documentation. Equally important is responsiveness—the obligation to act swiftly when issues such as bias or unintended harm are identified, including updating, retraining, or adjusting the system. Together, these principles ensure that AI remains transparent, correctable, and aligned with human values, ethical standards, and legal accountability
Relevance for this use case
The outcomes of the utilisation of AI tools such as text and media-based analysis can directly impact already marginalised communities. These AI models must not act as a "black box" without human review for context-heavy, often culturally nuanced judgements. The AI tools should allow third-party audits which would enable external checks on fairness and compliance, and there must be a willingness to act when problems are identified.
Source(s) that suggest inclusion of this principle
Alkiviadou, N. (2022). ARTIFICIAL INTELLIGENCE AND ONLINE HATE SPEECH MODERATION. Sur: Revista Internacional de Derechos Humanos, 19(32).Gier-Reinartz, N. R., Zimmermann-Janssen, V. E., & Kenning, P. (2023). AI-Assisted Hate Speech Moderation—How Information on AI-Based Classification Affects the Human Brain-In-The-Loop. In NeuroIS Retreat (pp. 45-56). Cham: Springer Nature Switzerland. Hee, M. S., Sharma, S., Cao, R., Nandi, P., Nakov, P., Chakraborty, T., & Lee, R. (2024). Recent advances in online hate speech moderation: Multimodality and the role of large models. Findings of the Association for Computational Linguistics: EMNLP 2024, 4407-4419. Balendra, S. (2025, January). Meta's AI moderation and free speech: Ongoing challenges in the Global South. In Cambridge Forum on AI: Law and Governance (Vol. 1, p. e21). Cambridge University Press. Alkiviadou, N. (2022). ARTIFICIAL INTELLIGENCE AND ONLINE HATE SPEECH MODERATION. Sur: Revista Internacional de Derechos Humanos, 19(32). Akhgar, B., Bayerl, P.S., Bailey, K., Dennis, R., Gibson, H., Heyes, S., Lyle, A., Raven, A., & Sampson, F. (2022b) AP4AI Framework Blueprint. Available at: https://www.ap4ai.eu/node/14; Akhgar, B., Bayerl, P.S., Bailey, K., Dennis, R., Heyes, S., Lyle, A., Raven, A., Sampson, F., & Gercke, M. (2022a) AP4AI Report on Expert Consultations. Available at: https://www.ap4ai.eu/node/6
Any disagreements between use partners on inclusion of this principle?
None

UC6: Accountability and responsibility
Definition
Accountability means that individuals and organisations must take responsibility for their decisions, actions, and their consequences.
Relevance for this use case
In biomedical ethics, it ensures that healthcare professionals, researchers, and institutions are answerable to patients, peers, regulatory bodies, and society. It fosters trust, integrity, and safety in medical practice and research.
Source(s) that suggest inclusion of this principle
Discussion between partners; Hoek et al. J Med Ethics 2025;51:481–486. For the point about curating data of loved ones, see: Fabry RE , Alfano M. The affective scaffolding of grief in the digital age: the case of deathbots. Topoi (Dordr) 2024;1–13. For the risk of overattachment see previous studies on griefbots: Krueger J , Osler L . Communing with the dead online: chatbots, grief, and continuing bonds. J Conscious Stud 2022;29:222–52. & Xygkou A , Siriaraya P , Covaci A . "The "conversation" about loss: understanding how chatbot technology was used in supporting people in grief". CHI '23; Hamburg Germany, April 19, 2023:1–15.
Any disagreements between use partners on inclusion of this principle?
Yes, because the article by Hoek et al. about deepfake therapy is focused on AI; while the stakeholders think that the person behind the laptop and the specific clinical setting are more important than the technology. Their work is focused on evaluating deepfake therapy in the PTSD setting which is a very different dynamic than grief therapy. People who grieve cannot go on because they do not want to; while with PTSS you want to let go but cannot. The two issues are totally different.

B.4 ROBUSTNESS AND RELIABILITY (UC2)

UC2: Robustness, safety and reliability, with special focus on accountability and traceability of safety decisions
Definition
Accountability and traceability of safety decisions refer to the ethical, technical, and organisational mechanisms that ensure every safety-related decision - whether made by humans, AI systems, or a combination of both – is attributable, explainable, and verifiable throughout the system lifecycle. They guarantee that the origin, rationale, data, and authority behind each safety judgement can be reconstructed and audited, enabling responsibility to be clearly assigned and corrective action to be taken when necessary.
Relevance for this use case
In semi-automated AI-supported safety analysis, traceability ensures that engineers can see the data and logic behind AI recommendations. Accountability ensures that a qualified human ultimately validates or rejects those outputs before they enter the safety case. Ethically, they prevent '"responsibility gaps" and automation bias, supporting informed oversight + public trust.
Source(s) that suggest inclusion of this principle
EU AI Act (2024, Art. 12–15), Ethics Guidelines for Trustworthy AI, ISO 26262:2018, ISO/PAS 21448:2022 (SOTIF), UNECE R155 / R156, Rowe, F., Jeanneret Medina, M., Benoit Journé, Coëtard, E., & Myers, M. (2023). Understanding responsibility under uncertainty: A critical and scoping review of autonomous driving systems. Journal of Information Technology, 39(3), 587-615, Tanja Pavleska, Massimiliano Masi, Giovanni Paolo Sellitto, Helder Aranha, Architecture-based governance for secure-by-design Cooperative Intelligent Transport Systems, Vehicular Communications, Volume 55, 2025.
Any disagreements between use partners on inclusion of this principle?
No disagreements; consensus achieved.

B.5 TRANSPARENCY AND EXPLAINABILITY (UC1, UC3)

UC1: Transparency and explainability
Definition
Transparency and explainability ensure that AI systems are understandable, traceable, and justifiable to all relevant stakeholders. They require accessibility, i.e., that clear information about an AI system's design, training data, and functioning is available in a form that can be meaningfully interpreted and used, not only by experts but also by affected individuals. High-risk AI systems must include human–machine interfaces that allow effective oversight and control throughout their use. Finally, justifiability ensures that AI outputs can be ethically, clinically, and legally defended, aligning decisions with professional standards and patient values to support trust and accountability in healthcare.
Relevance for this use case
Transparency is only meaningful when information can be practically used, understood and justified by different stakeholders. Without accessibility, transparency risks remaining abstract, limiting both safe adoption in daily practice and effective oversight of AI tools. Explainability holds direct relevance due to the need to have design features that allow humans to supervise, interpret and intervene in safeguarding against automation bias, misuse and ensure responsibility for outcomes. Finally, the operations and working of AI systems should be based on valid, defensible reasons that can be ethical justified.
Source(s) that suggest inclusion of this principle
Fehr, J., Citro, B., Malpani, R., Lippert, C., & Madai, V. I. (2024). A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, *6*, 1267290; Vo, V., Chen, G., Aquino, Y. S. J., Carter, S. M., Do, Q. N., & Woode, M. E. (2023). Multi-stakeholder preferences for the use of artificial intelligence in healthcare: A systematic review and thematic analysis. Social Science & Medicine, 338, 116357. Marey, A., Arjmand, P., Alerab, A. D. S., Eslami, M. J., Saad, A. M., Sanchez, N., & Umair, M. (2024). Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology. *Egyptian Journal of Radiology and Nuclear Medicine*, *55*(1), 183. >> Nouis, S. C., Uren, V., & Jariwala, S. (2025). Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making: a qualitative study of healthcare professionals' perspectives in the UK. BMC Medical Ethics, 26(1), 89. More technical: Houssein, E. H., Gamal, A. M., Younis, E. M., & Mohamed, E. (2025). Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review. Cluster Computing, 28(7), 469. Hildt E. What Is the Role of Explainability in Medical Artificial Intelligence? A Case-Based Approach. Bioengineering (Basel). 2025 Apr 2;12(4):375; Muralidharan, A., Savulescu, J. & Schaefer, G.O. AI and the need for justification (to the patient). Ethics Inf Technol 26, 16 (2024). Also: https://blogs.bmj.com/medical-ethics/2022/03/02/three-observations-about-justifying-ai/?utm_source=chatgpt.com; More Legal: Malgieri, G., & Pasquale, F. (2024). Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. Computer Law & Security Review, 52, 105899.
Any disagreements between use partners on inclusion of this principle?
Explainable AI (XAI) can be defined in different ways. At the model level, it refers to methods that show how the system works internally (its reasoning and logic) which is relatively standard practice. At the post hoc level, explainability focuses on specific cases, for example: given a CT scan, why does the system suggest that this patient has breast cancer? This form of explanation is more complex technically to achieve, but directly relevant to clinical use and liability, since responsibility often requires explanations at the inference level. Put differently, the technical definition of explainability focuses on features and system behaviour, while the clinical perspective emphasizes justification in practice. In the AIOLIA project, with its focus on industry partners and stakeholders, our discussions highlighted that the clinical and patient perspectives on explainability are still largely missing. Justifiability is difficult to translate into concrete technical measures, as it concerns whether users can meaningfully scrutinize AI outputs. One open question is whether AI systems could actively support

clinicians in justifying their decisions to colleagues or patients. This is highly relevant in healthcare, where clinicians must defend treatment choices and ensure that AI-supported decisions can be explained in ethically and professionally acceptable ways. Unlike explainability, which focuses on making the AI's reasoning understandable, justifiability asks whether those reasons are adequate grounds for clinical action. Industry partners argued that much depends on the level of granularity. For example, in neural networks trained on images, it may be possible to justify outputs by showing why certain features are included or excluded. However, at the single-patient level, such as interpreting one unique medical image, this becomes virtually impossible. Unlike large datasets where features can be compared, there is only one case, and subtle differences may not provide a clear or defensible justification.

UC3: Transparency and explainability
Definition
Transparency and explainability is considered split into three components: openness, accessibility / access to information, and documentation, traceability, and auditability. Openness is the degree to which the organisation provides clear, accessible and non-technical information about the existence, purpose and main functioning of the phishing-vulnerability measurement system, including what data it uses, how results may affect employees, and who is responsible for its governance. Accessibility / access to information means both ensuring that all employees can easily access information about the phishing-vulnerability measurement system, its purpose, data inputs, outputs, rights, safeguards and potential impacts, and that such information is easy to understand by all employees. Both entail providing information through channels and formats adapted to different roles, languages, digital literacy levels and accessibility needs. Documentation, traceability, and auditability includes the availability of clear, up-to-date documentation and logs describing the design, data sources, model versions, configuration changes and decision logic of the phishing-vulnerability measurement system, as well as traceable records that allow reconstruction and external review of how specific vulnerability scores or decisions were produced.
Relevance for this use case
A system that measures individual vulnerability to phishing can directly affect employees' trust, perceived autonomy and sense of fairness, especially when outputs may feed into training plans, performance discussions or HR-relevant decisions. Openness about the system's purpose, data sources, safeguards and governance reduces fears of hidden surveillance or punitive use, enables informed participation, and supports accountability. It is also closely aligned with transparency obligations in the EU AI Act for high-risk AI systems, and with GDPR principles on transparency, information duties towards data subjects and fair processing of employee data. Vulnerability assessments relate directly to individual employees. Therefore, all users, regardless of role, location, language or technical ability, must have equal access to information about how the system affects them. Clear and accessible communication supports informed participation, reduces misconceptions or fears of surveillance, and enables users to exercise their GDPR rights (access, rectification, objection, information). It is also consistent with the AI Act's emphasis on user-facing transparency and with fundamental-rights safeguards requiring that information be understandable, available and actionable for all impacted individuals. Organisations must be able to explain and demonstrate how the system works and why a given outcome occurred. Robust documentation and traceability enable internal and external audits, support investigations of potential bias or errors, and provide the basis for meaningful contestation by affected individuals. This is aligned with the EU AI Act requirements on technical documentation, logging and oversight for high-risk AI systems, and with GDPR principles of accountability, transparency and the ability to substantiate compliance when processing employee data.

Source(s) that suggest inclusion of this principle
Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. Science and engineering ethics, 26(6), 3333-3361. Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. Information and Software Technology, 159, 107197.Yanamala, K. K. R. (2023). Transparency, privacy, and accountability in AI-enhanced HR processes. Journal of Advanced Computing Systems, 3(3), 10-18. Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021, May). Expanding explainability: Towards social transparency in ai systems. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-19).
Any disagreements between use partners on inclusion of this principle?
None

B.6 OVER-RELIANCE AND DESKILLING (UC3)[8]

UC2: Over-reliance and deskilling
Definition
Two components are relevant to over-reliance and deskilling: dependence, and contestability and human oversight. Dependence is the degree to which security, HR or management decisions rely solely or predominantly on phishing-vulnerability scores and automated outputs, instead of combining them with human judgement, contextual information and other complementary security indicators. Contestability and human oversight necessitate clear, accessible mechanisms for employees and relevant stakeholders (e.g. HR, security, worker representatives) to review, question and correct vulnerability assessments and related decisions, supported by documented human-in-the-loop oversight for high-impact uses of the system.
Relevance for this use case
A system that measures individual vulnerability to phishing can create a false sense of precision and encourage staff to treat scores as definitive truth about an employee's behaviour or reliability. Over-dependence on automated outputs increases the risk of unfair treatment, misallocation of training or controls, and blind spots where the system is less accurate (e.g. specific roles or contexts). Managing dependence, by keeping human oversight, contextual review and clear usage boundaries, aligns with the AI Act requirements to avoid inappropriate reliance on high-risk AI systems, and with GDPR principles of fairness, proportionality and meaningful human involvement in decisions affecting employees. Vulnerability scores and risk flags can influence how employees are treated (e.g. training assignments, performance discussions, escalation to HR). Therefore, affected individuals should be able to understand and contest outcomes that they perceive as inaccurate or unfair. Human oversight prevents blind reliance on automated scoring, helps detect hidden biases or data quality issues, and aligns with AI Act obligations on human oversight for high-risk AI systems as well as GDPR principles on fairness, transparency and rights related to automated decision-making.

[8] UC2 did provide definition of Transparency, not Over-reliance and deskilling, due to a shift in focus at a later stage towards deskilling. Hence, only the UC3 definition is provided here.

Source(s) that suggest inclusion of this principle
Mujtaba, D. F., & Mahapatra, N. R. (2024). Fairness in AI-driven recruitment: Challenges, metrics, methods, and future directions. arXiv preprint arXiv:2405.19699. Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: "automation bias" and "selective adherence" to algorithmic advice. Journal of Public Administration Research and Theory, 33(1), 153-169. Holzinger, A., Zatloukal, K., & Müller, H. (2025). Is human oversight to AI systems still possible?. New Biotechnology, 85, 59-62. Ottun, A. R. O., & Flores, H. (2025). Trustworthy AI in Practice: A Comprehensive Review of Human Oversight and Human-in-the-Loop Approaches. Authorea Preprints.
Any disagreements between use partners on inclusion of this principle?
None

B.7 NON-MALEFICENCE (UC1, UC6)

UC1: Non-maleficence
Definition
Non-maleficence means ensuring that AI systems in healthcare do not cause harm through inaccuracy, bias, or misuse of personal data. It requires validity and accuracy, so that outputs reliably reflect the patient's true clinical state and support safe medical decisions. AI systems must also avoid bias, maintaining fair performance across all patient groups to prevent reinforcing health inequities. Finally, privacy must be protected by processing sensitive health data lawfully, securely, and with respect for individuals' rights. Together, these principles ensure that healthcare AI supports patient safety, fairness, and trust while upholding fundamental ethical and legal standards
Relevance for this use case
Non-maleficence protects patient safety, confidentiality and challenges underrepresentation amongst certain patient groups. Without validity and accuracy, surgeons would lack trust in the system to reflect real clinical situations which would lead, e.g., to poor clinical choices and follow-up schedules for patients. It would be impossible to introduce AI automation in any radiology department without a thorough assessment of algorithm's performance on the local population and appropriate clinical settings. If training data presents biases, then certain patient groups may be underrepresented within the cohort, leading to less reliable clinical decisions or follow up suggestions for patients who present with atypical anatomies or comorbidities. Finally, data privacy underpins the ability for patient data to be utilised in the development, calibration and monitoring of AI systems. Without privacy safeguards the data from chest x-rays, CT scans and additional follow up data cannot be made available, leading to inefficient training of the algorithms, biased validation, challenging post-market surveillance and less optimal outcomes for patients.
Source(s) that suggest inclusion of this principle
Holm S. On the Justified Use of AI Decision Support in Evidence-Based Medicine: Validity, Explainability, and Responsibility. Cambridge Quarterly of Healthcare Ethics. Published online 2023:1-7. Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., ... & Naganawa, S. (2024). Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese journal of radiology*, *42*(1), 3-15. Bak, M. A., Ploem, M. C., Tan, H. L., Blom, M. T., & Willems, D. L. (2023). Towards trust-based governance of health data research. Medicine, Health Care and Philosophy, 26(2), 185-200. Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Computers in Biology and Medicine, 158, 106848.
Any disagreements between use partners on inclusion of this principle?
No disagreements

B.8 HUMAN OVERSIGHT (UC2)

UC2: Human oversight and autonomy, with additional focus controllability
Definition
Human oversight and controllability refer to the design, governance, and operational measures that ensure humans can understand, supervise, intervene in, and ultimately take responsibility for the actions and decisions of an AI-enabled system – especially in safety-critical contexts such as automotive development, testing, or driving. They ensure that AI remains a decision-support tool rather than an autonomous authority, and that humans can prevent or mitigate harm when the system behaves unexpectedly or operates outside its validated conditions.
Relevance for this use case
In this use case, human engineers use AI to identify hazards or assess risk. Oversight ensures they can verify, contest, or refine AI suggestions, preventing automation bias. This helps preserving human responsibility for safety certification and incident response, prevent over-reliance on opaque AI recommendations, and enable continuous learning while maintaining regulatory compliance.
Source(s) that suggest inclusion of this principle
EU AI Act (Art. 14), Ethics Guidelines for Trustworthy AI (HLEG, 2019), ISO 26262:2018, ISO/PAS 21448:2022 (SOTIF), UNECE R157 / R156, Suzuki Wataru (2025). Safety Analysis and Design Improvement for Semi-Automatic Train Operation (STO) in High-Speed Rail Using STPA, PhD Thesis, MIT, Cappelli, M.A., Di Marzo Serugendo, G. A semi-automated software model to support AI ethics compliance assessment of an AI system guided by ethical principles of AI. AI Ethics 5, 1357–1380 (2025), Andreas Holzinger, Kurt Zatloukal, Heimo Müller, Is human oversight to AI systems still possible?, New Biotechnology, 85, 2025.
Any disagreements between use partners on inclusion of this principle?
None reported.

B.9 SAFETY/HUMAN SAFETY (UC5)

UC5: Safety/Human Safety
Definition
Safety as an ethical principle ensures that AI systems protect users from harm and operate within secure, controlled, and ethically guided boundaries. It requires robust user protection measures to prevent the generation or dissemination of harmful, violent, or psychologically distressing content. Security mechanisms–including age verification, multi-tier content classification, and active monitoring of harmful usage patterns–must be in place to detect and prevent unsafe or violent outputs, issuing warnings or bans when necessary. Equally, human oversight remains essential: trained moderators must be able to intervene when automated systems encounter complex or borderline scenarios. This combination of technical safeguards and human judgment ensures that AI systems act responsibly, uphold user well-being, and prevent the normalization or spread of harmful material.
Relevance for this use case
Safety is a paramount ethical principle in the development and deployment of conversational AI systems, particularly those designed for personal or emotionally intimate interactions. Users often share sensitive information and engage in deeply personal exchanges, which increases the risk of exposure to harmful, disturbing, or psychologically distressing content. Ensuring robust security measures—such as monitoring, age verification, and content classification—is essential for both user protection and legal compliance. Equally, human oversight plays a critical role in safeguarding users, enabling moderators to identify and manage complex or borderline cases that automated systems cannot reliably address. Human supervision also helps detect jailbreaking attempts, evaluate edge cases, and differentiate between legitimate user preferences and harmful behaviour. In environments

where user-generated content and AI characters can be shared or reused commercially, such oversight is indispensable to prevent misuse and uphold the safety and well-being of all users
Source(s) that suggest inclusion of this principle
EU AIA recital 5 + 9; Felipe Romero Moreno (2024) Generative AI and deepfakes: a human rights approach to tackling harmful content, International Review of Law, Computers & Technology, 38:3, 297-326. EU AIA Art. 5; Chawki, M. (2025). AI Moderation and Legal Frameworks in Child-Centric Social Media: A Case Study of Roblox. *Laws*, *14*(3), 29. EU AIA Art. 14; Ethics Guidelines for Trustworthy AI (HLEG, 2019); Lena Enqvist (2023) 'Human oversight' in the EU artificial intelligence act: what, when and by whom? Law, Innovation and Technology, 15:2, 508-535.
Any disagreements between use partners on inclusion of this principle?
None

B.10 PRIVACY, CONSENT AND DATA PROTECTION (UC5)

UC5: Privacy and data protection
Definition
Privacy and data protection ensure that individuals retain control over their personal information and that AI systems handle data lawfully, transparently, and securely. Users must be given clear and understandable information about how their data—both directly provided and inferred from behaviour—will be processed, stored, and shared, enabling genuine informed consent. In line with the principle of data minimisation, only data necessary for a specific, legitimate purpose should be collected, and it must not be stored longer than required. Any third-party sharing of data must comply with established legal and ethical standards, such as the GDPR, ensuring accountability, lawful processing, and respect for user privacy throughout the AI system's lifecycle.
Relevance for this use case
Privacy is both a fundamental human right and a core ethical value that must be upheld in the design and use of AI systems. While user consent is required to collect, process, and transfer personal data—such as for payment or verification purposes—transparency in how this data is handled is essential to determining whether privacy is genuinely protected. Legal frameworks, including the GDPR, require that providers collect only data that is strictly necessary for defined purposes and handle it responsibly. In this regard, it is crucial to acknowledge the risk of inference of additional sensitive attributes beyond intentionally discloses user information. Users, in turn, expect platforms to respect their privacy and autonomy, which is crucial for maintaining trust and legitimacy. However, in complex environments involving multi-tier moderation systems and safety obligations, continuous monitoring of user interactions may be necessary.
Source(s) that suggest inclusion of this principle
Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 1, 389–399 (2019). GDPR; EU AIA Art. 10 + Recital 69; Yanamala, A. K. Y., Suryadevara, S., & Kalli, V. D. R. (2024). Balancing innovation and privacy: The intersection of data protection and artificial intelligence. *International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence*, *15*(1), 1-43.; Brown & Truby. (2022). Mending lacunas in the EU's GDPR and proposed artificial intelligence regulation, 10.2478/eustu-2022-0003. GDPR; EU AIA Art. 11, 50 + Annex IV.
Any disagreements between use partners on inclusion of this principle?
Currently extensive monitoring is conducted to fulfil the multi-tier moderation system, which requires, among other things, the processing of chat histories. It remained unclear if this is communicated transparently enough to the users.

B.11 FREEDOM OF EXPRESSION AND NON-CENSORSHIP (UC4)

UC4: Freedom of expression and non-censorship
Definition
Freedom of expression and non-censorship uphold the right of individuals to form, express, and share opinions without fear of punishment or unjust interference. This principle safeguards autonomy and agency, ensuring that diverse perspectives can be voiced openly. Any restriction on expression must follow the principle of proportionality, applying only the least restrictive measures necessary to prevent genuine harm. Equally, non-discrimination requires that all speech is treated fairly and consistently, regardless of the speaker's identity, background, or viewpoint.
Relevance for this use case
Freedom of speech/freedom of expression is a fundamental human right, recognised within the EU and internationally (Article 10 of the ECHR). In upholding that, AI systems must be trained to avoid bias that supports discrimination against or suppression of specific groups or political standpoints, they should not be able to remove of alter legitimate speech due to its unpopularity or controversy. If AI flags hate speech, misinformation, disinformation, anti-political rhetoric or harmful narratives, it should only do so when the content meets clear thresholds, and not due to perceived offence or disagreement.
Source(s) that suggest inclusion of this principle
Scanlon, T. (1972). A theory of freedom of expression. Philosophy & Public Affairs, 204-226. Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. Computer science, 7, e598. Roberts, S. T. (2019). Behind the screen. Yale University Press. Tsakyrakis, S. (2009). Proportionality: An assault on human rights?. International Journal of Constitutional Law, 7(3), 468-493. Thiago Dias Oliva, Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression, Human Rights Law Review, Volume 20, Issue 4, December 2020, Pages 607–640, https://doi.org/10.1093/hrlr/ngaa032. Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. Neurocomputing, 546, 126232. Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., & Smith, N. A. (2021). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. arXiv preprint arXiv:2111.07997. Heinrichs, B. (2022). Discrimination in the age of artificial intelligence. AI & society, 37(1), 143-154. Nascimento, F. R., Cavalcanti, G. D., & Costa-Abreu, M. D. (2025). Gender bias detection on hate speech classification: an analysis at feature-level. Neural Computing and Applications, 37(5), 3887-3905.
Any disagreements between use partners on inclusion of this principle?
None

B.12 HUMAN WELL-BEING (UC5)

UC5: Human Well-being
Definition
Human Well-being encompasses the obligation to ensure AI systems promote users' physical and psychological health, support their flourishing as individuals, and avoid causing harm through inappropriate advice, reinforcement of unhealthy patterns, or failure to recognize crisis situations. In one context studied in UC5, this includes providing genuinely helpful habit formation support while preventing the AI from encouraging harmful behaviours or failing to identify when users need professional intervention beyond AI capabilities. In another context studied in UC5, this means at least protecting users from physical or psychological harm during and as a consequence of the interaction with the AI characters.
Relevance for this use case
The purpose is explicitly to enhance user well-being through improved habits and personal development. However, this creates significant responsibility: the AI operates in domains (health behaviours, mental wellness, habit formation) where harmful advice could have serious consequences. Thus, this principle requires the partner to be genuinely helpful while
Source(s) that suggest inclusion of this principle
--
Any disagreements between use partners on inclusion of this principle?
None

Appendix C: Definition of components as provided by use cases

UC1: Validity / accuracy (from technical perspective)
Definition
Non-maleficence means ensuring that AI systems in healthcare do not cause harm through inaccuracy, bias, or misuse of personal data. It requires validity and accuracy, so that outputs reliably reflect the patient's true clinical state and support safe medical decisions. AI systems must also avoid bias, maintaining fair performance across all patient groups to prevent reinforcing health inequities. Finally, privacy must be protected by processing sensitive health data lawfully, securely, and with respect for individuals' rights. Together, these principles ensure that healthcare AI supports patient safety, fairness, and trust while upholding fundamental ethical and legal standards
Relevance for this use case
Validity and accuracy are essential in medical AI because patient safety depends on reliable outputs. In radiology, this means chest x-ray tools must correctly identify normal images and findings so that triage and safety-net functions work as intended. In vascular surgery, accuracy is needed for measurements and treatment planning so that graft choices and follow-up schedules reflect the patient's real clinical situation. In both cases, reliable performance should allow clinicians to trust the system and use it safely in daily practice.
Source(s) that suggest inclusion of this principle
Holm S. On the Justified Use of AI Decision Support in Evidence-Based Medicine: Validity, Explainability, and Responsibility. Cambridge Quarterly of Healthcare Ethics. Published online **2023**:1-7.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC1: Bias
Definition
Bias means that AI systems are developed and used in ways that avoid creating or amplifying unfair differences in performance across patient groups, so that recommendations do not cause harm or reinforce health disparities.
Relevance for this use case
Bias in medical AI threatens patient safety when systems perform unevenly across populations. In radiology, underrepresentation in training data may cause Oxipit to miss findings in certain groups, weakening its role as a safety net. In vascular surgery, the AAA decision-support tool may give less reliable graft or follow-up suggestions for patients with atypical anatomies or comorbidities.
Source(s) that suggest inclusion of this principle
Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., ... & Naganawa, S. (**2024**). Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese journal of radiology*, *42*(1), 3-15.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC1: Privacy
Definition
Privacy means that patients' health data, as special category personal data under the GDPR, must be processed lawfully and securely, with individuals retaining rights over access, use, and disclosure to safeguard their fundamental rights (GDPR)
Relevance for this use case
Privacy is essential in medical AI because these systems depend on sensitive health data for development and use. In radiology, Oxipit processes large volumes of chest x-rays, while in vascular surgery the AAA decision-support system relies on CT scans and follow-up data. Both of which emphasizing the need for safeguards to maintain confidentiality. Privacy protections must also hold (or be adaptable) over time, as regulations evolve and data are repurposed. Strong measures are needed to prevent adversarial misuse.
Source(s) that suggest inclusion of this principle
Bak, M. A., Ploem, M. C., Tan, H. L., Blom, M. T., & Willems, D. L. (**2023**). Towards trust-based governance of health data research. Medicine, Health Care and Philosophy, 26(2), 185-200. >>More technical > **Khalid**, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (**2023**). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Computers in Biology and Medicine, 158, 106848.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC1: Auditibility
Definition
Auditability means that high-risk AI systems are designed and documented to ensure traceability of processes and outputs, with logging and records that allow independent examination, verification, and allocation of accountability. (EU AI ACT)
Relevance for this use case
Auditability is central under the EU AI Act for high-risk medical AI. For Oxipit's radiology tools, it means that automated reports and triage decisions can be traced and reviewed. For Afliant's AAA decision-support system, it ensures that e.g., vessel measurements and treatment suggestions remain transparent and verifiable. Auditability in both cases supports accountability and fosters trust among clinicians and patients when AI influences clinical decisions.
Source(s) that suggest inclusion of this principle
Auditability is central under the EU AI Act for high-risk medical AI. For Oxipit's radiology tools, it means that automated reports and triage decisions can be traced and reviewed. For Afliant's AAA decision-support system, it ensures that e.g., vessel measurements and treatment suggestions remain transparent and verifiable. Auditability in both cases supports accountability and fosters trust among clinicians and patients when AI influences clinical decisions.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC1: Human oversight
Definition
Human agency and oversight means that AI systems are developed and used as a tool that serves people, respects human dignity and personal autonomy, and that is functioning in a way that can be appropriately controlled and overseen by humans. (EU AI Act)

Relevance for this use case
Human oversight is vital to ensure that AI supports rather than substitutes professional judgment. In radiology, systems like Oxipit influence how images are reported and triaged, while in vascular surgery Afliant's tools inform treatment planning. Oversight ensures that these outputs remain subject to clinical review, preventing overreliance on automation and safeguarding patient safety. It also underpins accountability and maintains clinicians' role as the final decision-makers.
Source(s) that suggest inclusion of this principle
Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (**2024**, June). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2495-2507). >>> **Enqvist**, L. (**2023**). 'Human oversight'in the EU artificial intelligence act: what, when and by whom?. Law, Innovation and Technology, 15(2), 508-535.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC1: Accessibility
Definition
Accessibility means that information about AI systems, including their development, training data, and functioning is made available and presented in a form that can be understood and used by different stakeholders, so that transparency is not only technical but also practical. (in line with transparency obligations EU AI-Act)
Relevance for this use case
Accessibility is essential in medical AI because transparency is only meaningful when information can be practically used by different stakeholders. Clinicians need AI outputs, such as risk scores or image overlays, to be available in clear formats that fit their workflows, while managers and regulators require accessible documentation on data sources, validation, and performance to ensure accountability. Without such accessibility, transparency risks remaining abstract, limiting both safe adoption in daily practice and effective oversight.
Source(s) that suggest inclusion of this principle
Fehr, J., Citro, B., Malpani, R., Lippert, C., & Madai, V. I. (2024). A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. Frontiers in Digital Health, 6, 1267290. >>>> Vo, V., Chen, G., Aquino, Y. S. J., Carter, S. M., Do, Q. N., & Woode, M. E. (2023). Multi-stakeholder preferences for the use of artificial intelligence in healthcare: A systematic review and thematic analysis. Social Science & Medicine, 338, 116357.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC1: Explainability
Definition
High-risk AI systems should be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use (AI Act).
Relevance for this use case
Requires design features that allow humans to supervise, interpret, and (if necessary) intervene to safeguard against automation bias, misuse, and ensure responsibility for outcomes. In radiology, features such as heatmaps or overlays show why Oxipit flagged an x-ray, enabling radiologists to cross-check the system's outputs. In vascular surgery, evidence tables that link vessel measurements to graft suggestions allow surgeons to critically assess whether the recommendation is appropriate.

Source(s) that suggest inclusion of this principle
Marey, A., Arjmand, P., Alerab, A. D. S., Eslami, M. J., Saad, A. M., Sanchez, N., & Umair, M. (2024). Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology. Egyptian Journal of Radiology and Nuclear Medicine, 55(1), 183. >> Nouis, S. C., Uren, V., & Jariwala, S. (2025). Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making: a qualitative study of healthcare professionals' perspectives in the UK. BMC Medical Ethics, 26(1), 89. >> More technical> Houssein, E. H., Gamal, A. M., Younis, E. M., & Mohamed, E. (2025). Explainable artificial intelligence for medical imaging systems using deep learning: a comprehensive review. Cluster Computing, 28(7), 469.
Any disagreements between use partners on inclusion of this principle?
Discussion: Explainable AI (XAI) can be defined in different ways. At the model level, it refers to methods that show how the system works internally—its reasoning and logic—which is relatively standard practice. At the post hoc level, explainability focuses on specific cases, for example: given a CT scan, why does the system suggest that this patient has breast cancer? This form of explanation is more complex but directly relevant to clinical use and liability, since responsibility often requires explanations at the inference level. Put differently, the technical definition of explainability focuses on features and system behaviour, while the clinical perspective emphasizes justification in practice. In the AIOLIA project, with its focus on industry partners and stakeholders, our discussions highlighted that the clinical and patient perspectives on explainability are still largely missing.

UC1: Justifiability
Definition
Justifiability means that AI systems and their outcomes in healthcare are supported by reasons that align with ethical, clinical, legal, and patient values, so that their usage can be morally, professionally, and legally defended.
Relevance for this use case
AI system outputs and decisions should be based on valid, defensible reasons that can be ethically (and clinically) justified.
Source(s) that suggest inclusion of this principle
Hildt E. What Is the Role of Explainability in Medical Artificial Intelligence? A Case-Based Approach. Bioengineering (Basel). **2025** Apr 2;12(4):375. 2). >>> **Muralidharan**, A., Savulescu, J. & Schaefer, G.O. AI and the need for justification (to the patient). Ethics Inf Technol 26, 16 (**2024**). Also https://blogs.bmj.com/medical-ethics/2022/03/02/three-observations-about-justifying-ai/?utm_source=chatgpt.com >>>> More Legal> **Malgieri**, G., & Pasquale, F. (**2024**). Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. Computer Law & Security Review, 52, 105899.
Any disagreements between use partners on inclusion of this principle?
DISCUSSION: Justifiability is difficult to translate into concrete technical measures, as it concerns whether users can meaningfully scrutinize AI outputs. One open question is whether AI systems could actively support clinicians in justifying their decisions to colleagues or patients. This is highly relevant in healthcare, where clinicians must defend treatment choices and ensure that AI-supported decisions can be explained in ethically and professionally acceptable ways. Unlike explainability, which focuses on making the AI's reasoning understandable, justifiability asks whether those reasons are adequate grounds for clinical action.Industry partners argued that much depends on the level of granularity. For example, in neural networks trained on images, it may be possible to justify outputs by showing why certain features are included or excluded. However, at the single-patient level, such as interpreting one unique medical image, this becomes virtually impossible. Unlike large datasets where features can be compared, there is only one case, and subtle differences may not provide a clear or defensible justification.

UC2: Technical robustness and resilience
Definition
Technical robustness and resilience refer to the ability of an AI system to operate reliably, securely, and predictably under both normal and adverse conditions, and to withstand, detect, and recover from errors, perturbations, or malicious attacks that could compromise safety.
Relevance for this use case
Technical robustness ensures that the analyses of the AI system are reliable, reproducible, and traceable, so that human experts can safely base critical decisions on them. If the AI system provides erroneous or misleading analysis (e.g., missed fault paths, false safety correlations), this can directly undermine safety certificationand create ethical responsibility gaps. In thsi sense, technical robustness and resilience ensure that the AI system strengthens, rather than weakens, human judgment.
Source(s) that suggest inclusion of this principle
ISO 26262, ISO/PAS 21448 (SOTIF), EU AI Act (Art. 15), HLEG AI Ethics Guidelines (2019), Wäschle M, Thaler F, Berres A, Pölzlbauer F, Albers A. A review on AI Safety in highly automated driving. Front Artif Intell. 2022 Oct 3;5:952773, Nicola Tamascelli, Alessandro Campari, Tarannom Parhizkar, Nicola Paltrinieri, Artificial Intelligence for safety and reliability: A descriptive, bibliometric and interpretative review on machine learning, Journal of Loss Prevention in the Process Industries, Volume 90, 2024.
Any disagreements between use partners on inclusion of this principle?
No disagreements; consensus achieved.

UC2: Reliability through lifecycle testing and monitoring
Definition
Reliability through lifecycle testing and monitoring refers to the continuous assurance that an AI-based or automated system performs its intended safety and functional tasks consistently, accurately, and predictably across all phases of its lifecycle — from design and validation to deployment, operation, and maintenance. It integrates systematic testing, validation, and in-operation monitoring to detect performance degradation, failures, or unsafe behaviours early, enabling timely corrective actions and sustained compliance with safety standards.
Relevance for this use case
In automotive engineering, safety assurance does not end at validation — new vehicle configurations, software updates, and changing operational conditions constantly affect risk profiles. An AI system helping to identify hazards, perform risk classification, or model fault trees must therefore remain reliable throughout its lifecycle, not only at initial deployment. It should also remain reiliable when adding new data. This is also necessary for supporting traceability, certification and auditability.
Source(s) that suggest inclusion of this principle
ISO 26262:2018, ISO/PAS 21448:2022 (SOTIF), UNECE R156, EU AI Act (Art. 15), Ethics Guidelines for Trustworthy AI (HLEG, 2019), Shailesh Hegde, Dinesh Selvaraj, Josie Esteban Rodriguez Condia, Nicola Amati, Carla Chiasserini, Francesco Deflorio, and Matteo Sonza Reorda. 2025. Early Reliability Assessment of AI-based Automotive Systems. ACM Trans. Internet Things Just Accepted, R. Schnitzer, L. Kilian, S. Roessner, K. Theodorou and S. Zillner, "Landscape of AI Safety Concerns - A Methodology to Support Safety Assurance for AI-Based Autonomous Systems," 2024 8th International Conference on System Reliability and Safety (ICSRS), Sicily, Italy, 2024, pp. 501-510, https://semiengineering.com/the-importance-of-safety-analysis-in-automotive-systems-engineering/
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC2: Fairness in safety assurance
Definition
Fairness in safety assurance refers to the ethical and procedural commitment to ensure that safety evaluation methods, criteria, and decisions are applied consistently, transparently, and without unjust bias toward any group of users, contexts of operation, or system components. It means that all stakeholders — designers, operators, passengers, pedestrians, and other road users — receive an equitable level of protection, consideration, and accountability throughout the system's lifecycle, regardless of demographic, geographic, or technological differences.
Relevance for this use case
Semi-automated safety systems rely on AI-assisted data analysis (e.g., risk classification, failure pattern recognition). If these algorithms are biased or inconsistently applied, some operating contexts or user groups could face higher residual risks. Ethically, this aligns with the principles of justice, non-discrimination, and transparency set out in the EU AI Act and the Ethics Guidelines for Trustworthy AI (HLEG, 2019).
Source(s) that suggest inclusion of this principle
European Commission High-Level Expert Group on AI. (2019). Ethics Guidelines for Trustworthy AI, ISO/PAS 21448:2022, ISO 26262:2018, Filip Cano Cordoba (2025). Towards Responsible AI: Advances in Safety, Fairness, and Accountability of Autonomous Systems. PhD Thesis, University of Graz, Dimitrios I. Tselentis, Eleonora Papadimitriou, Pieter van Gelder, The usefulness of artificial intelligence for safety assessment of different transport modes, Accident Analysis & Prevention, Volume 186, 2023.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC2: Human oversight and controllability
Definition
Human oversight and controllability refer to the design, governance, and operational measures that ensure humans can understand, supervise, intervene in, and ultimately take responsibility for the actions and decisions of an AI-enabled system — especially in safety-critical contexts such as automotive development, testing, or driving. They ensure that AI remains a decision-support tool rather than an autonomous authority, and that humans can prevent or mitigate harm when the system behaves unexpectedly or operates outside its validated conditions.
Relevance for this use case
In this use case, human engineers use AI to identify hazards or assess risk. Oversight ensures they can verify, contest, or refine AI suggestions, preventing automation bias. This helps preserving human responsibility for safety certification and incident response, prevent over-reliance on opaque AI recommendations, and enable continuous learning while maintaining regulatory compliance.
Source(s) that suggest inclusion of this principle
EU AI Act (Art. 14), Ethics Guidelines for Trustworthy AI (HLEG, 2019), ISO 26262:2018, ISO/PAS 21448:2022 (SOTIF), UNECE R157 / R156, Suzuki Wataru (2025). Safety Analysis and Design Improvement for Semi-Automatic Train Operation (STO) in High-Speed Rail Using STPA, PhD Thesis, MIT, Cappelli, M.A., Di Marzo Serugendo, G. A semi-automated software model to support AI ethics compliance assessment of an AI system guided by ethical principles of AI. AI Ethics 5, 1357–1380 (2025), Andreas Holzinger, Kurt Zatloukal, Heimo Müller, Is human oversight to AI systems still possible?, New Biotechnology, Volume 85, 2025.
Any disagreements between use partners on inclusion of this principle?
No disagreements; consensus achieved.

UC2: Accountability and traceability of safety decisions
Definition
Accountability and traceability of safety decisions refer to the ethical, technical, and organisational mechanisms that ensure every safety-related decision - whether made by humans, AI systems, or a combination of both - is attributable, explainable, and verifiable throughout the system lifecycle. They guarantee that the origin, rationale, data, and authority behind each safety judgement can be reconstructed and audited, enabling responsibility to be clearly assigned and corrective action to be taken when necessary.
Relevance for this use case
In semi-automated AI-supported safety analysis, traceability ensures that engineers can see the data and logic behind AI recommendations. Accountability ensures that a qualified human ultimately validates or rejects those outputs before they enter the safety case. Ethically, they prevent 'responsibility gaps and automation bias, supporting informed oversight and public trust.
Source(s) that suggest inclusion of this principle
EU AI Act (2024, Art. 12–15), Ethics Guidelines for Trustworthy AI, ISO 26262:2018, ISO/PAS 21448:2022 (SOTIF), UNECE R155 / R156, Rowe, F., Jeanneret Medina, M., Benoit Journé, Coëtard, E., & Myers, M. (2023). Understanding responsibility under uncertainty: A critical and scoping review of autonomous driving systems. Journal of Information Technology, 39(3), 587-615, Tanja Pavleska, Massimiliano Masi, Giovanni Paolo Sellitto, Helder Aranha, Architecture-based governance for secure-by-design cooperative intelligent Transport Systems, Vehicular Communications, Volume 55, 2025
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC2: Transparency of safety-critical performance and limits
Definition
Transparency refers to the obligation to ensure that the capabilities, reliability, and boundaries of an AI-based or automated system — including when, where, and how it may fail — are clearly documented, communicated, and understandable to all relevant stakeholders (engineers, operators, regulators, and end-users). It requires that both the performance metrics (e.g., detection accuracy, fault tolerance) and the operational limits (e.g., environmental conditions, uncertainty thresholds) are explicit, interpretable, and verifiable throughout the system's lifecycle.
Relevance for this use case
In the context of AI-assisted safety analysis or vehicle operation, this principle ensures that engineers understand the confidence and reliability of AI-generated safety assessments; operators and regulators know under which conditions system outputs are valid; human supervisors can identify when intervention is needed due to uncertainty, data drift, or unexpected scenarios; the limits of automation are clear enough to prevent over-trust or misuse — a recurring cause of safety-critical failures in semi-automated systems.
Source(s) that suggest inclusion of this principle
EU AI Act (Art. 13), Ethics Guidelines for Trustworthy AI (EC HLEG, 2019), ISO 26262:2018, ISO/PAS 21448:2022 (SOTIF), UNECE Regulation 157 (Automated Lane Keeping Systems), SAE J3016 (Levels of Driving Automation), Daniel Omeiza, Raunak Bhattacharyya, Marina Jirotka, Nick Hawes, Lars Kunze, A transparency paradox? Investigating the impact of explanation specificity and autonomous vehicle imperfect detection capabilities on passengers, Transportation Research Part F: Traffic Psychology and Behaviour, Volume 109, 2025, Nagadivya Balasubramaniam, Marjo Kauppinen, Antti Rannisto, Kari Hiekkanen, Sari Kujala, Transparency and explainability of AI systems: From ethical guidelines to requirements, Information and Software Technology, Volume 159, 2023.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC2: Preservation of human skill and expertise
Definition
Preservation of human skill and expertise refers to the ethical and organisational commitment to ensure that the introduction of AI and automation does not erode the human operators' knowledge, judgement, diagnostic capability, and situational awareness that are essential for maintaining system safety, adaptability, and accountability. It means that humans remain competent, engaged, and capable of independent decision-making, even as automation performs parts of their tasks.
Relevance for this use case
In semi-automated safety analysis, supported by AI, this principle is highly relevant, because if engineers lose the ability to reason about causal chains or safety goals, systematic errors could remain undetected, compromising certification. Also, preservation of skills ensures engineers can cross-check AI outputs using established methods and domain logic. It also allows recognition of atypical or emergent hazards that fall outside trained data. It sustains the ability to interpret weak signals, anomalies, or ethical implications not codified in the AI model.
Source(s) that suggest inclusion of this principle
EU AI Act (2024), Art. 14 & Recital 47, Ethics Guidelines for Trustworthy AI, ISO 26262:2018, ISO/PAS 21448:2022 (SOTIF), OECD (2023), OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market, OECD Publishing, Paris, https://doi.org/10.1787/08785bba-en, Martina Benvenuti, Angelo Cangelosi, Armin Weinberger, Elvis Mazzoni, Mariagrazia Benassi, Mattia Barbaresi, Matteo Orsoni, Artificial intelligence and human behavioral development: A perspective on new skills and competences acquisition for the educational context, Computers in Human Behavior, Volume 148, 2023.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC2: Feedback and learning loops for human adaptation
Definition
Feedback and learning loops for human adaptation refer to the continuous, bidirectional exchange of information between human users and automated or AI-based systems that allows both sides to learn and improve over time.
Relevance for this use case
In AI-assisted safety engineering, feedback and learning loops are essential to ensure that automation remains aligned with evolving human expertise, regulatory frameworks, and contextual realities of automotive systems.
Source(s) that suggest inclusion of this principle
Ethics Guidelines for Trustworthy AI, EU AI Act (2024), Articles 14 & 15, ISO 26262:2018, ISO/PAS 21448:2022, Glickman, M., Sharot, T. How human–AI feedback loops alter human perceptual, emotional and social judgements. Nat Hum Behav 9, 345–359 (2025), Sumon Biswas, Yining She, and Eunsuk Kang. 2023. Towards Safe ML-Based Systems in Presence of Feedback Loops. In Proceedings of the 1st International Workshop on Dependability and Trustworthiness of Safety-Critical Systems with Machine Learned Components (SE4SafeML 2023), Tsiakas, K., & Murray-Rust, D. (2022). Using human-in-the-loop and explainable AI to envisage new future work practices. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2022 (pp. 588-594).
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC2: Training, education, and continuous skill development
Definition
Training, education, and continuous skill development refer to the systematic process of building, updating, and maintaining human competence necessary to understand, supervise, and complement AI-based or automated systems throughout their lifecycle. It ensures that all personnel involved — engineers, operators, safety managers, and decision-makers — possess and sustain the knowledge, cognitive skills, and ethical awareness required to perform their roles safely, effectively, and responsibly in a technologically evolving environment.
Relevance for this use case
The automotive safety domain — governed by strict standards — demands competent human oversight at every stage of the system lifecycle. When AI tools are introduced for semi-automatic safety evaluation, continuous education becomes essential to preserve human expertise and accountability, ensure correct and safe use of AI tools, adapt to evolving technologies and standards, support workforce resilience and inclusion.
Source(s) that suggest inclusion of this principle
Ethics Guidelines for Trustworthy AI, EU AI Act (Articles 4-15), IEEE 7000-2021. Ethical System Design Process, Farooqi, M. T. K., Amanat, I., & Awan, S. M. (2024). Ethical Considerations and Challenges in the Integration of Artificial Intelligence in Education: A Systematic Review. Journal of Excellence in Management Sciences, 3(4), 35–50,
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC2: Organisational policies for shared responsibility
Definition
Organisational policies for shared responsibility refer to the formal structures, procedures, and cultural norms established within an organisation to ensure that responsibility for AI-supported decisions is clearly distributed, traceable, and collectively managed across humans, systems, and organisational entities. The principle recognizes that in complex socio-technical systems—such as semi-automated safety analysis in automotive—no single individual or component holds full control over outcomes. Therefore, safety, ethics, and accountability must be supported by organisational mechanisms that promote collaboration, transparency, and joint ownership of safety assurance.
Relevance for this use case
Semi-automated safety analysis systems distribute decision-making across multiple layers: data engineers, model developers, safety analysts, project managers, and sometimes external auditors. Without structured shared-responsibility policies, accountability gaps or duplication of oversight may arise.
Source(s) that suggest inclusion of this principle
EU AI Act (Articles 9–14), ISO 26262:2018, Part 2, ISO/PAS 21448:2022 (SOTIF), IEEE 7000:2021, M. L. Cummings, "Identifying AI Hazards and Responsibility Gaps," in IEEE Access, vol. 13, pp. 54338-54349, 2025, Falco, G., Shneiderman, B., Badger, J. et al. Governing AI safety through independent audits. Nat Mach Intell 3, 566–571 (2021), ankins, S., Ocampo, A. C., Marrone, M., Restubog, S. L. D., & Woo, S. E. (2024). A multilevel review of artificial intelligence in organisations: Implications for organisational behavior research and practice. Journal of Organisational Behavior, 45(2).
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC3: Diversity
Definition
Inclusion of data and behavioural patterns that are representative of the organisation's workforce (across roles, locations, languages and protected groups), and integration of multi-stakeholder perspectives (e.g. HR, IT/security, worker representatives) in the design, monitoring and deployment of phishing-risk models.
Relevance for this use case
Diversity is essential to avoid systematically over- or under-estimating risk for specific groups (e.g. by role, location, language, gender or contract type). Ensuring diverse data and perspectives helps prevent discriminatory patterns in alerts, coaching actions or HR-relevant decisions, and aligns the system with EU AI Act and GDPR principles on fairness, purpose limitation and non-discrimination.
Source(s) that suggest inclusion of this principle
Kavvadias, A., & Kotsilieris, T. (2025). Understanding the Role of Demographic and Psychological Factors in Users' Susceptibility to Phishing Emails: A Review. Applied Sciences, 15(4), 2236. https://doi.org/10.3390/app15042236; Frank L. Greitzer, Wanru Li, Kathryn B. Laskey, James Lee, and Justin Purl. 2021. Experimental Investigation of Technical and Human Factors Related to Phishing Susceptibility. Trans. Soc. Comput. 4, 2, Article 8 (June 2021), 48 pages. https://doi.org/10.1145/3461672; Diaz, A., Sherman, A. T., & Joshi, A. (2020). Phishing in an academic community: A study of user susceptibility and behavior. Cryptologia, 44(1), 53-67; Monsoro, N., Martinie, C., Palanque, P., Saubanère, T. (2025). A Systematic Task and Knowledge-Based Process to Tune Cybersecurity Training to User Learning Groups: Application to Email Phishing Attacks. In: Clarke, N., Furnell, S. (eds) Human Aspects of Information Security and Assurance. HAISA 2024. IFIP Advances in Information and Communication Technology, vol 721. Springer, Cham. https://doi.org/10.1007/978-3-031-72559-3_12
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC3: Representativeness/Inclusivity
Definition
Extent to which the system's risk signals, features, and interventions reflect the realities of different employee groups (roles, locations, contract types, digital literacy levels, accessibility needs), and are designed so that all users can understand, access and benefit from the system on equal terms.
Relevance for this use case
Because phishing exposure and digital behaviours vary significantly across roles, seniority, locations and digital skills, systems that measure vulnerability to phishing must be representative and inclusive to avoid systematically overlooking or penalising specific groups. Ensuring that risk scores, thresholds and training/coaching content are understandable and usable for all employees supports non-discrimination, mitigates disparate impact in HR-relevant decisions, and aligns such systems with the EU AI Act fairness requirements and GDPR principles of fairness and data minimisation.
Source(s) that suggest inclusion of this principle
Diaz, A., Sherman, A. T., & Joshi, A. (2020). Phishing in an academic community: A study of user susceptibility and behavior. Cryptologia, 44(1), 53-67; Microsoft Design. (2025). Secure by Design: A UX Toolkit. Microsoft. Retrieved from https://microsoft.design/articles/secure-by-design-a-ux-toolkit/
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC3: Objectivity
Definition
Use of transparent, evidence-based and standardised criteria to assess phishing vulnerability, minimising subjective judgments or ad-hoc decisions in how risk signals are generated, aggregated and interpreted across employee groups.
Relevance for this use case
Because vulnerability assessments can influence how employees are perceived, prioritised for training, or flagged as "high risk", relying on objective criteria is essential to avoid arbitrary or biased treatment. Clear, documented and auditable methods for scoring and segmenting users reduce the risk of discrimination, support contestability of decisions, and align with the EU AI Act's requirements on robustness and governance, as well as GDPR principles of fairness, accountability and transparency in automated processing.
Source(s) that suggest inclusion of this principle
Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. Big Data & Society, 8(1), 2053951720983865; Usman, Q., & Jackson, M. (2022). Ethical AI in Cybersecurity: Addressing Bias and Fairness in Automated Threat Detection Systems; Bahangulu, J. K., & Owusu-Berko, L. (2025). Algorithmic bias, data ethics, and governance: Ensuring fairness, transparency and compliance in AI-powered business analytics applications. World J Adv Res Rev, 25(2), 1746-63.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC3: Non-stigmatizing use / Proportionality
Definition
Use of vulnerability scores in ways that are proportionate to the security objective and avoid labelling or penalising individuals or groups, focusing on support and risk reduction rather than blame.
Relevance for this use case
If vulnerability scores are used to stigmatise certain employees or groups (e.g. "unreliable", "untrustworthy"), this can create unfair treatment, workplace tension and potential discrimination. Anchoring the system in proportional, non-punitive use ensures that outputs serve security and resilience goals while respecting dignity and equal treatment. This is consistent with the AI Act's risk-based and fundamental-rights-oriented approach, and with GDPR principles such as purpose limitation, fairness and data minimisation in processing employee data.
Source(s) that suggest inclusion of this principle
Capasso M, Arora P, Sharma D, Tacconi C. On the Right to Work in the Age of Artificial Intelligence: Ethical Safeguards in Algorithmic Human Resource Management. Business and Human Rights Journal. 2024;9(3):346-360. doi:10.1017/bhj.2024.26
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC3: Openness
Definition
Degree to which the organisation provides clear, accessible and non-technical information about the existence, purpose and main functioning of the phishing-vulnerability measurement system, including what data it uses, how results may affect employees, and who is responsible for its governance.

Relevance for this use case
A system that measures individual vulnerability to phishing can directly affect employees' trust, perceived autonomy and sense of fairness, especially when outputs may feed into training plans, performance discussions or HR-relevant decisions. Openness about the system's purpose, data sources, safeguards and governance reduces fears of hidden surveillance or punitive use, enables informed participation, and supports accountability. It is also closely aligned with transparency obligations in the EU AI Act for high-risk AI systems, and with GDPR principles on transparency, information duties towards data subjects and fair processing of employee data.
Source(s) that suggest inclusion of this principle
Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. Science and engineering ethics, 26(6), 3333-3361; Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. Information and Software Technology, 159, 107197.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC3: Accessibility/Access to information
Definition
Ensuring that all employees can easily access and understand information about the phishing-vulnerability measurement system, its purpose, data inputs, outputs, rights, safeguards and potential impacts, through channels and formats adapted to different roles, languages, digital literacy levels and accessibility needs.
Relevance for this use case
Because vulnerability assessments relate directly to individual employees, all users, regardless of role, location, language or technical ability, must have equal access to information about how the system affects them. Clear and accessible communication supports informed participation, reduces misconceptions or fears of surveillance, and enables users to exercise their GDPR rights (access, rectification, objection, information). It is also consistent with the AI Act's emphasis on user-facing transparency and with fundamental-rights safeguards requiring that information be understandable, available and actionable for all impacted individuals.
Source(s) that suggest inclusion of this principle
Yanamala, K. K. R. (2023). Transparency, privacy, and accountability in AI-enhanced HR processes. Journal of Advanced Computing Systems, 3(3), 10-18; Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. Information and Software Technology, 159, 107197; Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021, May). Expanding explainability: Towards social transparency in ai systems. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-19).
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC3: Documentation, traceability & auditability
Definition
Availability of clear, up-to-date documentation and logs describing the design, data sources, model versions, configuration changes and decision logic of the phishing-vulnerability measurement system, as well as traceable records that allow reconstruction and external review of how specific vulnerability scores or decisions were produced.

Relevance for this use case
Organisations must be able to explain and demonstrate how the system works and why a given outcome occurred. Robust documentation and traceability enable internal and external audits, support investigations of potential bias or errors, and provide the basis for meaningful contestation by affected individuals. This is aligned with the EU AI Act requirements on technical documentation, logging and oversight for high-risk AI systems, and with GDPR principles of accountability, transparency and the ability to substantiate compliance when processing employee data.
Source(s) that suggest inclusion of this principle
Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. Information and Software Technology, 159, 107197; Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021, May). Expanding explainability: Towards social transparency in ai systems. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-19).
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC3: Dependence
Definition
Degree to which security, HR or management decisions rely solely or predominantly on phishing-vulnerability scores and automated outputs, instead of combining them with human judgement, contextual information and other complementary security indicators.
Relevance for this use case
A system that measures individual vulnerability to phishing can create a false sense of precision and encourage staff to treat scores as definitive truth about an employee's behaviour or reliability. Over-dependence on automated outputs increases the risk of unfair treatment, misallocation of training or controls, and blind spots where the system is less accurate (e.g. specific roles or contexts). Managing dependence, by keeping human oversight, contextual review and clear usage boundaries, aligns with the AI Act requirements to avoid inappropriate reliance on high-risk AI systems, and with GDPR principles of fairness, proportionality and meaningful human involvement in decisions affecting employees.
Source(s) that suggest inclusion of this principle
Mujtaba, D. F., & Mahapatra, N. R. (2024). Fairness in AI-driven recruitment: Challenges, metrics, methods, and future directions. arXiv preprint arXiv:2405.19699; Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: "automation bias" and "selective adherence" to algorithmic advice. Journal of Public Administration Research and Theory, 33(1), 153-169.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC3: Contestability & human oversight
Definition
Existence of clear, accessible mechanisms for employees and relevant stakeholders (e.g. HR, security, worker representatives) to review, question and correct vulnerability assessments and related decisions, supported by documented human-in-the-loop oversight for high-impact uses of the system
Relevance for this use case
Because vulnerability scores and risk flags can influence how employees are treated (e.g. training assignments, performance discussions, escalation to HR), affected individuals should be able to understand and contest outcomes that they perceive as inaccurate or unfair. Human oversight

prevents blind reliance on automated scoring, helps detect hidden biases or data quality issues, and aligns with AI Act obligations on human oversight for high-risk AI systems as well as GDPR principles on fairness, transparency and rights related to automated decision-making.
Source(s) that suggest inclusion of this principle
Holzinger, A., Zatloukal, K., & Müller, H. (2025). Is human oversight to AI systems still possible?. New Biotechnology, 85, 59-62; Ottun, A. R. O., & Flores, H. (2025). Trustworthy AI in Practice: A Comprehensive Review of Human Oversight and Human-in-the-Loop Approaches. Authorea Preprints.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC4: Autonomy and Agency
Definition
The right of individuals to form, express and share opinions without risk or fear of punitive interference
Relevance for this use case
AI systems should not remove or alter legitimate speech just because it is unpopular or controversial
Source(s) that suggest inclusion of this principle
Scanlon, T. (1972). A theory of freedom of expression. Philosophy & Public Affairs, 204-226; Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. Computer science, 7, e598; Roberts, S. T. (2019). Behind the screen. Yale University Press.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC4: Proportionality
Definition
Any interference or restriction should be the least restrictive possible to reduce the likelihood of harm
Relevance for this use case
If AI flags hate speech, it should do so only when the content genuinely meets clear hate speech thresholds, not for perceived offence or disagreement
Source(s) that suggest inclusion of this principle
Tsakyrakis, S. (2009). Proportionality: An assault on human rights?. International Journal of Constitutional Law, 7(3), 468-493; Thiago Dias Oliva, Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression, Human Rights Law Review, Volume 20, Issue 4, December 2020, Pages 607–640, https://doi.org/10.1093/hrlr/ngaa032; Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. Neurocomputing, 546, 126232.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC4: Non-descrimination
Definition
Equal treatment of all speech, regardless of sources identity, background or viewpoint
Relevance for this use case
AI systems must be trained to avoid bias that supports discrimination against or suppression of specific groups or political standpoints

Source(s) that suggest inclusion of this principle
Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., & Smith, N. A. (2021). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. arXiv preprint arXiv:2111.07997; Heinrichs, B. (2022). Discrimination in the age of artificial intelligence. AI & society, 37(1), 143-154; Nascimento, F. R., Cavalcanti, G. D., & Costa-Abreu, M. D. (2025). Gender bias detection on hate speech classification: an analysis at feature-level. Neural Computing and Applications, 37(5), 3887-3905.
Any disagreements between use partners on inclusion of this principle?
[none reported]
UC4: Equality and Impartiality
Definition
All groups adn individuals should be treated with equal respect and dignity
Relevance for this use case
AI systems must not flag or suppress speech disproportionately from any demographic group
Source(s) that suggest inclusion of this principle
Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC4: Representation and Inclusivity
Definition
Diverse voices, dialects and cultures should be recognised and reflected
Relevance for this use case
Training datasets should include varied linguistic and cultural expressions to prevent marginalisation or discrimination of particularly underrepresented communities
Source(s) that suggest inclusion of this principle
Maronikolakis, A., Baader, P., & Schütze, H. (2022). Analyzing hate speech data along racial, gender and intersectional axes. arXiv preprint arXiv:2205.06621; Peterson-Salahuddin, C. (2024). Repairing the harm: Toward an algorithmic reparations approach to hate speech content moderation. Big Data & Society, 11(2), 20539517241245333.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC4: Transparency of criteria
Definition
Users have a right to understand how decisions are made
Relevance for this use case
Clear explanation of why content is flagged helps prevent perceptions of bias and enables appeal of decision
Source(s) that suggest inclusion of this principle
Gonçalves, J., Weber, I., Masullo, G. M., Torres da Silva, M., & Hofhuis, J. (2023). Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. new media & society, 25(10), 2595-2617; Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. Science and

engineering ethics, 26(6), 3333-3361; Hayes, P., van de Poel, I., & Steen, M. (2023). Moral transparency of and concerning algorithmic tools. AI and Ethics, 3(2), 585-600.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC4: Human Oversight
Definition
Responsibility requires 'humans in the loop' for sensitive or borderline cases
Relevance for this use case
Automated hate speech detection should not function as a “black box” without human review for context-heavy judgments.
Source(s) that suggest inclusion of this principle
Alkiviadou, N. (2022). ARTIFICIAL INTELLIGENCE AND ONLINE HATE SPEECH MODERATION. Sur: Revista Internacional de Derechos Humanos, 19(32); Gier-Reinartz, N. R., Zimmermann-Janssen, V. E., & Kenning, P. (2023). AI-Assisted Hate Speech Moderation—How Information on AI-Based Classification Affects the Human Brain-In-The-Loop. In NeuroIS Retreat (pp. 45-56). Cham: Springer Nature Switzerland.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC4: Auditability/Evaluation
Definition
External parties (regulators, watchdogs) evaluating performance independently
Relevance for this use case
Hate speech models should allow third-party audits, enabling external checks on fairness and compliance.
Source(s) that suggest inclusion of this principle
Hee, M. S., Sharma, S., Cao, R., Nandi, P., Nakov, P., Chakraborty, T., & Lee, R. (2024). Recent advances in online hate speech moderation: Multimodality and the role of large models. Findings of the Association for Computational Linguistics: EMNLP 2024, 4407-4419; Balendra, S. (2025, January). Meta’s AI moderation and free speech: Ongoing challenges in the Global South. In Cambridge Forum on AI: Law and Governance (Vol. 1, p. e21). Cambridge University Press.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC4: Responsiveness
Definition
Rapid updates, retraining, or adjustments should follow evidence of bias in moderation outcomes
Relevance for this use case
Willingness to act when problems are identified
Source(s) that suggest inclusion of this principle
Alkiviadou, N. (2022). ARTIFICIAL INTELLIGENCE AND ONLINE HATE SPEECH MODERATION. Sur: Revista Internacional de Derechos Humanos, 19(32).
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC5: User consent & Understanding
Definition
Users have to be provided with clear and understandable terms and conditions based on which they consent to the processing, storage, and sharing of their user data (directly submitted by users) and inferred data (derived from user behaviour and activity patterns)
Relevance for this use case
Privacy as value to uphold and right to be protected (see literature). While user consent is necessary to store relevant data, process it for and transfer it to payment platform, transparency about this process is rather a condition for impairing if user data is protected or not.
Source(s) that suggest inclusion of this principle
EU AIA Art. 50; Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
Any disagreements between use partners on inclusion of this principle?
Currently extensive monitoring is conducted to fulfill the multi-tier moderation system. It remained unclear if this is communicated transparently enough to the users.

UC5: Data minimization, data use & storage
Definition
Collect user data only necessary for a specific purpose and store it not longer than required (GDPR) to prevent excessive data collection
Relevance for this use case
Legal regulations require providers to collect only data necessary for specific purposes. Users expect providers to respect their privacy and autonomy thereby enhancing user trust. However, multi-tier moderation system and safety obligations require monitoring of user interactions and activity.
Source(s) that suggest inclusion of this principle
GDPR; EU AIA Art. 10 + Recital 69; Yanamala, A. K. Y., Suryadevara, S., & Kalli, V. D. R. (2024). Balancing innovation and privacy: The intersection of data protection and artificial intelligence. *International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence*, *15*(1), 1-43.; http://data.europa.eu/eli/reg/2016/679/oj; Brown, R., & Truby, J. (2022). Mending lacunas in the EU's GDPR and proposed artificial intelligence regulation, 10.2478/eustu-2022-0003
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC5: Third-party sharing & compliance
Definition
Sharing of data with third-parties and compliance with agreed standards
Relevance for this use case
Payment platforms act as de-facto regulators
Source(s) that suggest inclusion of this principle
GDPR; EU AI Art. 11 + Annex IV
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC5: User protection
Definition
Protection of the user from harmful, violent, or psychologically distressing content generated through the AI characters
Relevance for this use case
User protection was rated by the provider of the personalized character app as the ethical principle with highest priority. Users engage in personal, intimate, and sensitive conversations with AI characters that have an emotional and conversational nature. This carries the risk of the exposure to harmful, disturbing, and potentially traumatic content.
Source(s) that suggest inclusion of this principle
EU AIA recital 5 + 9; Felipe Romero Moreno (2024) Generative AI and deepfakes: a human rights approach to tackling harmful content, International Review of Law, Computers & Technology, 38:3, 297-326, DOI: 10.1080/13600869.2024.2324540
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC5: Security measures
Definition
Measures that monitor and ensure compliance with the conditions such as age verification, multie-tier classification models that classify content into safety tiers at scale, monitoring and responding to harmful usage patterns through warnings and bans. The measures should prevent content that references physical violence toward living beings.
Relevance for this use case
Monitoring and ensuring security measures are the basis for use and legal protection both for the provider and the user side.
Source(s) that suggest inclusion of this principle
EU AIA Art. 5; Chawki, M. (2025). AI Moderation and Legal Frameworks in Child-Centric Social Media: A Case Study of Roblox. *Laws*, *14*(3), 29. https://doi.org/10.3390/laws14030029
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC5: Human oversight
Definition
The possibility that a human moderator can take over automated monitoring at any time, e.g., in critical situations or when the AI system faces complex scenarios beyond its design limits. Human moderators intervene when it comes to borderline cases (e.g., potentially disturbing but not explicitly violent behaviour) and decide about further action.
Relevance for this use case
In the context of conversational AI systems designed to be personalized through user data and user-created characters that can be commercially reused by other users, human oversight is crucial to prevent harmful consequences that may affect users and others. Moreover, human oversight is necessary to detect and evaluate edge cases, jailbreaking attempts, and distinguish between legitimate user preferences and harmful behaviours.
Source(s) that suggest inclusion of this principle
EU AIA Art. 14; Ethics Guidelines for Trustworthy AI (HLEG, 2019); Lena Enqvist (2023) 'Human oversight' in the EU artificial intelligence ct: what, when and by whom?, Law, Innovation and Technology, 15:2, 508-535, DOI: 10.1080/17579961.2023.2245683

Any disagreements between use partners on inclusion of this principle?
[none reported]

UC5: Informed consent
Definition
Ability to make informed choices based on transparent terms and conditions of an application. Possibility to refuse consent without facing negative consequences.
Relevance for this use case
Personalized AI characters require the expression of personal preferences and consent to specific types of content. At the same time, the user has to understand the limitations imposed by the moderation system and be able to evaluate personal limits to avoid potentially harmful usage and consumption.
Source(s) that suggest inclusion of this principle
EU AIA Art. 50; Definition under Art 3 (59)
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC5: System customization
Definition
Customization and personalization of application within safe boundaries.
Relevance for this use case
Users seeking explicitly for customized AI applications want the possibility of free expression of personal preferences. The provider enables this while balancing legitimate user expressions and harmful behaviour and attempts to circumvent built-in limitations (e.g., jailbreaking).
Source(s) that suggest inclusion of this principle
Xi Yang and Marco Aurisicchio. 2021. Designing Conversational Agents: A Self-Determination Theory Approach. In CHI Conference on Human Factors in Computing Systems (CHI '21), May 08–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3411764.3445445
Any disagreements between use partners on inclusion of this principle?
[none reported]
UC5: Transparency & user understanding
Definition
Provision of clear, concise, and easy to understand information about use of user behaviour/interaction data
Relevance for this use case
[not reported]
Source(s) that suggest inclusion of this principle
EU AIA Art 50; Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019, May). Guidelines for human-AI interaction. In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1-13).
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC6: Subsidiarity and proportionality
Definition
Prevent overreliance on experimental tools in contexts where less intrusive, non-experimental methods are available as first option for a particular patient. Balance patient therapeutic needs against potential harms.
Relevance for this use case
Deepfake therapy's benefits should outweigh the harms (proportionality) and therapy should involve the least intrusive alternative (subsidiarity) compared with similar technologies such as virtual reality therapy, imagery rehearsal and traditional exposure therapy. Potential harms to the patient: these may begin to arise in the process of collecting data for the creation of deepfakes, which largely relies on what the patient brings along. If these data are not readily available, patients should not be placed—especially in the sexual violence case—in dangerous situations in order to collect videos or photographs. Even if the data are available, the retrieval of photographs from personal archives or social media may be a 'triggering' moment for a patient: it may cause emotional distress as it arouses feelings or memories associated with the trauma.
Source(s) that suggest inclusion of this principle
Discussion between partners; Hoek et al. J Med Ethics 2025;51:481–486.
Any disagreements between use partners on inclusion of this principle?
Yes, the risk related to collecting media (dangerous situations or 'triggering') is not seen as an issue by the stakeholder as in the setting that is now being tested, the deepfake only requires a couple of photos and is trained on the one that works best. Bringing a photo is common practice for other types of exposure therapy too.

UC6: Effectiveness
Definition
Ensuring that actions, policies, or interventions achieve their intended purpose with the least harm possible. An ineffective action that introduces risks without benefits would violate non-maleficence.
Relevance for this use case
Implementation should follow established clinical protocols, adapted for PTSD contexts and shown to be effective. Right now, deepfakes are introduced in psychotherapy mostly in an experimental context, but this will likely change if their effectiveness is demonstrated. In the use case, deepfake therapy is not the first line of PTSD treatment but only used for residual issues mainly in the area of 'moral injury' like persistant shame or a damaged self image. It has, at the moment, been tested on three people wih one session each, and one of them wanted to have multiple sessions. The experimental question is what the therapist should say so that the session is helpful. The effectiveness of the the technological aspects of AI is also important; there was a mention of a case where AI twisted the picture after an update, causing it not to resemble the perpetrator anymore, which was stressful.
Source(s) that suggest inclusion of this principle
Discussion between partners; Hoek et al. J Med Ethics 2025;51:481–486.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC6: Societal well-being
Definition
Considering the broader impact on communities and society as a whole, beyond individual outcomes. This includes promoting collective safety, fairness, and sustainable practices while minimizing harm at the societal level.

Relevance for this use case
Societal implications of normalizing deepfake use for well-being purposes (risk of misuse outside clinical settings). The field of mental health has seen an increase in the use of (commercial) technologies such as chatbots: one day those chatbots might be extended with deepfaked voices and faces. In general, deepfakes might harm societal trust and diminish the evidential value of video material, but this is not so much the case for the controlled clinical setting.
Source(s) that suggest inclusion of this principle
Discussion between partners; Hoek et al. J Med Ethics 2025;51:481–486.
Any disagreements between use partners on inclusion of this principle?
Yes, because this is not seen as an issue within a controlled therapeutic setting.

UC6: Transparency
Definition
Ensuring that patients (or individuals) receive complete, clear, and honest information about treatment options, risks, and outcomes. Transparency supports informed consent, allowing individuals to make autonomous choices.
Relevance for this use case
Patients must clearly understand the intervention, its experimental nature, and that the deepfake is not “real”. Informed consent must cover use of personal data, therapeutic goals, and foreseeable risks. Deepfake researchers and therapists should be transparent about the limits and uncertainties of the method.
Source(s) that suggest inclusion of this principle
Discussion between partners; Hoek et al. J Med Ethics 2025;51:481–486.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC6: Privacy
Definition
Respecting the confidentiality of personal and medical information, protecting individuals from unauthorized access or misuse of their data. Privacy safeguards autonomy by allowing individuals to control how their information is shared and used.
Relevance for this use case
Deepfakes are often generated without consent. This use of a natural person’s image without their consent raises the question of whether there are legal and ethical grounds to object to such use within the context of therapy. In cases of data breach, the deepfake may also cause ‘reputational injury’ to the perpetrator, as it impacts their social identity and reputation when a deepfake video is seen and believed by others. Consider legal/privacy implications of using images without consent from perpetrators, to balance patient therapeutic needs against third-party privacy rights. In grief counseling: what is new about deepfake therapy, is that people might start to curate video content of their living loved ones specifically for the purpose of future grief counselling, which can positively or negatively impact those relationships
Source(s) that suggest inclusion of this principle
Discussion between partners; Hoek et al. J Med Ethics 2025;51:481–486. For the point about curating data of loved ones, see: Fabry RE , Alfano M The affective scaffolding of grief in the digital age: the case of deathbots. Topoi (Dordr) 2024;1–13.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC6: Risk of over-attachment and dependency
Definition
Recognizing that individuals may become overly reliant on caregivers or in this case even on technological systems (i.e. the deepfake). Excessive dependency can limit a person's ability to make independent choices or foster emotional reliance that affects autonomous decision-making.
Relevance for this use case
The simulated confrontation with the subject of a person's trauma may cause specific risks related to dependencies and a blurring of reality. The patient–therapist relationship itself may be impacted given that the therapist is controlling the deepfake: the patient may start to associate or even identify the therapist with the perpetrator or deceased loved one. This could possibly lead to confusion, feelings of unsafety or an unhealthy attachment to the therapist. Second, a patient could become 'addicted' to communicating with the generated image. This is most likely to happen in the grief counselling case. PTSD case: Could victims of sexual trauma regain a misplaced and perhaps dangerous semblance of trust towards the perpetrator of their trauma?
Source(s) that suggest inclusion of this principle
Discussion between partners; Hoek et al. J Med Ethics 2025;51:481–486.; for the risk of overattachment see previous studies on griefbots: Krueger J , Osler L . Communing with the dead online: chatbots, grief, and continuing bonds. J Conscious Stud 2022;29:222–52. & Xygkou A , Siriaraya P , Covaci A . "The "conversation" about loss: understanding how chatbot technology was used in supporting people in grief". CHI '23; Hamburg Germany, April 19, 2023:1–15.
Any disagreements between use partners on inclusion of this principle?
Yes, because the article by Hoek et al. about deepfake therapy is focused on AI; while the stakeholder think that the person behind the laptop and the specific clinical setting are more important than the technology. Their work is focused on evaluating deepfake therapy in the PTSD setting which is a very different dynamic than grief therapy. People who grieve cannot go on because they do not want to; while with PTSS you want to let go but cannot. The two issues are totally different.

UC6: Human agency and responsibility
Definition
In psychotherapy, the therapist and institution remain responsible for providing good care, regardless of the AI used.
Relevance for this use case
Deepfake technology should remain a tool in the hands of a therapist—not a standalone or addictive chatbot. Most national legal systems contain the general 'duty of providing good care' which also plays an important role in jurisprudence. Good care implies that the care provided is of a good quality, that is, safe, effective and efficient and tailored to the patient's real needs.
Source(s) that suggest inclusion of this principle
Medical Treatment Act (NL). Discussion between partners; Hoek et al. J Med Ethics 2025;51:481–486.
Any disagreements between use partners on inclusion of this principle?
Not a disagreement, but national differences. Healthcare is predominantly governed by bioethical principles, national legislation and professional guidelines, and what a therapist has to take into account in terms of quality standards, treatment plans, and general rights and obligations will vary country by country. Fulfilling one's duty of care in general implies that healthcare providers, when caring for their patients, adhere to the medical-professional standard, including guidelines, protocols, medical ethical codes and the like, established by the profession itself, as well as to health legislation and other documents guaranteeing the rights and interests of the patient involved.

UC6: Professional competence
Definition
Therapists and researchers must maintain up-to-date knowledge, clinical skills, and ethical standards. Using AI competently ensures that decisions are made with expertise, reducing the risk of harm and upholding trust.
Relevance for this use case
Skilled clinical judgment is essential: knowing when to confront, when to validate, and when to avoid harm. A lot depends on the skills of the deepfake therapist, who will be a specially trained therapist conducting the session after an intake with the patient who is already treated by a regular therapist.
Source(s) that suggest inclusion of this principle
Discussion between partners; Hoek et al. J Med Ethics 2025;51:481–486.
Any disagreements between use partners on inclusion of this principle?
[none reported]

UC6: Oversight
Definition
Mechanisms of monitoring, review, and regulation ensure that practices meet ethical and professional standards.
Relevance for this use case
Important to ensure that it fits current clinical standards and it effective and not harmful.
Source(s) that suggest inclusion of this principle
Discussion between partners; Hoek et al. J Med Ethics 2025;51:481–486.
Any disagreements between use partners on inclusion of this principle?
The stakeholder notes that there is disagreement between RECs and the maker of deepfake software about whether it would be a medical device as meant in the Medical Device Regulation. In any case, as the software is not taking decisions independently, it may not need to adhere to the highest risk level. Article by Hoek et al: "- Binding regulations of the European Union (EU) on medical devices (particularly relevant here is the Medical Device Regulation) and AI (AI Act) include rules that ensure that unsafe, defective or harmful medical devices do not enter medical practice or the market: we find that according to these regulations, deepfakes for therapeutic purposes would classify as medical devices and thus as high-risk AI systems which are subject to stricter requirements.

Appendix D: Technical and organisational measures as provided by use cases

PRACTICAL MEASURES PROVIDED BY USE CASE 1

Table 37: UC 1 – Practical measures to achieve non-maleficence with respect to Validity/Accuracy

Technical measures to address Validity/Accuracy				
Describe the measure	Document FP/FN and communicate: record when the AI is right or wrong for each case and share patterns with clinicians. In the viewer, show the AI output with confidence and overlays/centreline, provide a one-click FP/FN tag, and compile monthly reports plus a small library of real examples for clinical review meetings.	Detect and alert anomalies: automatically flag unusual scans or outputs that fall outside validated ranges (for example, implausible measurements or poor image quality). When flagged, hold for human review and show a clear warning.	Safeguards against over-reliance: require an independent clinician first read before showing the AI, then ask for accept/adjust/reject with a reason for high-impact suggestions. Periodic "AI-off" spot checks help keep clinical skills sharp. Same in autonomous case, just for 10-15% of the cases instead of 100%.	Performance monitoring and oversight: track real-world performance per site with defined KPIs, thresholds, and named owners; maintain a runbook that explains what to do when a metric drifts or an incident occurs.
Why is it relevant?	Prevents repeat errors and helps clinicians learn where the AI can fail.	Reduces patient risk from rare or out-of-scope cases and corrupted inputs.	Limits automation bias and keeps clinicians accountable for final decisions.	Enables early detection of drift and meets post-market monitoring expectations.
How can it be achieved?	Error logging, libraries, viewer UI (FP/FN button; structured reason codes; monthly confusion-matrix reports), automated periodic reports.	Use uncertainty thresholds, range checks, data-integrity checks, and simple rules that trigger "hold for review." Tune thresholds to clinically	Use first-read mode, staged reveal of AI suggestions, and mandatory acknowledgment for high-impact outputs;	Define KPIs (sensitivity, specificity, measurement error, low-confidence rate), set alert thresholds, and run regular governance reviews

		meaningful limits. Another layer of AI algorithms that detect anomalies in inputs.	include training and refreshers for clinicians.	with a documented incident workflow.
How can be assessed whether this measure has been fulfilled?	Audit FP/FN logs, monthly reports, and review minutes; verify that feedback leads to product or model changes.	Review alert logs and outcomes; confirm flagged cases were held and reviewed; validate thresholds against retrospective samples.	Check acknowledgment logs, override rates, and training completion; verify AI-off audits were completed.	Check KPI dashboards, alert history, incident tickets, and corrective action records.
What are (potential) challenges to fulfilment?	Extra steps in workflow, clinician time, and keeping the failure library curated and useful.	Alert fatigue and difficulty setting thresholds that generalize across scanners and sites. New scanner types installed.	Clinician buy-in if the UI feels slow or intrusive.	Sustained resourcing, data integration across sites, and clarity on who owns follow-up actions.
What are risks if not fulfilled?	Recurring errors stay hidden, so trust erodes; no structured learning from failures.	Silent failures reach clinical decisions, increasing risk.	Automation bias, over-trust in AI, and reduced clinical vigilance.	Delayed detection of drift/regressions; late fixes and compliance gaps.
Which are the core function/role/ stakeholders responsible?	Product owner, clinicians	MLOps, safety officer	Governance board and clinicians	Safety officer, MLOps and clinicians
Specific requirements?	ISO 82304-1 (health software product safety)	ISO 14971, AI Act	IEC 62366-1, AI Act (human oversight)	AI Act (post-market monitoring)

Table 38: UC 1 – Practical measures to achieve non-maleficence with respect to Bias and Privacy

Technical measures to address Bias		**Technical measures to address Privacy**		
Describe the measure	Validate performance across diverse populations and anatomies (age, sex, ethnicity, comorbidities, scanner/protocol) and report subgroup metrics with confidence intervals. Publish a model card and monitor subgroup performance after deployment. Validate for specific geographic regions and clinical settings separately – including AI boards of specific institutions in the design phase and review phases.	Protect the AI pipeline against malicious or corrupted inputs and model misuse. Use threat modelling, adversarial testing, and hardened deployment to prevent manipulation. Keep regularly/frequently updating.	Maintain audit trails of AI access and use, with clear rules about who can view outputs and when. Log access by user, role, time, and purpose.	Protect data integrity during processing with DICOM checksums, provenance records, and immutable storage for critical logs so inputs and outputs cannot be altered without detection.
Why is it relevant?	Reduces unfair outcomes and hidden performance drops for underrepresented groups.	Prevents unsafe recommendations caused by tampered inputs or security incidents.	Supports accountability and compliance by making AI usage traceable.	Prevents tampering and supports forensic review if results are questioned.
How can it be achieved?	Use stratified datasets, prospective multi-site evaluation, drift checks by site/scanner, and a clear subgroup monitoring plan. Retrospective/controlled prospective audits on potential client sites.	Run penetration tests and red-team exercises; validate input quality; secure infrastructure and rate-limit risky access.	Implement append-only logs, RBAC tied to hospital identity systems, and periodic access reviews.	Verify checksums on ingest and output; store signed model artifacts and SBOM; use secure storage for key logs.
How can be assessed whether this measure has been fulfilled?	Subgroup performance reports exist; dashboards show stable metrics; review outcomes documented.	Security test reports and mitigations are tracked; no open critical findings.	Audit log retrieval works; access anomalies are reviewed; retention policy is met.	Integrity checks pass; provenance records reproduce inputs/outputs.
What are (potential) challenges to fulfilment?	Limited labelled data for rare anatomies; data-sharing constraints (incl. Privacy aspects) across sites.	Evolving attack patterns; balancing security with clinical speed; cost.	Costs, privacy	Hospital IT constraints and added processing overhead; cost. Storage limits.
What are risks if not fulfilled?	Biased care and unequal clinical outcomes.	Unsafe outputs, security breaches, or loss of data confidentiality.	Untraceable access and compliance failures.	Corrupted data leads to wrong outputs with weak evidence for investigation.

Which are the core function/role/ stakeholders responsible?	Safety officer, data scientists, clinicians, healthcare administration	Security lead	DPO, IT, auditors	DevOps, Security
Specific requirements?	AI Act, GDPR	ISO 27001, MDR	GDPR, AI Act	NIS2, ISO 27001

Table 39: UC 1 – Practical measures to achieve accountability and responsibility with respect to Auditability

Technical measures to address Auditability			
Describe the measure	Decision-making auditable/explainable - Provide case-level explanations that link outputs to evidence (overlays/centerlines, confidence, measurement trace) and store them with the report.	Logs/docs per output - Record structured logs per case (input metadata, model version, parameters, user actions, overrides, timestamps) and link them to the clinical record.	Model updates tracked & documented. Track model and data versions, release notes, validation evidence, and SBOM; ensure a tested rollback path.
Why is it relevant?	Allows clinicians and auditors to understand why a decision was made.	Enables reconstruction of decisions during incidents or audits.	Ensures reproducibility and accountability across updates. Prevents undocumented changes from causing unsafe drift.
How can it be achieved?	Use explanation panels in the viewer and store evidence snapshots with each report.	Use append-only logging with unique case IDs and secure retention policies.	Semantic versioning; release notes; signed artifacts; environment hashes.
How can be assessed whether this measure has been fulfilled?	Random audit samples show explanations are available and usable; audit retrieval succeeds.	Logs are complete and retrievable; retention meets policy.	Every record includes version IDs; change logs exist; rollback tested.
What are (potential) challenges to fulfilment?	Explainability can add workflow time and UI complexity – may not fully align with clinical protocols. Some AI systems are not fully explainable on individual pass level, yet provide most accurate performance (e.g. transformers), or explainability itself is unreliable.	Storage volume, privacy controls, and cross-system identifiers.	Config drift; multi-site rollouts; vendor coordination.
What are risks if not fulfilled?	Opaque decisions; reduced adoption; audit failures.	Inability to reconstruct decisions; compliance findings.	Undetected regressions; inconsistent behaviour across sites.
Which are the core function/role/ stakeholders responsible?	Product owner; UX lead; Radiology lead; QA; MLOps	IT admin; QA; DPO; Product owner	MLOps; DevOps; QA; Safety officer
Specific requirements?	AI Act transparency & logging; IEC 62366-1 usability; hospital clinical governance	GDPR (logging/retention); AI Act logging; hospital audit policies	AI Act change management; ISO 13485 concepts; NIS2 security; SBOM best practices

Table 40: UC 1 – Practical measures to achieve accountability and responsibility with respect to Human Oversight

Technical measures to achieve Human Oversight			
Describe the measure	Provide a clinician appeal/contest workflow to flag questionable AI outputs and request secondary review with clear timelines.	Allow clinicians to override AI outputs at any time; capture structured reasons and log the decision.	Clearly label AI-generated content versus clinician-authored conclusions in the viewer and report, including who signed off.
Why is it relevant?	Provides a safety exit when AI outputs look wrong.	Keeps clinicians as final decision-makers and preserves accountability.	Avoids confusion about who made the recommendation.
How can it be achieved?	Appeal button in the viewer; route to QA/safety queue; document outcomes and communicate back to users. Analytics dashboard with ability to investigate all cases individually by auditing clinician.	Override controls with reason codes; regular QA review and feedback to the model team; short refresher training.	Use badges, metadata tags, and clinician sign-off fields in the report.
How can be assessed whether this measure has been fulfilled?	Appeals logged with resolution times; audit trail complete.	Override logs with reason codes are present and correctly recorded; trend reviews and corrective actions documented.	UI shows clear attribution; clinician surveys confirm understanding.
What are (potential) challenges to fulfilment?	Resource burden to review appeals; possible workflow delays.	Consistency in reason coding; staff time for review. Insufficient training/education about AI limitations.	Viewer/vendor constraints; consistent labelling across systems and sites.
What are risks if not fulfilled?	Uncorrected misuse and loss of clinician trust. Worse clinical outcomes.	Missed systemic issues	Misattribution; legal ambiguity; loss of trust.
Which are the core function/role/ stakeholders responsible?	QA, clinical board, clinicians	QA, clinicians	Product owner; Viewer/PACS integrator; Legal/Risk mgmt
Specific requirements?	Hospital policy and clinical governance requirements	QA SOPs	AI Act transparency; medical record policies

Table 41: UC 1 – Practical measures to achieve accountability and responsibility with respect to Liability

Technical measures to address Liability			
Describe the measure	Monitor key performance metrics in near real time (sensitivity/specificity, measurement error, low-confidence rate) with thresholds and alerts, plus regular review meetings.	Define and communicate a RACI for AI-related errors and incidents across manufacturer, hospital, and clinical roles.	Provide case-level flagging for uncertain or wrong outputs with escalation and tracking to closure.
Why is it relevant?	Detects drift early and reduces patient risk.	Clarifies ownership and speeds response to issues.	Creates a feedback loop for improvement and safety.
How can it be achieved?	Operational dashboards, alert thresholds, and incident tickets; documented runbooks for response.	Publish RACI, train staff, and review quarterly; align with hospital governance.	Flag/report button, triage queue, second opinion workflow, and QA review to closure.
How can be assessed whether this measure has been fulfilled?	Alert precision/recall; MTTR; stability of metrics.	RACI approved and known; incident response KPIs meet targets.	Time to triage; closure rates; model updates driven by feedback.
What are (potential) challenges to fulfilment?	Data pipelines quality; alert fatigue; data consistency (in particular across sites).	Overlapping responsibilities between teams; governance updates.	Triage capacity and clinician engagement.
What are risks if not fulfilled?	Silent failures; degraded care quality.	Accountability gaps; delayed incident handling.	Unresolved issues; erosion of trust.
Which are the core function/role/ stakeholders responsible?	MLOps/SRE; Safety officer; Product manager	Governance board; Legal; Safety officer; Department head	QA; Clinical champions; Product owner
Specific requirements?	AI Act post-market monitoring; ISO 13485/14971 concepts	Hospital governance policies; AI Act risk management	Hospital QA; AI Act monitoring

Table 42: UC 1 – Practical measures to achieve Transparency with respect to Accessibility and Explainability

Technical measures to address Accessibility				**Technical measures to address Explainability**
Describe the measure	Provide patient-friendly summaries of AI-supported findings in the portal using plain language, visuals, and an accessible reading level.	Inform patients when AI contributed to the report with a short, clear statement about what the AI did and did not do, plus a contact point for questions.	Limitations clearly communicated to doctors? Show model limitations at point of use: validated scope, contraindications, and uncertainty ranges.	Provide clinician-facing explanations such as overlays/centerlines, confidence, measurement provenance, and side-by-side comparisons with manual measurements.
Why is it relevant?	Supports trust, consent, and shared decision-making.	Builds transparency and meets disclosure expectations.	Prevents out-of-scope use and unsafe over-reliance.	Improves understanding and appropriate use.
How can it be achieved?	Translate results, include short FAQs, and test readability with target users.	Use standard disclosure text in reports, patient leaflets, and consent materials; keep it consistent across sites.	Provide a model card in the app, guardrail warnings, and quick reference cards for clinicians.	Use evidence panes, overlays, and “what changed” views; include short training on how to interpret them.
How can be assessed whether this measure has been fulfilled?	Patient comprehension surveys and reduced confusion or complaints.	Documented disclosures, patient surveys, and low complaint rates about AI use.	UI checks show limitations visible; incident reviews show no out-of-scope use.	Clinician comprehension surveys, usage metrics, and peer review outcomes.
What are (potential) challenges to fulfilment?	Health literacy and translation needs; avoiding information overload.	Alignment with hospital communication policies and legal review. Cost for translations.	Keeping content current after updates and across vendors. Maintain easy-to-use UI.	Information overload; different clinician preferences and time pressure.
What are risks if not fulfilled?	Distrust, confusion, and poor adherence.	Patients unaware of AI involvement and consent concerns.	Out-of-scope use; patient harm; liability.	Misinterpretation or over-trust in AI outputs.
Which are the core function/role/ stakeholders responsible?	Patient engagement lead; Clinicians	Patient engagement lead; Clinicians; Legal	Product owner; Safety officer; Clinicians	Clinicians, product owner, UX leads
Specific requirements?	GDPR transparency	GDPR transparency; AI Act Art. 13	AI Act instructions for use; IEC 62366-1	AI Act transparency

Table 43: UC 1 – Practical measures to achieve transparency with respect to Justifiability

Transparency: Technical measures to address Justifiability	
Describe the measure	Enable contextual justification by linking AI outputs to evidence and clinical guidelines; provide a structured note field so clinicians can record the rationale in plain language.
Why is it relevant?	Supports accountability and defensible clinical decisions for peers, patients, and auditors.
How can it be achieved?	Provide short guideline summaries, checklists, and templated justification fields embedded in the reporting workflow.
How can be assessed whether this measure has been fulfilled?	Reports include justification notes; random audit samples show adequate rationale; peer review confirms quality.
What are (potential) challenges to fulfilment?	Added workload, variation in writing style, and limited standardization.
What are risks if not fulfilled?	Decisions appear unsupported; audit and complaint risk.
Which are the core function/role/ stakeholders responsible?	QA, clinicians, governance board
Specific requirements?	AI Act, guidelines, clinical documentation standards

PRACTICAL MEASURES PROVIDED BY USE CASE 2

Table 44: UC2 – Practical measures to achieve robustness with respect to Technical Robustness and Resilience

Technical measures to address Technical robustness and resilience				
Describe the measure	UI/UX for correcting the report	Elimination of Batch Approvals	Data quality, integrity, and validation	Resilient architecture
Why is it relevant?	Such interfaces make the system (self)-correcting through human supervision, a fundamental property of robust socio-technical systems. It also must mitigate the cognitive filtering burden caused by high volumes of "Correct but Useless" noise, which can make up to 45% of the AI's suggestions	It directly mitigates "automation complacency". The audit proved that under time pressure, "completion-driven" engineers adopted a "batch validation" mindset and missed 100% of high-priority seeded errors	Guarantee that the inputs used for training and inference are accurate, consistent, and representative of real-world use contexts.	To ensure that one faulty component does not compromise the entire safety analysis; to separate components clearly and to limit fault propogation; to include mechanisms for version control, rollback, and adaptive re-validation.
How can it be achieved?	Display confidence scores or uncertainty metrics for each AI-generated safety statement. Highlight low-confidence outputs using color-coded cues (e.g. green = validated, amber = pending review, red = disputed). to assist the engineer in identifying where human judgment is most critical	Remove UI features that allow bulk approvals and enforce a strict protocol where any rating other than "Correct and Useful" requires a mandatory justifying comment.	Full traceability of datasets (source, preprocessing, labelling, and version control); datasets covering the intended operational domain.	By following ISO 26262 and ISO 21448 (SOTIF); by using modular and layered design; by introducing redundancy and diversity.
How can be assessed whether this measure has been fulfilled?	Checking the implementation in the analysis software.	Through review of system logs confirming individual, time-stamped ratings and justifications for every UCA before approval.	Comparison of AI inferences with expert-labelled safety cases; data management plan; data validation reports	By checking system design documentation (architecture diagrams, redundancy layers, fallback logic, safety goals).

What are (potential) challenges to fulfilment?	Data exchange between AI engines, report generators, and human-validation modules must remain deterministic and auditable. Generating clear explanations in the UI without cognitive overload is a design challenge. Engineers experience mental fatigue from reading repetitive or generic entries, increasing the likelihood of skipping critical errors under time pressure.	It introduces workflow friction and increases the time required for analysis, which faces resistance in high-pressure environments where rapid software releases are prioritized.	Heterogeneous and fragmented data sources; incomplete or noisy data; lack of ground truth for validation	Designing multi-layer, redundant, and self-monitoring systems introduces high complexity in both hardware and software; Realistic resilience testing (e.g. sensor loss, corrupted data, cyberattack) is expensive and time-consuming.
What are risks if not fulfilled?	Potentially unreliable results/ decisions; operators may ignore AI alerts ("alarm fatigue") or accept outputs uncritically ("automation bias"). Engineers can suffer output fatigue, adopt a path of least resistance, and fail to verify specific safety logic.	Engineers will skim baseline documentation, leading to the acceptance of residual faults and logically flawed inputs that could result in unsafe vehicle behaviour	Inaccurate or unreliable safety analysis; propagation of hidden bias and noise; audit and certification failure	Unsafe or misleading safety reports; potential for accidents or delayed hazard detection
Which are the core function/role/ stakeholders responsible?	System architects, Safety engineers, Functional safety engineers, System engineers, Software developers responsible for implementing transparent, interpretable interfaces that enable human review and correction of AI-generated safety analyses.	System architects, UI/UX designers, software developers, and safety managers.	Data engineers, Data analysts, ensure dataset traceability, quality assurance, and compliance with data management requirements under ISO 26262, ISO/PAS 21448, and EU AI Act.	System architects, Safety engineers, Functional safety engineers, System engineers, and DevOps engineers design and verify modular, fault-tolerant architectures with proper version control, rollback, and redundancy mechanisms.
Specific requirements?	Standards and regulations (see Identification of components)	Compliance with ISO 26262 and ISO 21448 (SOTIF) requirements for human oversight.	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)

Table 45: UC2 – Practical measures to achieve robustness with respect to Technical Robustness and Resilience

Organisational measures to address Technical robustness and resilience			
Describe the measure	Safety culture in the organisation (adherence to process, attitude, training...)	AI quality management system	Cross-disciplinary reviews
Why is it relevant?	Promotes consistent human oversight, adherence to safety standards, and responsible use of AI tools. Reduces the likelihood of errors and oversight gaps in STPA analysis	Ensures AI-assisted safety analysis (e.g., UCAs and safety requirements) is reliable, reproducible, and compliant with industry standards	Allows diverse perspectives to identify missed hazards, hidden biases, or process gaps, improving robustness, fairness, and safety accountability
How can it be achieved?	Regular training sessions on STPA and AI-assisted analysis. Promoting a culture of safety-first decision-making; clearly documented roles and responsibilities	Establish SOPs for AI use, version control of datasets and models, validation of AI outputs, audit trails, and periodic reviews	Schedule formal review workshops, document decisions, include structured disagreement analysis, and involve multiple stakeholders for each safety-critical deliverable
How can be assessed whether this measure has been fulfilled?	Training attendance logs, adherence audits, observation of process compliance during STPAi workflow	Internal audits of AI processes, review of change management logs, assessment of AI output traceability and QA checks	Minutes of meetings, documented corrective actions, evidence of stakeholder engagement and consensus on critical safety decisions
What are (potential) challenges to fulfilment?	Resistance to change, inconsistent adherence to processes, lack of awareness of AI-specific risks	Integrating AI processes into existing quality management systems, maintaining up-to-date validation datasets and documentation	Scheduling conflicts, insufficient expertise across disciplines, potential for groupthink or overlooked biases
What are risks if not fulfilled?	Human errors in reviewing AI outputs, unsafe safety decisions, compliance issues	AI outputs may be inaccurate, biased, or non-compliant, safety cases may be unreliable	Missed hazards, uncorrected biases, lack of accountability, reduced trust in AI-assisted safety analysis
Which are the core function/role/ stakeholders responsible?	Safety managers, Functional safety engineers, Project leads, HR/Training coordinators	AI developers, Functional safety engineers, QA and compliance reams, Project safety leads	Safety engineers, Functional safety engineers, Regulatory/ Compliance officers, Project managers
Specific requirements?	ISO 26262, ISO/PAS 21448 (SOTIF), EU AI Act (2024), internal corporate safety policies	ISO 26262, ISO/PAS 21448 (SOTIF), EU AI Act (2024), ISO 9001/ISO 13485, ASPICE (quality management)	ISO 26262, ISO/PAS 21448 (SOTIF), EU AI Act, internal governance policies for cross-functional safety reviews

Table 46: UC2 – Practical measures to achieve robustness with respect to Reliability through lifecycle testing and monitoring

Technical measures to address Reliability through lifecycle testing and monitoring			
Describe the measure	Integrating, in the AI system, the safety process including the STPAI tool	Continuous performance validation	Versioned lifecycle traceability
Why is it relevant?	By embedding the safety process (e.g., STPAI) into the AI system (lifecycle), one ensures that every lifecycle stage — from data acquisition to model updates — is guided by explicit safety constraints and hazard analyses.	Ensures the AI model continues to produce accurate, consistent, and safe results as new data or conditions arise throughout its lifecycle.	Enables reproducibility, accountability, and compliance by linking every model version to its dataset, parameters, and validation evidence.
How can it be achieved?	By embedding safety reasoning, hazard traceability, and control feedback into every phase of AI development and deployment.	Regular validation cycles (using benchmark datasets, and expert-labelled cases.	Version control tools to record dataset versions, model weights, code commits, and validation reports; formal change management procedures.
How can be assessed whether this measure has been fulfilled?	Existence of a Safety & AI Lifecycle Integration Plan, traceable approval workflow, periodic safety review meetings where AI behaviour and STPAI findings are jointly discussed	Validation reports showing maintained or improved accuracy across time and domains; approved by a human safety expert or quality assurance reviewer.	Traceability logs and audit trails; ability to reproduce results or reconstruct past versions.
What are (potential) challenges to fulfilment?	Opaque models, lack of toolchain integration, limited data coverage	Requires maintaining large, up-to-date test datasets; time and resource constraints for periodic evaluation.	Maintaining documentation discipline; integrating diverse data and software repositories.
What are risks if not fulfilled?	Losing end-to-end traceability, human oversight, and regulatory defensibility.	Model degradation goes undetected; system reliability declines.	Inability to reproduce or justify safety analysis decisions; regulatory non-compliance; loss of accountability.
Which are the core function/role/ stakeholders responsible?	Functional safety engineers, AI developers, and Safety managers are responsible for embedding safety workflows and constraints into the AI development pipeline. They ensure that hazard identification, control feedback, and risk assessment are continuously linked to AI updates.	V&V engineers, QA teams, are in charge of setting up recurring test and evaluation cycles, maintaining benchmark datasets, and interpreting drift or anomaly reports.	Configuration management, Data managers, and Project safety leads are responsible for managing repositories (code, models, datasets, documentation), versioning all artifacts, and maintaining auditable links across lifecycle stages.
Specific requirement to consider?	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)

Table 47: UC2 – Practical measures to achieve robustness with respect to Technical Robustness and Resilience

Technical measures to address Fairness in safety assurance				
Describe the measure	Diverse and representative training data	Bias testing and model auditing	Multi-stakeholder validation	Continuous monitoring of fairness drift
Why is it relevant?	A step towards systemic bias prevention (in data, models, and decision weighting)	A step towards systemic bias prevention (in data, models, and decision weighting)	A step towards systemic bias prevention (in data, models, and decision weighting)	A step towards systemic bias prevention (in data, models, and decision weighting)
How can it be achieved?	Datasets used for model training reflect diversity in operational conditions (e.g. vehicle types, road environments, weather, lighting, sensor configurations, human operator profiles. Use of data audits to detect underrepresented cases (edge scenarios, rare hazards).	Through tests for systematic deviations (e.g., always underrating specific hazard types); through the use sensitivity analysis to check fairness across different input distributions.	Through the inclusion of diverse expert groups in validating tool outputs. Through "disagreement analysis" sessions — if AI and humans diverge, identify bias sources	Through a fairness drift detector which triggers model retraining if imbalance occurs.
How can be assessed whether this measure has been fulfilled?	Data source traceability audit	Existence of a bias testing plan document (versioned and signed). Implementation records in MLOps pipeline logs (e.g., "bias_test.py" execution results). QA sign-off from the AI assurance lead	Minutes of validation workshops after each major tool iteration.	Through a review of system architecture and design reports; by checking AI model validation plan and monitoring logs.
What are (potential) challenges to fulfilment?	Constantly evolving operational contexts, manual audits costly	Lack of access to sufficiently diverse safety data; no standard fairness metrics in safety analysis; evolving operating conditions and system architectures	Absence of unified metrics for AI reliability; fragmented responsibility; Traditional safety validation (e.g. FMEA, STPA) assumes deterministic logic, while AI systems introduce probabilistic outputs and adaptive behaviour.	Real-world safety analysis data are uneven — some scenarios (e.g. urban pedestrians, night-time incidents) are underrepresented; automotive AI pipelines are often designed for performance monitoring, not fairness or ethical indicators; continuous monitoring requires significant compute, logging, and storage resources to process ongoing data streams.

What are risks if not fulfilled?	The AI model performs well only on the narrow conditions it was trained on; under new environments, configurations, or vehicle types, it produces unreliable safety assessments; the AI tool consistently misclassifies or deprioritizes certain hazard types, which may lead to blind spots in safety analysis.	Hidden bias in AI model; non-compliance with AI Act Article 15; lack of accountability.	Validation limited to a single stakeholder's data or perspective may fail to capture real-world variability; in the absence of documented multi-stakeholder validation, it becomes impossible to identify who is responsible for faulty safety analyses.	Undetected bias accumulation; degradation of model performance over time; discriminatory safety outcomes.
Which are the core function/role/ stakeholders responsible?	Safety engineers and Functional safety engineers define hazards, losses, and system constraints; responsible for ensuring inputs represent relevant operational conditions.	Safety engineers and Safety managers review AI-generated UCAs, rate them, and correct errors; maintain traceability and ensure outputs are aligned with safety standards.	Safety engineers, Compliance officers and Safety managers are planned participants for broader validation beyond individual engineers	AI developers, Software engineers, Safety lead would implement pipelines for monitoring AI fairness over time.
Specific requirement?	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)

Table 48: UC2 – Practical measures to achieve Human Oversight and Autonomy with respect to Human oversight and controllability

Technical measures to address Human oversight and Controllability			
Describe the measure	Human-in-the-loop review of safety-critical outputs	Error Injection with Report Blocking	Manual override and safe-fail controls
Why is it relevant?	Keeping a human expert in the validation loop ensures that AI-generated hazard analyses, risk categorizations, or failure reports are critically reviewed before use in official documentation. This approach mitigates automation bias, and ensures accountability. It actively combats "automation complacency," where engineers may over-trust "good enough" AI outputs and fail to identify subtle errors under tight software release cycles.	During audits, experts missed highly critical safety contradictions (e.g., setting the gear to reverse during forward parking) because they assumed the AI was correct. Active testing automation bias is required to verify human vigilance.	Providing humans with clear override options and safe-fail mechanisms preserves the human as ultimate decision-maker, maintaining operational safety and system trustworthiness.
How can it be achieved?	Through the integration of a formal approval workflow requiring human sign-off on all AI-produced safety assessments. Through UI/UX flagging uncertain or high-risk outputs for mandatory human review. Through the definition of a role-based access (e.g., safety engineer, reviewer) to ensure appropriate levels of verification. Elimination of "batch approval" functionalities to force interaction. Engineers must engage with each individual entry to prevent skimming.	Seed contradictory UCAs into the workflow and configure the system to halt certification output and require multifactor authorizations or a deeper review if the engineer categorizes the seeded error as "correct".	Through the implementation of manual override or “pause” functions within the AI tool, allowing operators to stop automated reasoning or revert to previous states. Through the definition of safe-fail protocols (when confidence thresholds are breached, the system halts or requests human confirmation). Through the inclusion of procedural guidance in standard operating procedures defining when and how to override AI recommendations. Through the training of users to recognize override scenarios and to apply recovery actions safely.
How can be assessed whether this measure has been fulfilled?	Review of audit trails confirming that human validation occurred for every safety-critical report; evaluation of compliance metrics (e.g., percentage of reports validated by humans).	System logs tracking error detection rates by users, and QA reviews of blocked reports.	By reviewing system logs showing correct activation of manual override functions. By auditing training records demonstrating that personnel have been instructed on override protocols.
What are (potential) challenges to fulfilment?	Risk of human complacency due to overreliance on automation. Balancing efficiency vs. thoroughness in human review	Designing realistic seeded errors that effectively test vigilance without overly frustrating engineers or disrupting valid automation flows.	Determining optimal intervention points without disrupting valid automation. Designing UI elements that clearly communicate override states.

	Need for clear documentation standards to maintain consistency across reviewers.		
What are risks if not fulfilled?	In a time-pressured environment, "completion-driven" engineers may adopt a "batch validation" mindset, missing up to 100% of high-priority errors in their haste to finish AI-generated safety assessments could include undetected errors or unsafe assumptions. Accountability gaps and liability exposure in case of accidents or regulatory inspections.	Without active testing, engineers develop a false sense of security (the "complacency paradox"), assuming AI outputs are flawless, leading to critical safety failures.	Loss of human control in case of AI malfunction or erroneous automation. Potential for unsafe propagation of incorrect analyses through the workflow.
Which are the core function/role/ stakeholders responsible?	Safety engineers; Functional safety engineers review and validate AI-generated UCAs, hazards, and safety requirements. Safety managers and QA Teams ensure compliance with human review procedures and audit trails; Project leads oversee review processes and enforce adherence to safety protocols.	Safety engineers; Functional safety engineers, QA teams, and safety managers.	System architects and Software developers implement manual override, pause functions, and safe-fail protocols in the AI tool. Human safety Engineers trained to use override and recovery functions appropriately; Safety managers and QA teams audit correct activation of overrides and ensure procedures are followed.
Specific requirements?	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)

Table 49: UC2 – Practical measures to achieve Human Oversight and Autonomy with respect to Accountability and Traceability

Technical measures to address Accountability and Traceability			**Organisational measure**
Describe the measure	Decision logging and audit trails	Version control for models, data, and analyses	Role-based responsibility assignment
Why is it relevant?	Decision logs allow auditors and safety managers to reconstruct what inputs, models, and reasoning paths led to a given safety recommendation or hazard classification.	Version tracking enables consistent management of updates, helps identify when model changes affect results, and supports transparent safety validation throughout the AI system's lifecycle.	Accountability in safety analysis requires knowing who made or approved each decision. Clear responsibility assignment ensures that safety-critical judgments undergo proper expert review, supports ethical governance, and aligns with EU AI Act Article 9 (Risk Management System) and ISO 26262 Part 2 requirements for safety management. Sign-off procedures make accountability explicit, enabling traceable oversight and preventing ambiguous ownership of safety outcomes.
How can it be achieved?	Automated logging mechanisms that record all AI-generated recommendations, human validations, overrides, and contextual parameters (e.g., model version, dataset, timestamp).	Through the use dedicated AI lifecycle management tools; through the maintenance of metadata repositories and automated change tracking.	Through a role-based access control (RBAC) system assigning responsibilities (AI developer, safety engineer, reviewer, auditor). Through a digital sign-off required for each critical decision or report before final approval.
How can be assessed whether this measure has been fulfilled?	Existence of comprehensive log archives for all safety analyses conducted; audit of random samples to ensure all critical decisions have traceable justifications	Through a demonstration to reproduce any safety analysis using archived versions; through version identifiers in all safety analysis reports and dashboards.	Through verification that every final report includes at least one human reviewer's and one responsible engineer's digital signature. Through audit logs showing time-stamped validation and review actions.
What are (potential) challenges to fulfilment?	Managing large volumes of data generated by automated decision systems; risk of excessive documentation burden reducing agility.	Maintaining synchronization between datasets, models, and user interfaces. Resource demands for storage and documentation.	Potential resistance from users to new administrative responsibilities. Maintaining clarity when teams or roles change over time. Need to balance accountability with workload efficiency.
What are risks if not fulfilled?	Loss of accountability in case of accidents or safety non-compliance; inability to reconstruct or justify past safety analyses; legal vulnerability in product liability or conformity assessment.	Irreproducible results and unverified updates leading to invalid safety conclusions. Lack of clear responsibility for changes introduced to models or data.	Ambiguity over who approved safety analyses, leading to gaps in liability. Lack of traceable oversight undermining confidence in AI tool outputs.

		Increased risk of regulatory non-conformity and inability to pass safety audits.	Organisational confusion in case of errors, accidents, or external investigations.
Which are the core function/role/ stakeholders responsible?	Safety Engineers and Functional safety engineers ensure all AI-generated outputs, human validations, and overrides are logged. QA Teams and Safety managers review and audit logs to verify completeness and compliance. System architects and software developers implement automated logging infrastructure in the tool.	Configuration managers and Data managers manage datasets, model weights, and analysis artifacts. AI developers and DevOps engineers maintain version control systems and ensure proper tagging of releases. Safety managers Project leads will verify version consistency and compliance with safety standards.	Project leads and Safety managers define responsibilities and enforce role-based workflows. System administrators should enforce access, approval, and sign-off rules. Safety engineers execute tasks according to assigned roles and document approvals.
Specific requirements?	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)

Table 50: UC2 – Practical measures to achieve Human Oversight and Autonomy with respect to Transparency of safety critical performance

Technical measures to address Transparency of safety critical performance			**Organisational measure**
Describe the measure	Explicit performance boundaries and operating conditions	User communication and visualisation of model behaviour	Clear documentation
Why is it relevant?	Transparency about operational boundaries (e.g., accuracy, model confidence, data validity) ensures users know when the AI's predictions are valid and when human judgment must prevail.	If uncertainty is not clearly communicated, users may treat low-confidence results as reliable, leading to unsafe decisions. Visualizing uncertainty helps users interpret results appropriately, balance machine and human inputs, and comply with ethical principles of informed human oversight and robustness.	Transparent documentation ensures that all stakeholders — developers, safety engineers, auditors, and regulators — can understand the AI's intended use, known limitations, and validation evidence.
How can it be achieved?	Through the definition and documentation of Operational Design Domains for the AI tool (e.g., applicable system types, data sources, failure scenarios, or environmental conditions).	Through display warnings or fallback messages when the system operates outside its validated domain.	Through clear definition and documentation of performance metrics (accuracy, recall, false-positive rate) and confidence intervals for all outputs.
How can be assessed whether this measure has been fulfilled?	Through a review of the documentation of the system and testing.	Verify through usability tests that users correctly interpret uncertainty information; Audit user interface elements to ensure uncertainty is displayed for all safety-relevant outputs.	Through a review of the documentation.
What are (potential) challenges to fulfilment?	Defining clear and meaningful performance boundaries for complex, data-driven models. Ensuring ODDs remain up-to-date when the AI model is retrained or upgraded.	Quantifying uncertainty in complex or non-probabilistic AI models; Risk of user misinterpretation or cognitive overload if uncertainty displays are too complex.	High resource cost for maintaining documentation across multiple model versions; Technical difficulty in translating complex model behaviour into accessible language.
What are risks if not fulfilled?	Users may over-trust the AI beyond safe operational limits, leading to unsafe conclusions; Misinterpretation of AI outputs in unfamiliar contexts; Regulatory non-compliance with transparency and risk communication obligations.	Overconfidence in AI results leading to unsafe or non-compliant safety assessments. Inability to distinguish reliable from unreliable predictions; Loss of user trust and accountability during audits.	Poor understanding of AI limitations among users or auditors. Inconsistent safety decisions due to misinterpretation of model outputs.
Which are the core function/role/ stakeholders responsible?	System architects' Functional safety engineers will define Operational Design Domains (ODDs) and performance limits; AI developers implement model monitoring and enforce boundary checks in the AI system; Safety	UI/UX designers and Front-end developers design interfaces to communicate uncertainty, confidence scores, and operational limits; Safety engineers will validate that visualizations correctly convey risks and uncertainty.	Safety engineers and Project leads should ensure documentation accurately reflects safety-critical assumptions, validation results, and performance boundaries; Auditors, and QA Teams will

	managers will review boundary definitions and ensure compliance with safety standards.		review documentation for completeness and traceability.
Specific requirements?	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)

Table 51: UC2 – Practical measures to achieve Transparency with respect to Preservation of human skill and expertise

Technical measures to address Preservation of human skill and expertise			**Organisational measure**
Describe the measure	Regular human-in-the-loop training	Knowledge capture and mentorship programs	Adaptive training and continuous skill development
Why is it relevant?	Semi-automated safety tools risk deskilling engineers if humans become passive validators of AI-generated results rather than active problem-solvers. To preserve critical expertise, engineers must regularly exercise their analytical skills through independent reasoning and comparative validation.	Safety engineering depends on accumulated domain expertise and tacit reasoning, often gained through experience rather than formal training. The tool's ability to pre-fill analyses can lead to a deterioration of engineering expertise over time, turning practitioners into passive validators rather than active problem solvers.	Engineers must continuously adapt to new tools, validation methods, and data-driven processes while retaining their analytical mindset. Continuous training programs ensure humans remain effective supervisors, capable of understanding AI logic and detecting deviations.
How can it be achieved?	Include explanation and reasoning comparison sessions, where human conclusions are contrasted with AI outputs. Maintain feedback loops to improve both AI models (learning from human insights) and human expertise (learning from model reasoning). Integrate these activities into the continuous professional training plan for safety teams.	Implement structured mentorship programs pairing senior safety engineers with junior safety engineers to ensure the continuity of tacit reasoning and domain-specific knowledge that AI agents currently cannot replicate.	Through a modular training program combining technical (AI literacy, model interpretation) and domain-specific (safety engineering) modules. Through an interactive e-learning platforms for AI explainability and safety validation skills.
How can be assessed whether this measure has been fulfilled?	Evaluate engineers' skill retention through periodic technical assessments.	Review recorded knowledge assets (case analyses, "lessons learned" files) stored in internal repositories. Conduct surveys assessing staff perception of learning continuity and expertise transmission.	Through pre/post-training performance assessments in applied safety tasks or simulated decision-making. Through skill evolution evaluation through regular internal examinations or peer reviews.

What are (potential) challenges to fulfilment?	Additional workload and time pressure on safety engineers. Difficulty in designing fair comparisons across varied expertise levels. Ensuring organisational support and recognition for training efforts.	Limited time availability of senior experts for mentoring. Difficulties in codifying tacit knowledge (non-verbal or intuitive expertise). Risk of knowledge silos if documentation practices are not standardized.	Difficulty balancing training time with project delivery deadlines. Keeping training content up to date with evolving AI and safety standards. Ensuring engagement and motivation for continuous learning.
What are risks if not fulfilled?	Progressive loss of tacit safety expertise among engineers. Overreliance on AI leading to acceptance of erroneous outputs. Reduced human capacity to intervene effectively during unexpected system behaviours.	Discontinuity of organisational safety knowledge and experience. New engineers may lack contextual judgment for interpreting AI recommendations. Diminished innovation and learning capacity within safety organisations.	Skills lag between human operators and AI system capabilities. Reduced ability to detect anomalies or model drift in safety evaluations.
Which are the core function/role/ stakeholders responsible?	Safety engineers, Functional safety engineers and Training coordinators are responsible for planning, executing, and reviewing human-in-the-loop training sessions.	Safety engineers, Safety managers and Training managers are responsible for knowledge capture, mentorship execution, and monitoring knowledge continuity.	Safety engineers, AI developers, Training coordinators, QA teams are responsible for designing adaptive training, monitoring skill evolution, and integrating lessons learned into continuous skill development programs.
Specific requirements?	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)

Table 52: UC2 – Practical measures to achieve Transparency with respect to Feedback and learning loops for human adaptation

Technical measures to address Feedback and learning loops for human adaptation			Organisational measure
Describe the measure	User feedback capture and iterative interface refinement	Continuous human-AI performance monitoring and calibration	Organisational learning from human-AI Interaction outcomes
Why is it relevant?	Capturing feedback from engineers on usability, interpretability, and workflow friction ensures the AI evolves alongside human needs. Regular iterations based on user input enhance trust, prevent cognitive overload, and improve safety decision-making accuracy.	Monitoring how operators interpret, accept, or override AI recommendations identifies both over-reliance and under-use. It enables calibration of trust and ensures the AI remains a decision-support tool rather than a decision-maker. This continuous loop reinforces accountability and technical robustness throughout the system's lifecycle.	Systematic analysis of how humans and AI jointly perform safety tasks allows the company to refine procedures, training, and governance. Embedding feedback loops at the organisational level fosters resilience, transparency, and improvement of safety culture.
How can it be achieved?	Through in-tool feedback widgets Through a feedback database linking comments to AI model versions and updates.	Through monitoring dashboards tracking user interactions with AI (acceptance rates, override frequency, error correction patterns). Through explainability metrics (e.g., time to comprehension, correction rate) to detect drift in mutual understanding.	Through the integration of findings into process improvement documents, safety standards updates, and training materials. Through knowledge repositories with case studies highlighting both successful and problematic human-AI collaboration. Through organisational learning sessions and internal workshops.
How can be assessed whether this measure has been fulfilled?	Review design iteration logs to verify that user comments led to implemented improvements. Conduct usability testing pre- and post-iteration to quantify gains in clarity and satisfaction.	Through monitoring data to verify ongoing tracking of human-AI interactions.	Through the assessment of participation and feedback from employees on shared learning activities
What are (potential) challenges to fulfilment?	Collecting feedback without overburdening engineers during their workflows. Balancing conflicting suggestions from multiple user groups. Converting qualitative feedback into actionable technical changes.	Complexity of interpreting metrics	Need for dedicated resources and leadership commitment
What are risks if not fulfilled?	Persistent usability issues may cause erroneous safety assessments.	Diminished accountability if human decisions cannot be traced or evaluated.	Organisational stagnation and inability to evolve with technological change. Weakened ethical governance and reduced resilience to future challenges.

Which are the core function/role/ stakeholders responsible?	UX/UI designers, Safety engineers, AI developers are responsible for collecting feedback, analysing usability issues, and iteratively updating the interface to improve clarity, interpretability, and workflow efficiency	Safety engineers, Functional safety leads, QA engineers, AI developers are responsible for monitoring human-AI interactions, tracking overrides and corrections, evaluating trust levels, and calibrating AI outputs to support effective human oversight.	Organisational leaders, Safety managers, AI team leads, HR/training coordinators, and Process improvement teams are responsible for collecting insights from human-AI interactions, embedding them into organisational processes, and facilitating learning activities.
Specific requirements?	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)

Table 53: UC2 – Practical measures to achieve Transparency with respect to Training, education and continuous skill development

Organisational measures to address Training, education and continuous skill development		
Describe the measure	Structured onboarding and foundational AI-for-safety training	Continuous professional development and certification renewal
Why is it relevant?	Introducing AI into safety-critical workflows requires that engineers understand both the technology and its limits. Foundational training ensures users can interpret AI recommendations, identify potential biases, and maintain responsibility for safety decisions.	AI systems and safety standards evolve rapidly. Continuous professional development ensures that human experts remain aligned with new tools, regulations, and ethical norms. Regular renewal of competence maintains both safety assurance and workforce adaptability over the AI system's lifecycle.
How can it be achieved?	Implementation of a mandatory onboarding curriculum covering AI reliability, explainability, and specifically the risks of automation bias and complacency. This ensures engineers understand the boundaries and limitations of the tool before using it in production environments.	Through a CPD framework with annual or bi-annual refresher courses on updated ISO, UNECE, and EU AI Act requirements; new AI model versions, functionalities, or risk controls; emerging ethical and cybersecurity considerations. Through the implementation of learning credits or certification renewals tied to participation.
How can be assessed whether this measure has been fulfilled?	Through pre-/post-training tests measuring comprehension of AI functions and limitations. Throughout the audit of the onboarding documentation for completeness and updates.	Through completion rates of CPD modules and renewal certifications.
What are (potential) challenges to fulfilment?	Allocating time and resources amid tight project schedules. Maintaining up-to-date materials as AI models evolve.	Sustaining employee motivation for ongoing learning. Balancing training time with project delivery.
What are risks if not fulfilled?	Over- or under-reliance on automated outputs.	Skills obsolescence and regulatory non-compliance. Inconsistent application of safety procedures across teams.
Which are the core function/role/ stakeholders responsible?	Training coordinators, HR and L&D teams, Safety managers, AI tool experts, and Senior safety engineers responsible for designing, delivering, and evaluating onboarding programs.	HR/L&D teams, Safety managers, AI tool experts, and Safety engineers responsible for maintaining CPD frameworks, scheduling refresher courses, and tracking certification renewals.
Specific requirements?	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)

Table 54: UC2 – Practical measures to achieve Transparency with respect to Shared responsibility

Technical measure to address shared responsibility		Organisational measure to address shared responsibility
Describe the measure	Integration of shared responsibility into quality and incident management systems	Establishment of an AI governance and oversight committee
Why is it relevant?	Embedding shared responsibility within existing quality management and incident reporting systems ensures that accountability is not treated as an abstract ethical principle but as an operational practice. When safety anomalies or model errors occur, clear escalation and reporting pathways help identify root causes — whether human, procedural, or algorithmic.	A formal oversight structure ensures that shared responsibility is managed across disciplines, integrating technical experts, human-factors specialists, data scientists, and compliance officers. It institutionalizes collective accountability and ensures that ethical and safety decisions are reviewed through multiple lenses. This acts as a crucial safeguard against organisational pressures, as 80% of experts from the conducted survey believe management could misuse the system to prioritize productivity and speed over necessary safety rigor.
How can it be achieved?	Through an update of the Quality Management tools and documentation.	Define the committee's mandate to review the tool's impact on safety decisions, monitor ethical adherence, and protect engineers from operational pressures that demand speed over safety or use the tool to justify skipping manual checks
How can be assessed whether this measure has been fulfilled?	Through a review of Review updated quality management procedures for inclusion of AI-related responsibilities. Through a review of the associated tools.	Then periodic review sessions.
What are (potential) challenges to fulfilment?	Adapting legacy quality systems to accommodate AI-enabled workflows.	Through a review of documented decisions and policy updates issued by the committee. Overcoming resistance from management driven by the demand for rapid software releases and tight project deadlines.
What are risks if not fulfilled?	Poor visibility into the causes of safety incidents involving AI systems.	Fragmented governance and inconsistent application of safety or ethics principles. Reduced trust among stakeholders in the integrity of AI-assisted analyses. Management push for efficiency turns the AI into a tool for procedural compliance rather than substantive safety reasoning, compromising the integrity of the vehicle safety case.
Which are the core function/role/ stakeholders responsible?	Quality managers, Safety managers, Process owners, and Compliance officers are responsible for integrating shared responsibility into quality and incident management systems, defining clear escalation pathways, and ensuring accountability for both human and AI-related safety decisions.	Safety engineers, Quality managers, Compliance officers, Data scientists, and Human Factors specialists are responsible for defining AI governance policies, reviewing AI-related decisions, and ensuring shared responsibility across disciplines.
Specific requirements?	Standards and regulations (see Identification of components)	Standards and regulations (see Identification of components)

PRACTICAL MEASURES PROVIDED BY USE CASE 3

Table 55: UC3 – Practical measures to achieve diversity with respect to Non-bias, Fairness, and Non-discrimination

Technical measures to address Diversity				
Describe the measure	Diverse training data sampling	Fairness-aware modelling pipelines	Bias detection dashboards	Data quality monitoring across subgroups
Why is it relevant?	It reduces the risk that the model learns patterns mainly from overrepresented groups, which would mis-estimate vulnerability for less represented roles, locations or contract types.	It builds fairness checks directly into model development, so issues affecting specific groups are detected and mitigated before deployment, not only after problems appear in production.	They provide continuous visibility over how scores and errors are distributed across groups, making it easier to detect emerging disparities and trigger corrective actions in time.	It prevents situations where some groups systematically have poorer or missing data, which could lead to unreliable scores or exclusion from protections for those employees.
How can it be achieved?	Define representation targets per role/location. Use stratified sampling or weighting in the data pipeline. Regularly compare dataset composition against the workforce profile.	Integrate fairness libraries into training (e.g., evaluating subgroup accuracy). Use cross-group validation splits. Block model deployment if subgroup performance falls below thresholds.	Implement automated subgroup metric reports (e.g., false positives by role). Visualise group comparisons on monitoring dashboards. Configure alerts for statistically significant disparities.	Track missingness and anomalies per group. Run automated data completeness checks before training. Flag groups with systematically incomplete or lower-quality records.
How can be assessed whether this measure has been fulfilled?	Compare dataset demographics/roles/sites against the organisation's workforce profile. Check documented evidence of stratified sampling or weighting. Verify that under-represented groups are present above minimum thresholds.	Confirm model training logs include subgroup performance metrics. Verify that fairness thresholds were evaluated during development. Inspect model cards or documentation showing cross-group validation results.	Inspect dashboards for active subgroup comparisons (e.g., error rates). Confirm alerts or triggers are functioning when disparities appear. Review monthly/quarterly monitoring reports for bias analyses.	Check automated reports for missingness or anomalies per group. Compare data quality KPIs across roles/sites. Review remediation logs for groups flagged with low data completeness.

What are (potential) challenges to fulfilment?	Workforce composition may be highly unbalanced (e.g., many office workers, few frontline staff). Some groups may produce very little data (limited email usage, limited training interactions). Historical logs may not include diverse subgroups because the system was not previously used in all areas.	Limited technical expertise in fairness-aware ML techniques. Data protection constraints may limit use of sensitive attributes for fairness testing. Trade-offs between fairness and overall accuracy may be difficult to resolve.	Lack of resources to maintain or update monitoring tools. Difficulty defining meaningful subgroup categories (e.g., teams that change frequently). Low data volume for small groups can create noisy or inconclusive results.	Persistent data gaps for groups with different working patterns (e.g., offline or field workers). Multiple data sources with inconsistent formats reduce comparability. Lack of ownership: unclear who is responsible for addressing data gaps.
What are risks if not fulfilled?	System systematically overestimates or underestimates vulnerability for underrepresented groups. Certain roles or locations receive disproportionate alerts or attention. Hidden discrimination or disparate impact becomes embedded in the scoring logic. Increased exposure to AI Act non-compliance for high-risk systems	Models deployed with known or unknown disparities across groups. Potential harm to specific populations (e.g., frontline staff being flagged more often). Loss of employee trust in the system and perceptions of unfair treatment. Higher likelihood of complaints, escalations or regulatory scrutiny.	Slowly emerging disparities go unnoticed for long periods. Incorrect or biased scores influence HR or security decisions. Lack of evidence for audits or accountability reviews. Increased legal risk if negative impacts are discovered post-hoc.	Groups with incomplete data receive unreliable or unstable risk scores. Vulnerable populations may be excluded from protective measures. Systemic gaps mask operational or behavioural patterns for certain teams. Poor data quality undermines credibility of the entire system.
Which are the core function/role/ stakeholders responsible?	Data Science / ML Engineering. Data Engineering. IT / Security (as data owners) HR Analytics (for workforce representativeness checks)	Data Science / ML Engineering Responsible AI / Ethical AI teams Security Analytics Compliance / Risk Management (oversight)	Data Science / ML Engineering Data Analytics / BI teams Security Operations (SOC) / Cyber Risk Risk & Compliance (monitoring)	Data Engineering. Data Governance & Data Quality teams. HRIS (source of workforce metadata). Internal Audit (periodic verification).
Specific requirements?	Yes, EU AI Act (high-risk systems), GDPR, AI Management	Yes, EU AI Act (high-risk systems), GDPR, AI Management	Yes, EU AI Act (high-risk systems), GDPR, AI Management	Yes, EU AI Act (high-risk systems), GDPR, AI Management

	System Standards (e.g., ISO 42001, ISO 42005)	System Standards (e.g., ISO 42001, ISO 42005)	System Standards (e.g., ISO 42001, ISO 42005)	System Standards (e.g., ISO 42001, ISO 42005)

Table 56: UC3 – Practical measures to achieve diversity with respect to Non-bias, Fairness, and Non-discrimination

Organisational measures to address Diversity				
Describe the measure	Multistakeholder design reviews	Inclusive user research & testing	Diversity-focused risk assessment	Governance rules on model updates
Why is it relevant?	Bringing together HR, security, worker reps, D&I and legal ensures that diverse perspectives and use cases are considered, reducing blind spots and reinforcing legitimacy and trust.	It validates that the system works and is understood across different groups (e.g. frontline vs office, high vs low digital literacy), and helps identify design choices that could unintentionally disadvantage some users.	It embeds diversity and disparate-impact analysis into formal risk management, aligning with AI Act and GDPR expectations and documenting how risks to specific groups are identified and addressed.	They ensure that each significant change to the model triggers a new check of subgroup performance, preventing regression where an update improves overall accuracy but harms specific groups.
How can it be achieved?	Create a governance committee with HR, security, compliance, D&I and worker reps. Conduct structured review sessions at design, pre-deployment and update stages. Document concerns and design adjustments in minutes.	Recruit users from diverse roles, shifts, locations and literacy levels for testing. Conduct usability sessions in multiple languages when needed. Document insights and integrate them into design iterations.	Add a template section for subgroup impact analysis. Evaluate risks of disparate impact using test data and feedback. Include mitigations and follow-up actions in the DPIA and EU AI Act high risk systems guidelines.	Implement a mandatory fairness review before each major model update. Require approval from the governance committee prior to deployment. Keep model cards documenting subgroup performance across versions.
How can be assessed whether this measure has been fulfilled?	Review meeting minutes showing participation of HR, security, compliance, D&I, worker reps. Check review templates or decision logs documenting diversity-related concerns. Verify that identified issues have associated mitigation actions.	Inspect user research documentation showing participant diversity (roles, sites, languages). Check usability testing reports with feedback segmented by group. Confirm that design changes reflect insights from diverse users.	Verify that DPIA includes a dedicated analysis of subgroup risks. Check evidence that disparate-impact risks were evaluated with data. Confirm documented mitigations and follow-up actions.	Review model update documentation for recorded subgroup performance checks. Confirm governance approval (committee signatures or logs) before deployment. Validate versioned model cards detailing fairness comparisons across versions.

What are (potential) challenges to fulfilment?	Stakeholders may have conflicting priorities (security vs HR vs D&I). Some voices (e.g., worker reps) may be underrepresented or excluded.	Hard to recruit participants from all relevant groups, especially shift workers or field staff. Language diversity may require additional translation resources. Time constraints may limit the depth of testing.	Teams may lack expertise in identifying disparate-impact risks. Hard to quantify risks when subgroup sizes are small. DPIA may be treated as a compliance formality rather than a meaningful analysis.	Tight timelines for model updates may lead to bypassing fairness checks. Lack of ownership for fairness evaluation creates gaps in accountability. Documentation and model card updates may be overlooked under time pressure.
What are risks if not fulfilled?	Blind spots in system design affecting certain groups go undetected. Decisions reflect only the perspective of one department (e.g., security). Reduced employee trust and potential labour-relations conflicts. Missed compliance issues that another stakeholder could have caught.	Interfaces or communications may be inaccessible to some users. System misunderstood or misused by groups with different digital literacy levels. Increased likelihood of stigma or negative reactions among less-included groups. Missed opportunities to design for diverse working contexts.	Disparate-impact risks remain undocumented and unmanaged. Inability to demonstrate GDPR and AI Act compliance during audits. Employees may challenge automated decisions without clear justification available. Higher likelihood of harmful or discriminatory downstream effects.	Model updates degrade fairness without anyone noticing. Historical improvements in equality regress over time. No audit trail to explain why disparities suddenly increased. Regulatory exposure due to lack of documentation or oversight mechanisms.
Which are the core function/role/ stakeholders responsible?	Project Governance Committee / AI Governance Board. HR. IT/Security. Data Protection Officer (DPO). Worker representatives / Works Council. Diversity & Inclusion Office.	UX Research / UX Design. HR (facilitating access to diverse users). Security Awareness & Training teams. Employee Resource Groups / Worker reps (for diverse participation).	Data Protection Officer (DPO). Risk Management / Compliance Security / IT. HR. Legal Counsel. Responsible AI / Ethics teams (if available).	AI Governance Board or equivalent. Data Science / ML Engineering. Security / CISO Office. Compliance & Internal Audit. Legal / DPO (oversight on fairness and GDPR compliance).
Specific requirements?	Yes, EU AI Act (high-risk systems), GDPR, AI Management System Standards, Labour Regulation	Yes, EU AI Act (high-risk systems), GDPR, AI Management System Standards, UX Standards	Yes, EU AI Act (high-risk systems), GDPR, AI Management System Standards, Labour Regulation	Yes, EU AI Act (high-risk systems), GDPR, AI Management System Standards, Labour Regulation

Table 57: UC3 – Practical measures to achieve representativeness / inclusivity with respect to Non-bias, Fairness, and Non-discrimination

Technical measures to address representativeness / inclusivity		
Describe the measure	Group-representative data coverage	Subgroup performance & error analysis
Why is it relevant?	Ensures the system reflects how different groups actually experience phishing (e.g. frontline vs office, sites with different tooling, temporary staff). Reduces the risk that decisions are optimised for a "default office worker" and misrepresent vulnerability for other profiles.	Even with representative data, models can still have different error patterns across groups. Without subgroup analysis, you cannot see if some roles/sites are systematically over-estimated or under-estimated.
How can it be achieved?	Map the workforce structure. Set representation targets for training/validation data (e.g minimum share of each major region). Use stratified sampling or reweighting in the data pipeline to reach these targets where possible. Document known gaps (e.g. groups with little or no data yet) and treat them explicitly as limitations.	During validation and regularly in production, compute performance metrics (e.g. calibration, FPR/FNR, precision/recall) per subgroup (role, site, contract type, language, etc.). Build dashboards or reports that contrast metrics across groups. Define alert thresholds or triggers for investigation when differences exceed a certain level.
How can be assessed whether this measure has been fulfilled?	Compare dataset composition vs HR workforce statistics at least annually. Track coverage metrics: % of workforce included in training and in monitoring, by role/site/contract type. Record and review any groups flagged as under-represented against agreed thresholds.	Check that subgroup metrics are part of standard model validation and monitoring. Verify that reports are produced on a fixed cadence (e.g. quarterly) and reviewed in governance meetings. Confirm that disparities are logged with follow-up actions (e.g. further analysis, mitigation).
What are (potential) challenges to fulfilment?	Some groups generate very little digital trace (no regular email, shared accounts). Historical data may come from a subset of sites or roles only, especially in pilots. Legal and organisational constraints may limit use of certain contextual attributes	Small sample sizes for some subgroups make metrics noisy. L ack of access to sensitive attributes limits some fairness metrics (needs careful DPIA/FRIA).
What are risks if not fulfilled?	Scores generalised from one group (e.g. office staff) are inaccurate or unfair for others . Under-protection of groups that actually face higher exposure but have poor data coverage. Difficulty convincing employees and regulators that the system is fair and non-discriminatory.	Persistent higher false positives in some groups (more stigma, more unnecessary interventions). Higher false negatives in others (real risk not addressed). Increased exposure to claims of discriminatory impact, even if intent was neutral.
Which are the core function/role/	Data Science. Data Engineering. HRIS.	Data Science. BI/Analytics. Compliance.

stakeholders responsible?	Security. DPO.	Security.
Specific requirements?	EU AI Act (data representativeness), GDPR fairness & Art. 25.	EU AI Act (robustness, monitoring), GDPR fairness.

Table 58: UC3 – Practical measures to achieve representativeness / inclusivity with respect to Non-bias, Fairness, and Non-discrimination

Organisational measures to address representativeness / inclusivity		
Describe the measure	Engagement of diverse user groups in design & feedback	Governance process to review representativeness regularly
Why is it relevant?	People closest to the work can reveal practical constraints, cultural aspects and edge cases that designers miss. Ensures the system does not unintentionally disadvantage or alienate specific groups.	Workforce, tools and phishing tactics change over time; representativeness is not a one-off check. Regular review helps detect drift: the system gradually fitting some groups worse than others.
How can it be achieved?	Identify key groups to involve (e.g. frontline, remote, shift workers, multiple regions, varying digital skills). Run co-design workshops, interviews, and pilots with these groups before large-scale roll-out. Feed insights into requirements, UX, communication content and training design.	Integrate representativeness and subgroup performance into the agenda of the AI governance board or equivalent committee. At least yearly (ideally more often), review: data coverage vs workforce, subgroup metrics and error rates, feedback from user groups and contestations. Require documented actions and deadlines when gaps or issues are identified.
How can be assessed whether this measure has been fulfilled?	Review user research plans and reports: which groups were included. Check design documentation for traceability from feedback to design decisions. Ask worker reps or local HR whether they were consulted and how.	Governance terms of reference explicitly include representativeness/fairness. Minutes show that relevant metrics and feedback were discussed. Track whether action items (e.g. expand data, adjust models, change communication) were completed.
What are (potential) challenges to fulfilment?	Hard to free time for frontline/shift staff. Language and cultural differences may limit open feedback.	Lack of clear ownership for representativeness/fairness topics. Risk that reviews become formalities if KPIs and data are poor.
What are risks if not fulfilled?	System fits the reality of a small subset of workers only. Increased risk of resentment or resistance in neglected groups. Over time, unequal usability leads to unequal effectiveness and fairness.	Gradual degradation of fairness goes unnoticed. New business units, roles or sites get added without being properly reflected in data and design. Harder to show regulators that there is ongoing monitoring, not just initial compliance.
Which are the core function/role/	HR. Security Awareness. UX.	AI Governance board. Security. HR.

stakeholders responsible?	Worker reps. DPO.	Data Science. DPO. Compliance.
Specific requirements?	GDPR Art. 25 (DP by design), AI Act's human-centric design obligations	EU AI Act (post-market monitoring, change management), GDPR accountability.

Table 59: UC3 – Practical measures to achieve objectivity with respect to Non-bias, Fairness, and Non-discrimination

Technical measures to address objectivity			
Describe the measure	Standardised scoring and feature set	Empirical validation and benchmarking against ground truth	Automated consistency checks and controlled overrides
Why is it relevant?	To ensure that vulnerability levels are assigned based on consistent, evidence-based criteria rather than ad-hoc judgments or manual tweaks that could introduce bias.	To show that scores correlate with real-world phishing behaviour (e.g. simulated campaigns, reporting, training outcomes) and are not arbitrary or driven by spurious patterns.	To avoid subjective, undocumented manual interventions that undermine objective scoring (e.g. changing scores for particular users or teams without clear justification).
How can it be achieved?	By defining a documented scoring specification (features used, weighting, model type, thresholds), applying it uniformly across all users, and controlling changes through versioning and change	By defining quantitative validation metrics (e.g. calibration, precision/recall, ROC-AUC) and comparing predicted vulnerability to observed events across groups and time; by benchmarking against baselines.	By implementing rule-based consistency checks (e.g. score ranges, business constraints), restricting manual overrides to defined roles, and logging any changes with reasons.
How can be assessed whether this measure has been fulfilled?	By reviewing the formal scoring documentation, checking that implementation matches the specification, and verifying that there are no undocumented rules or "exceptions" for certain groups or individuals.	Through validation reports, test-set evaluations, and periodic re-validation, including subgroup-level performance results.	By auditing logs of overrides; checking that overrides are rare, justified and not concentrated in specific groups; and verifying that consistency checks are active in production.
What are (potential) challenges to fulfilment?	Legacy rules or manual exceptions; pressure from stakeholders to "tune" scores for specific groups; lack of discipline in documenting changes.	Limited or noisy ground truth data; changes in phishing tactics; mismatch between simulation data and real attacks; resource constraints for repeated validation.	Business pressure for exceptions. Lack of tooling for fine-grained access control and logging. Cultural acceptance of "manual fixes".
What are risks if not fulfilled?	Inconsistent scoring, hidden bias, difficulty explaining results to employees or	Scores may not reflect actual vulnerability, leading to misplaced controls, unfair labelling, or a false sense of security.	Hidden bias, favouritism or punitive use. Inability to explain why some users received different treatment.

	regulators, and increased risk of arbitrary or discriminatory outcomes.	Higher risk of ineffective or discriminatory interventions.	Non-compliance with transparency and accountability expectations.
Which are the core function/role/ stakeholders responsible?	Data Science. ML Engineering. Security / Risk teams. IT. Compliance / DPO (oversight).	Data Science / Analytics. Security (phishing / SOC teams). Risk Management. Compliance / Internal Audit (review).	IT / Platform Engineering. Data Science. Security / HR (as potential override users). Compliance / DPO.
Specific requirements?	EU AI Act requirements for technical robustness, documentation and traceability of high-risk AI systems; GDPR principles of fairness, transparency and accountability (Art. 5 and 25).	EU AI Act obligations on testing and monitoring of high-risk AI; GDPR accountability and data protection by design (Art. 25) requiring evidence that processing is appropriate and not excessive.	EU AI Act logging and post-market monitoring obligations; GDPR accountability and rights related to contesting automated decisions (Art. 22 and relevant recitals).

Table 60: UC3 – Practical measures to achieve objectivity with respect to Non-bias, Fairness, and Non-discrimination

Organisational measures to address objectivity		
Describe the measure	Policy on objective use and limitations of vulnerability scores	Independent review and validation of scoring methodology
Why is it relevant?	To make clear that scores are decision-support tools, not definitive judgments of character or performance, and to prevent arbitrary or excessive uses.	To ensure the methodology is scrutinised by a function that is not the direct developer/owner, reducing bias and blind spots.
How can it be achieved?	By issuing a written internal policy defining allowed uses (e.g. training prioritisation), prohibited uses (e.g. disciplinary decisions without further assessment), and the role of human judgment.	By assigning second-line or third-line functions (e.g. Risk, Compliance, Internal Audit) to review model design, data, performance and fairness, and to challenge assumptions.
How can be assessed whether this measure has been fulfilled?	Through existence and approval of the policy; evidence that processes (HR, security) reflect it. Interviews showing that managers understand the limitations.	Through validation reports, audit findings, and documented responses to recommendations. Evidence that go-live and major changes require independent sign-off.
What are (potential) challenges to fulfilment?	Misaligned expectations from management. "Function creep" of the system. Lack of communication or training on the policy.	Limited AI expertise in some business functions. Resource constraints. Potential tensions between speed-to-deploy and thorough review.
What are risks if not fulfilled?	Scores being used as de facto performance or trust metrics. Unfair HR consequences; loss of trust. Heightened risk of discrimination claims.	Methodological weaknesses or embedded bias go undetected. Regulators view the system as insufficiently governed. Increased risk of harmful or unfair outcomes.

Which are the core function/role/ stakeholders responsible?	HR. Security / CISO Office. Compliance / Legal. DPO. Senior management (policy owner).	Risk Management. Compliance. Internal Audit. DPO. AI Governance Board (if exists).
Specific requirements?	GDPR principles of purpose limitation and fairness; guidance from DPAs on employee monitoring and profiling; AI Act emphasis on appropriate, risk-based use of high-risk AI.	EU AI Act requirements for quality management and internal control; general governance principles in regulated sectors (e.g. model risk management in finance); GDPR accountability.

Table 61: UC3 – Practical measures to achieve non-stigmatising use / proportionality with respect to Non-bias, Fairness, and Non-discrimination

Technical measures to address non-stigmatising use / proportionality		
Describe the measure	Role-based access and aggregation of scores	Supportive, non-punitive action design (automated responses)
Why is it relevant?	Limiting who can see individual scores and favouring aggregated views reduces the risk that vulnerability information is used to label, shame or single out particular employees or groups.	If technical workflows automatically trigger punitive or escalatory actions, the system can become a tool for sanctioning rather than support, even if the model itself is unbiased.
How can it be achieved?	By implementing role-based access control (RBAC) so that only authorised functions (e.g. security/HR) can see individual scores; managers see only what they strictly need; and dashboards default to aggregated, anonymised or pseudonymised views where possible.	By designing score-based triggers to launch proportionate, supportive actions (e.g. tailored training, nudges, awareness campaigns) rather than sanctions. By requiring human review before any high-impact HR action; and by encoding “no-go” uses directly into the system logic.
How can be assessed whether this measure has been fulfilled?	By reviewing access control configurations, user roles and permissions; auditing access logs; and checking that most reporting interfaces use group-level or anonymised data by default.	By documenting all automated workflows; checking that technical rules link scores only to supportive measures; and auditing a sample of cases where scores influenced actions.
What are (potential) challenges to fulfilment?	Legacy systems with broad access. Pressure from management to see named rankings. Technical complexity of implementing fine-grained access and aggregation.	Organisational pressure to use scores for performance management. Function creep where notifications are gradually repurposed for disciplinary processes. Lack of clear mapping of which actions are allowed or banned.
What are risks if not fulfilled?	Scores may circulate informally, feeding stigma or reputational harm. Employees may be informally labelled “high risk” or “untrustworthy”. Increased risk of discrimination or workplace conflict.	Employees may be punished, excluded or informally downgraded based primarily on scores. Vulnerability assessment becomes de facto surveillance. Serious risk of discrimination and labour disputes.

Which are the core function/role/ stakeholders responsible?	IT / Identity & Access Management. Security. HR. DPO. Compliance / AI governance.	Product/solution owners. Security. HR. IT / Workflow owners. Legal / Compliance.
Specific requirements?	GDPR data minimisation and access limitation; GDPR Art. 88 and DPA guidance on employee monitoring; EU AI Act expectations around fundamental rights protection and appropriate, risk-based use of high-risk AI.	GDPR purpose limitation and fairness; guidance on automated decision-making and profiling; AI Act's risk-based approach and protection of fundamental rights.

Table 62: UC3 – Practical measures to achieve non-stigmatising use / proportionality with respect to Non-bias, Fairness, and Non-discrimination

Organisational measures to address non-stigmatising use / proportionality		
Describe the measure	Neutral, non-stigmatising presentation of scores in the UI	Policy on proportional, non-punitive use of vulnerability scores
Why is it relevant?	Labels, colours and wording in interfaces can strongly influence how people perceive themselves and others; stigmatising language or visual design can amplify blame and social pressure.	A clear policy sets boundaries so scores are used for support and risk reduction, not as hidden performance metrics or grounds for discrimination.
How can it be achieved?	By using neutral terminology (e.g. "current exposure level" instead of "weak user"); avoiding shaming visuals (e.g. red "danger" labels attached to names); providing contextual explanations focused on learning; and conducting UX testing on how employees perceive the interface.	By drafting and approving an internal policy that defines legitimate purposes (e.g. training prioritisation, awareness planning), explicitly prohibits certain uses (e.g. salary decisions, disciplinary action solely based on scores), and clarifies proportionality expectations.
How can be assessed whether this measure has been fulfilled?	Through UX reviews and content guidelines; user feedback sessions; checking screens for stigmatizing formulations or rankings; monitoring complaints or negative reactions to the interface.	By verifying existence and communication of the policy; checking alignment of HR and security procedures; and reviewing decisions or escalations for consistency with the policy.
What are (potential) challenges to fulfilment?	Tendency to use "traffic light" metaphors and competitive rankings; design patterns borrowed from performance dashboards; limited UX/ethics involvement.	Misalignment between security and HR priorities; ambiguity around "borderline" cases; insufficient dissemination of the policy to managers.
What are risks if not fulfilled?	Employees may feel shamed or humiliated; team dynamics can suffer; people may hide mistakes instead of reporting phishing, undermining security culture.	Function creep; use of scores for unfair HR decisions; higher likelihood of legal challenges, union disputes and loss of employee trust.
Which are the core function/role/ stakeholders responsible?	UX / Product Design. Security Awareness. HR. Ethics / Responsible AI. DPO (advisory).	HR. Security / CISO Office. Legal. DPO. Senior management (policy owner). Works Council / worker reps (where applicable).
Specific requirements?	GDPR fairness and data protection by design (Art. 25); AI Act's requirements for user-friendly information and fundamental-rights-respecting design; labour and occupational health norms on psychosocial risks may also be relevant.	GDPR purpose limitation, fairness and Art. 88 in the employment context; DPA guidance on employee monitoring; AI Act emphasis on proportional risk mitigation and fundamental rights.

Table 63: UC3 – Practical measures to achieve openness with respect to Transparency and Explainability

Technical measures to address openness		
Describe the measure	In-product transparency notices for employees	Data and processing registry for the system (data map)
Why is it relevant?	So that employees do not discover the system "by accident", but receive clear, timely information where they actually interact with it (e.g. training platform, phishing simulations, dashboards).	Openness towards employees and regulators depends on knowing what data is actually processed, from where, for what, and with which systems.
How can it be achieved?	Integrate short transparency notices and "learn more" links into the tools employees use: banners on training platforms, notices in phishing simulations, contextual info icons explaining that behaviour may be logged for security and training purposes.	Maintain a structured record of processing activities and data flows for the system: sources (e.g. simulations, logs, HR data), categories of data and recipients, retention periods, interfaces with other tools.
How can be assessed whether this measure has been fulfilled?	By reviewing the interfaces/screens; checking that notices appear before or at the time of data collection/use; validating the content against legal info requirements (purpose, data, recipients, rights).	By reviewing the registry or data map; checking consistency with what is communicated in privacy notices and transparency pages; verifying that it is used in DPIAs and AI governance.
What are (potential) challenges to fulfilment?	Limited UI real estate; tendency to hide information in long legal texts; inconsistent implementation across multiple tools.	Complex or changing technical architecture; multiple owners; legacy systems with undocumented data flows.
What are risks if not fulfilled?	Employees feel monitored without being informed, leading to distrust; potential non-compliance with GDPR transparency duties and AI Act transparency for high-risk systems.	Inability to explain processing to employees or authorities; gaps between reality and what notices say; higher risk of unlawful or excessive data use.
Which are the core function/role/ stakeholders responsible?	Product / platform owners. UX / UI design. Legal / DPO. Security awareness teams.	Data Governance. IT / Architecture; Security. DPO. Product owner. Compliance.
Specific requirements?	GDPR Arts. 12–14 on transparent information and privacy notices; AI Act obligations to inform users that they interact with or are subject to high-risk AI and what it does.	GDPR Art. 30 records of processing; data protection by design (Art. 25); AI Act data governance and documentation obligations.

Table 64: UC3 – Practical measures to achieve openness with respect to Transparency and Explainability

Organisational measures to address openness		
Describe the measure	Transparency & communication policy for vulnerability measurement	General employee information and consultation processes
Why is it relevant?	Ensures that employees who are personally identified as more vulnerable are addressed in a consistent, non-stigmatising, accurate way. Reduces the risk that managers or security staff overstate the meaning of a high score or create unnecessary fear.	Openness is not only about publishing information but about engaging with employees, answering questions and, where required, consulting representatives before introducing monitoring tools.
How can it be achieved?	Draft a short policy or guideline that covers, specifically for positive/high vulnerability cases: how to explain to an employee that they have been flagged as more vulnerable; what the intended consequences are (e.g. training, coaching, extra support); what the score does not mean.	Organise info sessions, Q&A, FAQs; involve worker representatives or works councils where legally required; ensure that new hires are informed as part of onboarding; set up channels for questions (HR, DPO).
How can be assessed whether this measure has been fulfilled?	Check that the policy explicitly refers to communication with employees flagged as high/positive for vulnerability. Review emails, templates, and guidance used in practice when notifying employees about their vulnerability status.	By reviewing records of information sessions and consultations; onboarding materials; FAQs; and feedback gathered from employees.
What are (potential) challenges to fulfilment?	Managers may be uncomfortable talking about employees being "high risk" and may improvise messaging.	Scheduling and reaching all groups (shift workers, field staff); overcoming mistrust; limited capacity for repeated sessions across locations.
What are risks if not fulfilled?	Employees who are flagged as more vulnerable receive confusing, harsh or inconsistent messages.	Employees may perceive the system as unilateral surveillance; unions or works councils may oppose deployment; weakened security culture and cooperation in phishing reporting.
Which are the core function/role/ stakeholders responsible?	HR. Communications. Security / CISO office. Legal / DPO; senior management. Works Council / worker reps where applicable.	HR. Works Council / worker reps. Security awareness. DPO. Communications.
Specific requirements?	GDPR transparency and information obligations; AI Act user information requirements; labour law and, in some countries, works council consultation duties.	GDPR, DPIA and consultation expectations; national labour laws and collective agreements (often requiring information/consultation for monitoring and scoring tools); AI Act focus on human oversight and user understanding.

Table 65: UC3 – Practical measures to achieve accessibility / access to information with respect to Transparency and Explainability

Technical measures to address accessibility / access to information		
Describe the measure	Multilingual and role-adapted transparency content	Self-service access to personal data, scores and explanations
Why is it relevant?	In multinational or diverse organisations, a single language or "generic office-worker" formulation will not be equally understandable for all employees (e.g. frontline staff, contractors, local-language workers).	Employees should not depend entirely on ad-hoc requests to understand what the system holds about them; a self-service portal supports ongoing access and exercises of rights.
How can it be achieved?	Provide transparency notices, FAQs and explanations in relevant working languages. Adapt examples and wording to different roles (frontline vs. office vs. management). Embed translations into the tools employees actually use (training platforms, intranet, mobile apps). Draft all texts in plain language, aiming for a reading age (e.g. short sentences, clear structure), and use basic readability tools during drafting to support this.	Provide a secure platform or interface where employees can see: their current vulnerability level, main factors, and key data categories used. Offer simple explanations and links to exercise rights (e.g. contact DPO, request rectification). Ensure appropriate authentication and access logging.
How can be assessed whether this measure has been fulfilled?	Check availability of translations for key documents and screens. Verify coverage of main languages/roles in the organisation. Gather feedback on comprehensibility from different locations and job families.	Existence and usability of the platform. Logs showing that employees access their own information. Internal testing of accuracy and completeness of what is shown.
What are (potential) challenges to fulfilment?	Limited translation budget or capacity. Keeping all language versions updated when the system changes. Over-simplification or loss of nuance in translations.	Technical complexity and security requirements. Aligning what is shown with legal and organisational constraints. Risk of overload if explanations are too technical or too detailed.
What are risks if not fulfilled?	Some employee groups remain uninformed or misinformed. Unequal ability to understand monitoring and exercise GDPR rights; Perceptions of second-class treatment for certain sites or roles.	Employees must resort to complex, slow rights-based requests for basic transparency. Lower trust in the system; perception that scoring is opaque and unaccountable. Difficulty demonstrating "easily accessible" information under GDPR.
Which are the core function/role/ stakeholders responsible?	HR (local HR, HR Ops). Corporate Communications. Security awareness teams. DPO / Legal (review of content consistency across languages).	IT / Product owners. Security / Identity & Access Management. HR. DPO / Legal (to ensure rights are properly reflected).

Specific requirements?	GDPR transparency principles (Arts. 12–14): information must be clear and understandable for the data subject. AI Act: user information must be adapted to the intended users, including their technical knowledge. National labour law or collective agreements may require information in specific languages.	GDPR rights of access and information (Arts. 12–15, 22 in practice). AI Act: instructions and relevant information to deployers/users for high-risk systems; Internal policies on employee monitoring and algorithmic transparency.

Table 66: UC3 – Practical measures to achieve accessibility / access to information with respect to Transparency and Explainability

Organisational measures to address accessibility / access to information		
Describe the measure	Structured process for handling GDPR information and access requests (employee-focused)	Accessibility KPIs for employee information
Why is it relevant?	Employees will sometimes want more detailed information than is available in portals or notices; a clear process avoids inconsistent or delayed responses.	Makes accessibility measurable, not just aspirational. Helps detect whether some groups (e.g. disabled employees, shift workers, non-desk workers, non-native speakers) are not being reached or cannot easily understand the information. Supports accountability and continuous improvement on transparency and accessibility, which is important for GDPR and AI Act compliance in practice.
How can it be achieved?	Define procedures and SLAs for access, rectification, objection and other rights requests related to the system. Provide contact points (HR, DPO) and templates for responding. Ensure coordination between HR, security and IT when retrieving data and explanations.	Define a small set of concrete KPIs, for example: % of sites/units that have received at least one targeted communication on the system. Assign ownership for collecting and reporting these KPIs. Integrate KPI review into periodic governance meetings for the system.
How can be assessed whether this measure has been fulfilled?	Existence of documented procedures. Logs of requests and response times. Quality checks on responses for completeness and clarity.	Check that a documented list of accessibility KPIs exists and is approved; there is a regular reporting cycle where KPIs are produced and discussed. Verify that KPI definitions, data sources and thresholds are clearly documented.
What are (potential) challenges to fulfilment?	Fragmented data across systems. Limited resources to process requests. Difficulty explaining technical aspects in understandable terms.	Difficulty collecting reliable data for some KPIs (e.g. feedback from small groups, or coverage for contractors). Risk of choosing only “easy” KPIs that do not reflect real accessibility. Limited analytical capacity or ownership to maintain the KPI reporting.

What are risks if not fulfilled?	Non-compliance with GDPR rights. Employee frustration and escalation to supervisory authorities or works councils. Perception of opacity and unfair treatment.	Lack of visibility on whether information about the system is actually accessible to all groups. Persistent blind spots where certain employees never receive or understand key information about monitoring and scoring. Difficulties demonstrating GDPR transparency and fairness in practice, especially for employees with disabilities or in non-standard roles.
Which are the core function/role/ stakeholders responsible?	DPO / Data protection office. HR. Security / IT (data retrieval). Legal / Compliance.	DPO / Data protection office. HR. Security / IT (data retrieval). Legal / Compliance.
Specific requirements?	GDPR Arts. 12–15, 21–22 regarding rights of access, rectification, objection and automated decision-making. Supervisory authorities' guidance on employee data and profiling.	GDPR accountability (Art. 5(2)): organisations must be able to demonstrate compliance. EU AI Act: for high-risk AI systems, emphasis on user-facing transparency and documentation that supports fundamental rights and effective oversight. National accessibility and anti-discrimination rules, especially where public-sector or equality obligations apply.

Table 67: UC3 – Practical measures to achieve documentation, traceability, and auditability with respect to Transparency and Explainability

Technical measures to address documentation, traceability, and auditability		
Describe the measure	Technical and functional documentation (system + model)	Versioning and configuration management (models and rules)
Why is it relevant?	Without clear documentation, it is very hard to explain the system to employees, regulators or auditors, or to show that fairness and data protection were considered in the design.	To know "which system" produced a given outcome at a given time, and to understand how changes may have affected fairness and accuracy.
How can it be achieved?	Maintain a "system dossier" or model card: purpose, data sources, features, training/validation, performance, limitations, version history. Include both technical (architecture, APIs) and functional (use cases, decisions supported) views.	Use version control (e.g. Git, model registry) for code, models and configuration. Record version identifiers in logs and documentation. Require change tickets and approvals for deploying new versions.
How can be assessed whether this measure has been fulfilled?	Check that documentation exists, is complete and up to date. Verify that what is described matches the actual implementation. Confirm that docs are used in DPIAs, AI governance and training.	Check existence of a model/Config registry. Confirm that each deployment is linked to a version and change record.
What are (potential) challenges to fulfilment?	Time pressure and lack of incentives for documentation. Multiple teams changing components without updating docs. Outsourced components with limited vendor transparency.	Legacy ad-hoc deployments. Multiple environments with inconsistent configurations.
What are risks if not fulfilled?	Inability to explain the system to employees or authorities. Hidden design choices that create fairness or GDPR issues. Higher risk of errors, inconsistent changes and "black box" perception.	Impossible to tie specific behaviours or issues to a particular version. Undetected regressions in fairness or robustness.
Which are the core function/role/ stakeholders responsible?	Product owner. Data Science / Engineering. IT Architecture. Security. DPO / Compliance	Data Science / ML Engineering. DevOps / MLOps. IT. AI Governance / Risk management.
Specific requirements?	EU AI Act: technical documentation obligations for high-risk AI systems. GDPR: accountability (Art. 5(2)), data protection by design (Art. 25) and records of processing (Art. 30).	EU AI Act: quality and risk management, post-market monitoring, documentation of changes. GDPR: accountability; updated DPIA if changes affect risk.

Table 68: UC3 – Practical measures to achieve documentation, traceability, and auditability with respect to Transparency and Explainability

Organisational measures to address documentation, traceability, and auditability		
Describe the measure	Periodic internal audits and reviews using system documentation and logs	Clear governance responsibilities for documentation and traceability
Why is it relevant?	Documentation and traceability only add value if they are used to verify that the system is behaving as intended and remains compliant over time.	Without clear ownership, key tasks (updating docs, maintaining logs, answering regulators) may fall between teams, leading to gaps.
How can it be achieved?	Include the system in internal audit / compliance monitoring plans. Define audit scope: data governance, fairness checks, access/use of scores. Use logs, docs and samples of decisions as evidence.	Define roles in an AI governance (owner, maintainer, reviewer). Integrate responsibilities into job descriptions and project charters. Ensure escalation paths for complex cases (e.g. DPO, legal).
How can be assessed whether this measure has been fulfilled?	Audit plans and reports explicitly mentioning the system. Action plans and follow-ups on findings. Evidence that issues are tracked and resolved.	Governance documents clearly naming system owner and support roles. Evidence that requests (from employees, regulators, audit) are handled in a coordinated way. Interviews confirming that teams know their responsibilities.
What are (potential) challenges to fulfilment?	Limited AI expertise in audit/compliance. Competing audit priorities.	Matrixed organisations with overlapping mandates. Turnover of key staff. "Shadow ownership" by teams that built the system but do not formally own it.
What are risks if not fulfilled?	Emerging issues (bias, misuse, security) go undetected. Documentation becomes "shelfware".	Slow or inconsistent responses to queries and audits. Loss of knowledge when people leave. Increased risk of non-compliance and reputational damage.
Which are the core function/role/ stakeholders responsible?	Internal Audit. Compliance / Risk management. DPO. Security. HR (for HR-impact aspects). Product owner.	AI Governance Board / equivalent. Product owner. Security / CISO. HR (for HR-impact aspects). DPO / Legal; Risk / Compliance.
Specific requirements?	EU AI Act: quality management and post-market monitoring. GDPR: accountability and ongoing monitoring of processing activities. Internal audit standards and model risk management practices.	EU AI Act: requirement for a quality management system and assignment of responsibilities for high-risk AI. GDPR: accountability (Art. 24) and role of controllers/processors and DPO.

Table 69: UC3 – Practical measures to avoid dependence with respect to Over-reliance

Technical measures to address dependence		
Describe the measure	Decision-support design	Multi-source decision views (no single-score dashboards)
Why is it relevant?	Prevents over-reliance by making it technically impossible to use the system as the single basis for important decisions affecting employees.	Encourages decision-makers to consider multiple sources of evidence, not just the score, reducing over-reliance on one metric.
How can it be achieved?	Configure workflows so that scores can only trigger low-impact, supportive actions automatically (e.g. training invitations), and all higher-impact actions require human review and explicit confirmation. Implement "hard stops" or business rules forbidding certain actions without additional inputs	Create composite views showing: risk score + exposure metrics + training history + contextual notes. Avoid "top N worst employees" lists, instead focus on risk profiles at team/role level. Provide space for human-entered notes or context when reviewing cases.
How can be assessed whether this measure has been fulfilled?	Review workflow configuration and business rules. Check that no API or integration allows direct, automatic disciplinary actions from scores. Sample decisions to confirm there is always a documented human step for high-impact use.	Review dashboards: are they multi-dimensional, or just lists of scores? Interview users about which information they actually rely on. Check report templates used in HR/security processes.
What are (potential) challenges to fulfilment?	Desire for automation and efficiency, even that the model's predictions are used as the final decision, without prior human supervision. Pressure from management to "act quickly" on high-risk scores. Legacy integrations that bypass new safeguards	Desire for simple rankings. Limited data integration capabilities. Time constraints in decision-making.
What are risks if not fulfilled?	Scores become de facto automated decision-making about employees. High risk of unfair treatment, discrimination claims and conflicts with GDPR Art. 22 and labour law.	Single-score rankings drive labelling and over-simplistic actions. Structural or contextual factors (workload, exposure, tooling) are ignored. Higher risk of blaming individuals instead of addressing systemic issues.
Which are the core function/role/ stakeholders responsible?	IT / Workflow and integration owners. Security / CISO office. HR. DPO / Legal / Compliance (oversight).	BI / Analytics teams. Security / HR process owners. Product owner. DPO / Compliance
Specific requirements?	EU AI Act: explicit focus on human oversight and avoidance of automation bias/over-reliance in high-risk AI; GDPR: provisions on automated decision-making and profiling (Art. 22) and requirement for meaningful human involvement where relevant.	EU AI Act: human oversight must be able to properly contextualise outputs; GDPR: fairness and proportionality, especially in the employment context.

Table 70: UC3 – Practical measures to avoid dependence with respect to Over-reliance

Organisational measures to address dependence		
Describe the measure	Policy on the role of the system in decision-making	Mandatory human review for high-impact uses. For example, a disciplinary action based on a "high-risk" score.
Why is it relevant?	Gives a clear signal to managers and security staff that over-reliance is not acceptable and defines boundaries for use.	Ensures decisions affecting employees' rights and working conditions are based on a holistic assessment, not just one metric. Reduces risk of unfair, arbitrary or discriminatory outcomes when the model is wrong, incomplete or biased.
How can it be achieved?	Draft and approve a policy stating: permitted uses, prohibited uses, and required human review. Link it to HR policies, security procedures, and ethics/AI governance frameworks. Communicate it to all relevant roles.	Define "high-impact" decisions in policy Require that, for such decisions: A named human role must review the case; they consult other information sources; they record a brief justification in the HR/security system (e.g. decision form or case file).
How can be assessed whether this measure has been fulfilled?	Existence and approval of the policy. Evidence that procedures (HR, security) reference it. Sample of decisions and escalations to see whether policy is respected.	Using a written rule (policy/procedure) clearly requiring human review for defined high-impact decisions.
What are (potential) challenges to fulfilment?	Disagreement between departments (security vs HR). Ambiguity in borderline use cases. L ow visibility of the policy in daily practice.	Operational pressure to automate: teams may want quick, automatic responses to "high risk" scores. Ambiguity about what counts as "high-impact" in practice, leading to inconsistent application. Limited time and resources for managers/HR to conduct proper reviews. Cultural bias toward trusting "objective" scores more than human judgment.
What are risks if not fulfilled?	Scores gradually used as hidden performance metrics. Over-reliance becomes normalised. Increased legal and reputational risk.	Vulnerability scores become de facto automated decisions about employees. Increased risk of unfair or disproportionate measures, particularly for groups where the model is less accurate.
Which are the core function/role/ stakeholders responsible?	HR. Security office Legal. DPO. Senior management.	HR. Security office Legal. DPO. Senior management.

	Works Council.	Works Council.
Specific requirements?	GDPR: purpose limitation, fairness, employee monitoring guidance. EU AI Act: requirement that deployers use high-risk systems as intended and with adequate human oversight.	GDPR: purpose limitation, fairness, employee monitoring guidance. EU AI Act: requirement that deployers use high-risk systems as intended and with adequate human oversight. Labour Law

Table 71: UC3 – Practical measures to achieve contestability and human oversight with respect to Over-reliance

Technical measures to address contestability and human oversight		
Describe the measure	Traceable logs for human reviews and overrides	Individual case view with explanations
Why is it relevant?	Shows that human oversight is actually happening, not just declared. Provides evidence for audits, investigations and learning from past cases. Enables detection of patterns (e.g. systematic overrides for certain groups).	Contestability requires that people know what the system is saying about them and why. Enables meaningful review by HR/managers and by the employee. Reduces "black box" perception and supports fairness under GDPR and AI Act.
How can it be achieved?	Extend logging to capture: who reviewed the case, what decision was made (confirm, adjust, disregard), brief justification. Link these logs to the underlying score/version and the challenge (if any). Ensure logs are access-controlled and retained per policy.	Implement an explanation API or layer that aggregates relevant features for each score. Design a case screen summarising: score, key drivers, time period, data categories. Use short, non-technical explanation
How can be assessed whether this measure has been fulfilled?	Sample contested cases: do they show a clear trace of review and outcome? Use analytics to see frequency and distribution of overrides.	Check that individual case views exist for employees/HR. Verify that explanations are consistent with the model logic and up to date.
What are (potential) challenges to fulfilment?	Additional complexity and performance overhead in logging. Privacy concerns if too much detail is logged. Ensuring consistent use by all reviewers.	Technical difficulty generating explanations that are both accurate and understandable. Risk of exposing too much internal detail or confusing users.
What are risks if not fulfilled?	Impossible to demonstrate that human oversight and contestation work in practice. Hard to detect biased override patterns.	Employees and HR cannot understand or challenge scores in any concrete way. Higher perception of arbitrariness; more resistance and distrust.
Which are the core function/role/ stakeholders responsible?	Platform / Data Engineering; Security. HR / Decision-makers (users of review tools). DPO / Compliance / Internal Audit	Data Science / ML Engineering (explanation logic). UX / Product (case view design). HR / Security (define which factors to show). DPO / Legal (review of content).

Specific requirements?	EU AI Act: logging obligations and human oversight. GDPR: accountability and security of processing; data minimisation in logs.	GDPR: transparency (Arts. 12–14), rights of access and information, profiling context. EU AI Act: information to deployers and affected persons about functioning and limitations; human oversight able to understand outputs.

Table 72: UC3 – Practical measures to achieve contestability and human oversight with respect to Over-reliance

Organisational measures to address contestability and human oversight		
Describe the measure	Oversight panel for contested and high-risk cases	Feedback loop from contestation outcomes into system design
Why is it relevant?	Brings multiple perspectives to difficult cases, reducing risk of biased judgments. Provides a focal point for analysing systemic issues revealed by individual contestations. Strengthens perceived legitimacy and fairness of outcomes.	Contestability is not only about correcting individual cases, but about detecting and fixing systemic problems (bias, mis-specification, confusing explanations). Shows regulators and employees that contestation leads to real improvements.
How can it be achieved?	Define membership, mandate and meeting frequency of the panel. Establish criteria for escalation (e.g. repeated contestations in same area, potential rights impact). Ensure decisions and recommendations are recorded and linked to system improvements.	Require that each resolved contestation is categorised (e.g. data error, misunderstanding, edge case, model limitation). Periodically analyse categories and frequency to identify patterns.
How can be assessed whether this measure has been fulfilled?	Terms of reference or charter for the panel. Meeting minutes and case summaries. Evidence that panel recommendations lead to policy, training or technical changes.	Existence of categorisation schema and reports. Evidence of changes made in response to patterns (release notes, updated model cards, revised policies).
What are (potential) challenges to fulfilment?	Time and capacity of senior stakeholders. Possible tension between security needs and worker rights. Ensuring confidentiality while including worker representation.	Fragmented recording of contestation outcomes. Lack of clear owner to translate insights into design changes.
What are risks if not fulfilled?	Difficult or systemic cases handled in an ad hoc, inconsistent manner. Opportunities to correct structural issues are missed. Greater perception of bias or defensiveness from management.	Same problems recur, generating repeated contestations and user frustration. Missed opportunity to improve fairness, accuracy and usability.
Which are the core function/role/	HR. Security / CISO office. DPO / Legal.	Product owner. Data Science / Engineering. HR.

stakeholders responsible?	Worker representatives. AI governance / Ethics department.	DPO / Compliance. AI governance department.
Specific requirements?	EU AI Act: governance and risk management for high-risk AI, including human oversight. GDPR: accountability and demonstrable safeguards for profiling and automated processing. Labour law relying on joint bodies/committees for oversight of monitoring tools in some jurisdictions.	EU AI Act: post-market monitoring and risk management cycle. GDPR: data protection by design and by default, requiring continuous improvement.

PRACTICAL MEASURES PROVIDED BY USE CASE 4

Table 73: UC4 – Practical measures to achieve Freedom of Expression and Non-Censorship with respect to Autonomy and agency

Technical measures to address Autonomy and agency		Organisational Measures
Describe the measure	Building Alternative narratives	'Human in the loop at every stage'
Why is it relevant?	The alternative narrative focuses on building agency and autonomy in groups/individuals that violent extremists seek to infringe upon	Ensuring that each alternative narrative is reviewed by a human ensuring the promotion of autonomy and agency
How can it be achieved?	Through focusing on the target (individual being radicalised) as opposed to target audience of hate	
How can be assessed whether this measure has been fulfilled?	Feedback loop built into evaluative process	Feedback forms
What are (potential) challenges to fulfilment?	Incidental worsening of narratives, lack of cultural sensitivity	Lengthy time to create
What are risks if not fulfilled?	Worsening of cultural divisions	Worsening of cultural divisions
Which are the core function/role/ stakeholders responsible?	Narrative developers based in target country	Narrative developers based in country
Specific requirements?	Laws that exist within that country	Provision of services within country

Table 74: UC4 – Practical measures to achieve Freedom of Expression and Non-Censorship with respect to Proportionality

Technical measures to address Proportionality		**Organisational Measures**
Describe the measure	Building Alternative narratives	Cultural sensitivity check
Why is it relevant?	Not intrusive and not punitive measure of addressing violent extremism	To ensure that the narrative being created is appropriate for the system it's being disseminated into
How can it be achieved?	Utilising designated trusted messengers within each country	Adding in an extra external check
How can be assessed whether this measure has been fulfilled?	Feedback loop built into evaluative process	Identity Mapping happening as part of the system
What are (potential) challenges to fulfilment?	Not having a full assessment of spread of hate narrative and underestimating the spread - therefore not creating appropriate narrative	Steps being missed due to time it takes to create an alternative narrative
What are risks if not fulfilled?	Worsening of cultural divisions	Worsening of cultural divisions
Which are the core function/role/ stakeholders responsible?	Narrative developers based in target country	Narrative developers based in country
Specific requirements?	The needs/laws and cultural codes within each country	Provision of services within country

Table 75: UC4 – Practical measures to achieve Freedom of Expression and Non-Censorship with respect to Non-discrimination

Technical measures to address Non-discrimination		Organisational Measures
Describe the measure	Building Alternative narratives	Fairness & Non-Discrimination Governance Framework
Why is it relevant?	The entire focus of the alternative narrative is not to focus on "wrong" or and "out group" and instead focuses on alternative viewpoints for the target	Ensures systematic prevention, detection, and remediation of discriminatory impacts in hate speech detection, beyond ad-hoc technical fixes
How can it be achieved?	By ensuring that the language used is non-discriminatory, avoiding blaming, villainising or shaming language	Establishing a formal governance structure with clear ownership. Appointment of an AI Fairness/Ethics Lead, creation of a cross-functional review body (ML, Legal, Policy, DEI, Trust & Safety) and integration of non-discrimination checks into the system lifecycle
How can be assessed whether this measure has been fulfilled?	Feedback loop built into evaluative process	Existence of documented policies and roles, regular review meetings, evidence of decisions and remediation actions as well as inclusion of fairness KPIs in governance reports
What are (potential) challenges to fulfilment?	The feelings of the narrative builder interfering with the narrative/take down measures being utilised by traditional organisations	Ambiguous accountability across teams; limited organisational expertise on bias; potential conflicts with business or moderation objectives
What are risks if not fulfilled?	Worsening of cultural divisions	Persistent discriminatory outcomes, unequal treatment of protected groups, legal and regulatory non-compliance. Organisation could also suffer reputational harm
Which are the core function/role/ stakeholders responsible?	Narrative developers based in target country	Narrative developers based in target country
Specific requirements?	Laws/regulations/cultural divides that exist within that country	Laws/regulations/cultural divides that exist within that country

Table 76: UC4 – Practical measures to achieve Non-bias, fairness and non-discrimination with respect to Equality and impartiality

Technical measures to address Equality and Impartiality			Organisational measures to address Equality and Impartiality
Describe the measure	Cultural sensitivity check	Ethnographic insight	Formal Policy for AI Systems
Why is it relevant?	To ensure that the developers preconceived and bias notions aren't spilling into the alternative narrative	The deep, contextual understanding of people's lived experiences, ensuring embedding into the everyday environments	Ensures consistent and unbiased treatment of individuals and groups affected by hate-speech detection
How can it be achieved?	Team meetings and external checks	Specific choosing of dissemination stakeholders	Adopt an organisation-wide policy defining equality principles, prohibited biases, and impartial decision-making standards; embed into AI governance and moderation policies
How can be assessed whether this measure has been fulfilled?	Identity mapping	Will naturally feed into narrative	Policy existence and accessibility and integration into operational procedures
What are (potential) challenges to fulfilment?	Cultural nuances being missed or misunderstood	Narrative being based on incorrect cultural nuances and incorrect disinformation	Translating abstract principles into practice and a resistance to change from the organisation and/or end users
What are risks if not fulfilled?	Worsening of cultural divides	Worsening of cultural divides	Unequal enforcement, discriminatory outcomes and regulatory and reputational risks
Which are the core function/role/ stakeholders responsible?	Narrative developers	Narrative developers	Narrative developers
Specific requirements?	The needs/laws and cultural sensitivity within each country	The needs/laws and cultural sensitivity within each country	The needs/laws and cultural sensitivity within each country + mandatory staff training / enforcement and escalation mechanisms

Table 77: UC4 – Practical measures to achieve Non-bias, fairness and non-discrimination with respect to Inclusivity

Technical measures to address Inclusivity	
Describe the measure	Identity Mapping
Why is it relevant?	To ensure the targets of the alternative narrative are appropriately picked
How can it be achieved?	AI Tools - that could not be disclosed due to data security
How can be assessed whether this measure has been fulfilled?	Narratives being disseminated and engaged with correctly
What are (potential) challenges to fulfilment?	Not having all the correct information, data security not allowing full access to that target group
What are risks if not fulfilled?	Worsening of cultural divides
Which are the core function/role/ stakeholders responsible?	Narrative developers
Specific requirements?	The needs/laws and cultural sensitivity within each country

Table 78: UC4 – Practical measures to achieve Inclusivity

Organisational measures to address Inclusivity	
Describe the measure	Inclusive Stakeholder Engagement and Review Process
Why is it relevant?	Ensures that perspectives of affected and marginalised communities inform system design and policy decisions
How can it be achieved?	Establish structured consultation with civil society, advocacy groups, linguists, and impacted users. Include feedback loops into policy and model updates
How can be assessed whether this measure has been fulfilled?	Documentation of consultations, diversity of participants and evidence of changes made based on feedback
What are (potential) challenges to fulfilment?	Identifying representative stakeholders and balancing conflicting views
What are risks if not fulfilled?	Blind spots in harm identification and exclusion of vulnerable voices
Which are the core function/role/ stakeholders responsible?	Narrative developers
Specific requirements?	Defined consultation cadence, participant selection criteria and feedback incorporation process

Table 79: UC4 – Practical measures to achieve Transparency

Technical measures to address Transparency		Organisational measures to address Transparency
Describe the measure	Model training on publicly available datasets	Publicly Documented Content Moderation and Classification Standards
Why is it relevant?	Users know exactly what data has gone into the model and what it has trained on, giving an insight into the decision making inside the model	Allows users and auditors to understand how hate speech is defined and enforced, supporting fairness and accountability
How can it be achieved?	Use public datasets available under license from sites such as HuggingFace; or if you are using a dataset created by you, publish your dataset	Publish clear moderation guidelines explaining criteria, examples, thresholds, and enforcement actions; maintain version history
How can be assessed whether this measure has been fulfilled?	Availability and visibility of datasets	Availability and clarity of documentation and user comprehension testing
What are (potential) challenges to fulfilment?	Opening the model to edge cases, manipulation of inputs and outputs, hosting requirements of self-generated datasets	Oversimplification and a risk of gaming the system
What are risks if not fulfilled?	Perceived arbitrariness; loss of user trust; difficulty contesting decisions	Perceived arbitrariness; loss of user trust; difficulty contesting decisions
Which are the core function/role/ stakeholders responsible?	Developers	Policy team; Legal; Communications; Trust & Safety
Specific requirements?	Hosting availability for datasets	Plain-language documentation; change logs; alignment with legal standards

Table 80: UC4 – Practical measures to achieve Accountability and responsibility

Technical measures to address Human Oversight		Organisational measure to address Auditability/Evaluation	Organisational measure to address Responsiveness
Describe the measure	Human checks at each stage of building narrative	Evaluative framework built into process	Dissemination of messengers
Why is it relevant?	Ensuring that the narratives are not entirely reliant on AI LLMS that could be trained incorrectly	The very function of the entire process means that it can be evaluated at every stage and then the identity mapping will intrinsically show whether the process is effective or not	Having the correct stakeholders disseminate the narrative is vital to the success of the narrative, as if it disseminated by the appropriate source it is much more likely to be accepted
How can it be achieved?	Ensuring security checks implemented after each output	Practitioners and narrative developers engaging in evaluation checkpoints	Cultural engagement and understanding of the systems within the target audience
How can be assessed whether this measure has been fulfilled?	Evaluation forms	Through the continued identity mapping	Identity mapping
What are (potential) challenges to fulfilment?	Time poor practitioners not realising the importance	Time poor practitioners not realising the importance	Time poor practitioners not realising the importance
What are risks if not fulfilled?	Worsening of cultural divides	Worsening of cultural divides	Worsening of cultural divides
Which are the core function/role/ stakeholders responsible?	Narrative developers	Narrative developers	Narrative developers
Specific requirements?	The needs/laws and cultural sensitivity within each country	The needs/laws and cultural sensitivity within each country	The needs/laws and cultural sensitivity within each country

PRACTICAL MEASURES PROVIDED BY USE CASE 5

Table 81: UC5 – Practical measures to achieve Privacy and data protection with respect to User consent and transparency

Technical measures to address User consent and transparency		
Describe the measure	Data encryption: Process of scrambling data into an unreadable format to protect it from unauthorized access	Anonymization techniques
Why is it relevant?	To protect stored conversation data and user information	Privacy expectations of users are crucial for customer retention
How can it be achieved?	Symmetric (shared key algorithm) and asymmetric (public key algorithm) encryption	De-identification of personal data for model training where possible
How can be assessed whether this measure has been fulfilled?	Internally or externally?	
What are (potential) challenges to fulfilment?	Different regulations/policies regarding data encryption dependent on region;	Complete removal of identifying elements is in conflict with monitoring system; possibility of re-identification is a risk for data privacy
What are risks if not fulfilled?	Data breaches and leaks to unauthorized parties; legal consequences; reputational damage/loss of user trust	Identity exposure, non-compliance with data protection laws, economic damage
Which are the core function/ role/stakeholders responsible?	Data collector; technical team	Data collector; technical team
Specific requirements?	GDPR in Europe; payment card industry data security standard (PCI DSS)	GDPR in Europe; payment card industry data security standard (PCI DSS)

Table 82: UC5 – Practical measures to achieve Privacy and data protection with respect to User consent and transparency

Organisational measures to address User consent and transparency			
Describe the measure	GDPR compliance program	Privacy policy transparency	Privacy impact assessment
Why is it relevant?	Mandatory for organisations processing personal data of EU residents	Mandatory for organisations processing personal data of EU residents	Ensures compliance with terms of use and users' privacy expectations
How can it be achieved?	Establishing a comprehensive compliance program that includes data protection policies, procedures, staff training, and regular audits	drafting clear, concise, and easy to access privacy policies on data collection, use, and sharing practices that are handed to users beforehand	regular evaluation of privacy risks in monitoring systems
How can be assessed whether this measure has been fulfilled?	internal audits, third-party evaluators to check for compliance with EU data protection requirements	reviewing and gathering user feedback	
What are (potential) challenges to fulfilment?	managing cross-border data flows, keeping up with latest regulations	Being as transparent as possible while handling the risk that users feel their privacy violated; extensive monitoring is required to fulfil safety requirements, but users do not need to know about any detail of their data processing	
What are risks if not fulfilled?	legal action, reputational and thus economical damage, loss of users (trust)	legal penalties, user mistrust and privacy complaints	
Which are the core function/role/ stakeholders responsible?	technical, legal, and compliance team	legal, and compliance team	technical, legal, and compliance team
Specific requirements?	Appointment of a Data protection officer	Compliance with EU AIA & GDPR requirements for transparency	

Table 83: UC5 – Practical measures to achieve Privacy and data protection with respect to Data minimisation, data use and storage

Technical measures for Data minimisation, data use and storage		Organisational measures
Describe the measure	Collection of only necessary data for functionality and safety	User data rights processes
Why is it relevant?	Reduces risk of data breaches, compliance with regulations	Mandatory for organisations processing personal data of EU residents
How can it be achieved?	clearly define data collection purposes, data retention policies	Implement transparent data policies and consent mechanisms; give users right to access, delete, and transfer their data
How can be assessed whether this measure has been fulfilled?	internally and by third-party evaluators	internal audits, user feedback, third-party evaluators
What are (potential) challenges to fulfilment?	Identifying what data is truly necessary and how long it has to be stored to fulfil safety obligations	managing cross-border data flows, keeping up with latest regulations, technical boundaries?
What are risks if not fulfilled?	increased vulnerability in case of data leaks, loss of user trust and reputation (economic damage)	legal penalties, loss of user trust
Which are the core function/role/ stakeholders responsible?	Data collector; technical team	technical, legal, and compliance team
Specific requirements?	GDPR in Europe; payment card industry data security standard (PCI DSS)	Compliance with GDPR

Table 84: UC5 – Practical measures to achieve Privacy and data protection with respect to Third party sharing and compliance

Technical measures to address Third party sharing and compliance			**Organisational measures**
Describe the measure	Audit logging	Access controls	Data sharing agreements
Why is it relevant?	Provides a trail of user and system activities, essential for detecting and responding to security incidents	Prevent unauthorized employees to access user data; ensure sensitive data is only accessed for valid reasons	Secure and legal data exchange necessary to protect user information
How can it be achieved?	Automatically record activities within the IT infrastructure, secure storage of logs	Clearly defined roles and role-based and context-based access control	Drafting contracts and agreements detailing data usage, protection measures and consequences if non-compliance
How can be assessed whether this measure has been fulfilled?		Access rights and documentation of accesses	
What are (potential) challenges to fulfilment?	Management of vast amounts of logging data, compliance with logging standards and regulatory requirements	ineffective role management; balancing security and usability -> effective moderation requires extensive access to conversation history	differing legal requirements across jurisdictions, managing multiple partners and their differing handling
What are risks if not fulfilled?	Inability to respond to security incidents, lack of accountability, non-compliance with requirements	Data breaches and leaks to unauthorized parties; legal consequences; reputational damage/loss of user trust	data breach, legal consequences
Which are the core function/role/ stakeholders responsible?	Data collector; technical team	Data collector; technical team	technical, legal, compliance team, business partners, payment platforms
Specific requirements?	GDPR in Europe; payment card industry data security standard (PCI DSS)	GDPR in Europe; payment card industry data security standard (PCI DSS)	compliance with GDPR, PCI DSS for payment data

Table 85: UC5 – Practical measures to achieve Safety/Human safety with respect to User protection

Technical measures to address User protection		**Organisational measures**	
Describe the measure	Tiered severity system	Cross-functional tier definition	Regular policy review
Why is it relevant?	Treat risks according to their severity	positive user experience depends on comprehensible and differentiated rules for violation classification;	Keep up with latest regulatory specifications and legal requirements
How can it be achieved?	implement, inspect, and update periodically	Safety categories developed collaboratively by manual moderators, data team, legal team, and payment platform representatives	Ongoing assessment of tier definitions against emerging patterns
How can be assessed whether this measure has been fulfilled?			
What are (potential) challenges to fulfilment?	interdependent with classification models	New risks and potential harmful behaviour arise time and again; monitoring of latest research findings especially on psychological risks and harms	No legal obligation
What are risks if not fulfilled?	Over-flagging and disruption of user satisfaction		Legal issues
Which are the core function/role/ stakeholders responsible?	Technical team and human moderators	Whole team	Legal team, leader
Specific requirements?	Escalation mechanism		

Table 86: UC5 – Practical measures to achieve Safety/Human safety with respect to Security measures

Technical measures to address Security measures				
Describe the measure	Multi-tier classification models	real-time content filtering	pattern detection algorithms	jailbreak detection
Why is it relevant?	Violations and potential threats to users vary in severity; maintain security and autonomy at the same time	respond in real-time to potential violations in conversation history	respond preventively to problematic behavioural patterns, apply escalation process from monitoring to highlighting to warning to user ban	prevent users from circumventing safety system
How can it be achieved?	Develop several AI models trained on human-labelled datasets to automatically classify content into safety tiers at scale	automated detection and filtering of prohibited content during conversations	systems to identify concerning behavioural patterns over time	measures to identify and prevent attempts to circumvent safety systems
How can be assessed whether this measure has been fulfilled?	implementation and regular inspection	implementation and regular inspection	implementation and regular inspection	implementation and regular inspection
What are (potential) challenges to fulfilment?	misaligned models or severity system	leaks in the filter system	could violate data usage and storage agreement	rapidly advancing technology offers new possibilities, being up to date
What are risks if not fulfilled?	overlooking security risks, harm to users	harm to users; difficulty of recapturing once harmful content has been spread	underestimating the long-term threat of behaviour that was classified as mild or no violation	harmful content enters circulation; Accountability gaps and liability exposure in case of accidents or regulatory inspections.
Which are the core function/ role/ stakeholders responsible?	technical and legal team	technical and legal team	technical and legal team	technical and legal team
Specific requirements?				

Table 87: UC5 – Practical measures to achieve Safety/Human safety with respect to Security measures

Organisational measures to address Security measures			
Describe the measure	progressive intervention protocol	payment platform alignment	legal compliance review
Why is it relevant?	respond to violations in an appropriate but consistent and systematic manner	coordination important for accountability and clarification of responsibilities	Moderation decisions must be in accordance with applicable law
How can it be achieved?	implement and communicate an escalation process from monitoring, to highlighting, to warning to user ban	Policies coordinated with payment processor requirements	Regular legal team assessments of moderation decisions
How can be assessed whether this measure has been fulfilled?	drafting clear protocol	contracts with payment platform	regular audits with legal team
What are (potential) challenges to fulfilment?	communication with users about these rules, balance between safety and autonomy		
What are risks if not fulfilled?	inconsistent and unsystematic responses to incidents can annoy users and harm reputation	unclear accountability, liability	decisions that bypass law, liability
Which are the core function/ role/ stakeholders responsible?	human moderators, technical team	compliance team	legal team
Specific requirements?			

Table 88: UC5 – Practical measures to achieve Safety/Human safety with respect to Human oversight

Technical measures to address Human oversight		Organisational measures
Describe the measure	Labelled training data sets	Human moderation oversight
Why is it relevant?	further development of multi-tier moderation system based on human feedback	edge cases, cases that are not clearly classifiable or need context for decision; automated systems may over-flag content that human reviewers would consider acceptable and vice versa
How can it be achieved?	curate datasets of examples labelled by human moderators to train the automated systems	employ a team of human moderators; supervise and provide further training on a regular basis
How can be assessed whether this measure has been fulfilled?		Review that human validation occurred for every edge case that was not automatically detected
What are (potential) challenges to fulfilment?	clear classification of cases that require a great deal of context and individual case decisions	Training of moderators, exposing moderators to potentially traumatic content only possible for a limited period of time; liability for decisions
What are risks if not fulfilled?	dependence on human decision-making is more costly and prone to error; friction with user autonomy	false positives and negatives in moderation decisions
Which are the core function/role/ stakeholders responsible?	human moderators, technical team	human moderators
Specific requirements?		

Table 89: UC5 – Practical measures to achieve Autonomy/User agency with respect to Agency

Technical measures to address Informed Consent		
Describe the measure	User consent mechanisms	Age verification system
Why is it relevant?	Consent legally required	prevention that minors enter the platform
How can it be achieved?	retrieving consent from users after enlightenment and potentially later on	technical measures to ensure only adults access the platform
How can be assessed whether this measure has been fulfilled?	no account generation possible without consent	UNCLEAR + UNCLEAR IF SUFFICIENT; verification via identity document
What are (potential) challenges to fulfilment?	grey areas where consent must be sought during conversations	users circumvent age verification, excessive data collection for verification conflicts with privacy; adult users can still show content to minors
What are risks if not fulfilled?	harm to users, liability	harm to minors, legal consequences
Which are the core function/role/ stakeholders responsible?	legal team, human moderators	legal team, technical team
Specific requirements?	GDPR in Europe	laws protecting minors

Table 90: UC5 – Practical measures to achieve Autonomy/User agency with respect to System customisation

Technical measures to address System customisation			Organisational measures
Describe the measure	Customization parameters / scope-limiting guardrails	Competitive analysis	Internal policy on permissible behaviour (change)
Why is it relevant?	platform has duty to care for safety by defining boundaries against harmful, violent, or potentially traumatizing content and to keep the model within intended context	strategic orientation, don't lose competitive edge	Transparency and documentation about what the application will and will not support regarding personalisation and scope; creates shared understanding and commitment
How can it be achieved?	boundaries defining permitted (character) behaviour modifications; key word filtering, output monitoring	Regular monitoring of competitor policies to assess market positioning	Documentation, internal memo, regular evaluation and ethical discussions
How can be assessed whether this measure has been fulfilled?	evaluation of moderation-system and test model performance		Internal only
What are (potential) challenges to fulfilment?	Not every possible borderline case is predictable, depending on post hoc assessment and intervention	competitors not transparent about own strategies	Internal disagreement about how to deal with areas, unclear legal situation
What are risks if not fulfilled?	users being exposed to potentially harmful content	being outdone, lose customers	Uncertainty about it jeopardizes the other measures
Which are the core function/role/ stakeholders responsible?	technical team, human moderators	Whole team	Whole team; regulation authorities
Specific requirements?	EU AIA recital 5 + 9	no	EU AIA

Table 91: UC5 – Practical measures to achieve Safety/Human Safety with respect to Transparency and user understanding

Technical measures to address Transparency and User understanding		**Organisational measures**	
Describe the measure	Transparent moderation notifications	Clear community guidelines & user education resources	Appeals process
Why is it relevant?	Transparency and explainability, user satisfaction	Without minimum amount of transparency, no informed consent	Necessary to adopt to user needs while keeping them safe
How can it be achieved?	Automated explanations when content is restricted of flagged	By published documentation of permitted and prohibited content	Implement mechanisms for users to context moderation decisions
How can be assessed whether this measure has been fulfilled?	Regular monitoring	Regular monitoring	-
What are (potential) challenges to fulfilment?	Tagging disrupts user satisfaction and providers could therefore decide against it	Varying engagement with guidelines by users	In conflict with the autonomy and decision-making authority of the provider
What are risks if not fulfilled?	Ambiguity, uncertainty about permissible interactions/content	Ambiguity, uncertainty about permissible interactions/content	Limited user satisfaction
Which are the core function/role/ stakeholders responsible?	Human moderators, technical team	Whole team	Legal team, human moderators
Specific requirements?	No	No	no

Table 92: Practical measures to achieve Promotion of user's health

Technical measures to address Promotion of user's health		Organisational Measures	
Describe the measure	Automated classifiers for harmful advice prevention	Advisory consultation	Ethics review for new features
Why is it relevant?	Need for technical system that automatically flags dangerous topics/contexts	Involve external expertise from mental health professionals and ADHD specialists to inform appropriate boundaries and approaches	Internal team discussion for reviewing ethical implications before launch to ensure new capabilities align with ethical principles
How can it be achieved?	Implement automated classifiers that detect topics such as weight loss, medication-related requests and decline advice in such cases	Consult external experts on a regular basis	Establish fixed procedure before launch of new features
How can be assessed whether this measure has been fulfilled?	Regular monitoring	internal	internal
What are (potential) challenges to fulfilment?	Users trying to jailbreak the system; defining harmful advice in borderline cases	-	Overlook minor changes that do not constitute a new feature that have consequences on AI behaviour in other contexts
What are risks if not fulfilled?	Over flagging content or overlooking harm	Failure of having boundaries and area of application professionally secured	Non-compliance with ethical regulations
Which are the core function/role/ stakeholders responsible?	Technical team	Founder, whole team	Whole team
Specific requirements?	No	no	-

Table 93: Practical measures to achieve Crisis Detection

Technical measures to address Crisis Detection		Organisational measures to address Crisis Detection
Describe the measure	External LLM solutions for crisis detection	User feedback collection
Why is it relevant?	Having an external solution specifically trained for crisis detection	Enabling users to report concerns or problems provides insightful feedback from real world usage contexts
How can it be achieved?	Implement crisis detection AI for suicidal ideation and self-harm identification; escalate borderline situations to this system	Set up reporting system and customer team interface
How can be assessed whether this measure has been fulfilled?	Regular monitoring	-
What are (potential) challenges to fulfilment?	Tagging disrupts user satisfaction as application essentially operates in a critical area where the boundaries of harmful behaviour must be carefully balanced	Lack of capacity to process queries and feedback givens start up size
What are risks if not fulfilled?	Failure to refer users to professional help in crisis situations in a timely manner	Missing crucial feedback, customer attrition
Which are the core function/role/ stakeholders responsible?	Human moderators, technical team	Technical team, human moderators/support
Specific requirements?	No	No

Table 94: Practical measures to achieve Scope Boundaries

Technical measures to address Scope Boundaries		Organisational measures to address Scope Boundaries
Describe the measure	Multi-layered scope control	Explicit scope boundaries
Why is it relevant?	Catching out of context content both input (human) and output (AI) levels	Clear organisational policy on what the partner supports vs excludes
How can it be achieved?	Keyword filtering to catch out of scope requests; training data constraints limiting model knowledge to appropriate domains; output monitoring to prevent harmful responses even if filters missed request	Policy formed through internal team discussion, terms of conditions
How can be assessed whether this measure has been fulfilled?	Internal, monitoring	Internal, official application orientation of the app
What are (potential) challenges to fulfilment?	Constant updating of technical implementation	Competitive disadvantage compared to other providers can complicate decision on policy
What are risks if not fulfilled?	Unacceptable interactions with risks for users	Unacceptable interactions with risks for users
Which are the core function/role/ stakeholders responsible?	Technical team	Whole team
Specific requirements?	EU AIA Art. 14 (behaviour manipulation banned)	EU AIA Art. 14 (behaviour manipulation banned)

PRACTICAL MEASURES PROVIDED BY USE CASE 6

Table 95: UC6 – Practical measures to achieve Non-maleficence with respect to Subsidiarity and proportionality and Effectiveness

Organisational measure to address Subsidiarity and proportionality		Organisational measures to address Effectiveness	
Describe the measure	Consider less intrusive (exposure) therapies first	Personalization of clinical protocol; one size may not fit all	Use only in the research context, currently
Why is it relevant?			Whether deepfake therapy has benefits over other forms of therapy requires further research. As long as it is unclear whether deepfake therapy meets principles for good care and the criteria of proportionality and subsidiarity, such therapy should be proposed only to patients in a clinical research context.
How can it be achieved?			
How can be assessed whether this measure has been fulfilled?			
What are (potential) challenges to fulfilment?			
What are risks if not fulfilled?			
Which are the core function/role/ stakeholders responsible?			
Specific requirements?			

Table 96: UC6 – Practical measures to achieve Non-maleficence with respect to Societal well-being

Organisational measures to address Societal well-being			
Describe the measure	Avoid introduction in chatbots that can be used without a medical purpose and without the presence of a therapist.	Deepfake therapy should be societally accepted before implementing it into practice, and this may require public dialogue	Broader fairness concerns such as equal access to deepfake therapy and the environmental impact, should be considered before deciding whether deepfakes should be introduced in mental healthcare.
Why is it relevant?			
How can it be achieved?			
How can be assessed whether this measure has been fulfilled?			
What are (potential) challenges to fulfilment?			
What are risks if not fulfilled?			
Which are the core function/role/ stakeholders responsible?			
Specific requirements?			

Table 97: UC5 – Practical measures to achieve Autonomy with respect to Transparency and Privacy

Organisational measure to Transparency		**Organisational measures to address Privacy**	
Describe the measure	Consent: information about the therapy to the person undergoing it (expectation management)	Consent from the depicted (conclusion: not desirable in PTSD case, may be an option in grief case when someone is terminally ill)	Family discussion/consent (grief therapy case)
Why is it relevant?			Potentially, the person undergoing deepfake therapy should discuss this with other living family members too, who might take offence to deepfake therapy due to worries about the distortion of their loved one's image or the instrumentalization itself. However, unclear if the family has a legitimate interest that would outweigh the interest in therapy (further study needed)
How can it be achieved?			
How can be assessed whether this measure has been fulfilled?			
What are (potential) challenges to fulfilment?		We discussed that consent is not desirable (and also not done for regular exposure therapy). If deepfake therapy can be beneficial for sexual-violence-related PTSD victims, as a last resort for a considerable health problem, we find that using a deepfake without consent would be acceptable in light of the legitimate interest of the patient.	
What are risks if not fulfilled?			
Which are the core function/role/ stakeholders responsible?		GDPR; national law. This consideration is also in line with a survey among the general public which finds that deepfake (voice) transformations are more accepted when they are used inside a therapeutic context doi:10.1098/rstb.2021.0083	
Specific requirements?			

Table 98: UC5 – Practical measures to achieve Safety/Human Safety with respect to Privacy

Technical measures to address Privacy					
Describe the measure	Minimal number of people with access	Encrypted transfer of the picture used to create the deepfake, and immediate deletion after use	From Hoek et al: given that therapists are unlikely to be directly involved in the deepfake creation process, it is imperative to carefully vet and select a third party with expertise in medical or therapeutic applications, and a comprehensive understanding of privacy concerns and cybersecurity.	From Hoek et al: therapists should ensure that in cases of a data breach it remains clear that the video is fake, for instance, by always using their own voice rather than a deepfaked voice, or by incorporating resilient watermarks in the deepfake	From Hoek et al: As deepfakes are a digital medium, the patient will interact with them through videocall software, which should be secure and therapy-specific, rather than relying on general-use commercial companies such as Zoom or Teams, which can be less transparent about their data storage, monitoring and use.7 Standard contractual clauses would be helpful for drafting agreements with technology companies creating deepfakes and proving the videocall software.
Why is it relevant?					
How can it be achieved?	Only 2 therapists see it				
How can be assessed whether this measure has been fulfilled?					
What are (potential) challenges to fulfilment?					
What are risks if not fulfilled?					
Which are the core function/role/ stakeholders responsible?					
Specific requirements?					

Table 99: UC5 – Practical measures to achieve Autonomy with respect to Risk of over-attachment and dependency

Technical measure to address Risk of over-attachment	**Technical measure to address Risk of over-attachment**				
Describe the measure	Limit the number of sessions	Avoid introducing false or wish-fulfilling narratives (e.g., imagined reconciliation or unrealistic "afterlife" dialogues).	Have a preparation meeting to manage expectation; ensure people know it is fake	Tailor therapy to the patient; do not use in patients who are e.g. prone to overattachment	qualitative research with patients and therapists should be conducted to explore the risk of overattachment and potential other unintended consequences for persons receiving deepfake therapy.
Why is it relevant?					
How can it be achieved?					
How can be assessed whether this measure has been fulfilled?					
What are (potential) challenges to fulfilment?					
What are risks if not fulfilled?					
Which are the core function/role/ stakeholders responsible?					
Specific requirements?					

Table 100: UC5 – Practical measures to achieve Accountability with respect to Human agency and responsibility and Professional competence

Organisational measures to address Human agency and responsibility			**Organisational measures to address Professional competence**
Describe the measure	Develop guidelines that highlight that the AI technology is secondary and the responsibility remains with the therapist	To respect autonomy, healthcare providers should thoroughly discuss the pros and cons of deepfake therapy with their patients and provide them the opportunity to refrain from it, similar to any treatment option.	Prevent re-traumatization by tailoring therapeutic scripts carefully and ensuring therapists are skilled in trauma dynamics.
Why is it relevant?			
How can it be achieved?			Training
How can be assessed whether this measure has been fulfilled?			
What are (potential) challenges to fulfilment?			
What are risks if not fulfilled?			
Which are the core function/role/ stakeholders responsible?			
Specific requirements?			

Table 101: UC5 – Practical measures to achieve Autonomy with respect to Oversight

Organisational measures to address Oversight					
Describe the measure	Independent ethical review (e.g., REC)	Involvement of ethicists/lawyers in design of deepfake therapy.	Continuous monitoring of patient experiences and adaptation of protocols.	Legal study on whether and how the MDR and AI Act would apply	To ascertain whether deepfake therapy can constitute good care, more ethical, legal, social and psychological research is needed into its effects, merits and drawbacks
Why is it relevant?			Evaluations after sessions, but also conducting interviews with experts and experience expert about acceptability and effectiveness		
How can it be achieved?					
How can be assessed whether this measure has been fulfilled?					
What are (potential) challenges to fulfilment?			People might have very different views and experiences		
What are risks if not fulfilled?					
Which are the core function/role/ stakeholders responsible?					
Specific requirements?					