Funded by the European Union

3

D2.2 Report on the selection of ethical principles and values

Deliverable under WP2: Understanding emerging trends and ethical challenges

Project Name	AIOLIA
Deliverable Title/Number	D2.2
Description	Report on the selection of ethical principles and values
Lead beneficiary	CEPS
Lead Authors	Sue Anne Teo; Nicoleta Kyosovska
Contractual delivery date:	30/06/2025
Actual delivery date:	30/06/2025
Sensitivity	PUBLIC

Document History

Name	Organisation	Role	Action	Date
Sue Anne Teo	CEPS	Task 2.2 lead; author	Conceptualisation, draft, literature review, bibliographic research and complete draft	March-June 2025
Nicoleta Kyosovska	CEPS	Co-author	Conceptualisation, bibliographic research, data analysis, literature review and draft	March-June 2025
Abril Armengol	CEPS	Co-author	Draft and data analysis	May-June 2025
Andrea Renda	CEPS	Internal reviewer	Internal review	3 June 2025
Miltos Ladikas	KIT	Consortium Reviewer	Review	18 June 2025
Alexei Grinbaum	CEA	Coordinator	Review	19 June 2025
Sue Anne Teo	CEPS	Lead author	Revision upon review	24 June 2025
Elena Lazutkaite, Atsuo Kishimoto	EUREC; The University of Osaka	Consortium partners	General consortium review and feedback	25 June 2025
Sue Anne Teo	CEPS	Task 2.2 lead	Submission of D2.2	30 June 2025

Dissemination level	
PU	Public, fully open

How to cite
Teo, S.A., Kyosovska, N. and Armengol, A. (June 2025). AIOLIA D2.2: Report on the selection of ethical principles and values.

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.

SUMMARY

AIOLIA deliverable 2.2 contains a selection of ethical principles and values in relation to the use cases and research areas in the project (see section "Selection tables"). The selection of the most relevant principles is in relation to use cases from EU partners. International partners were tasked to select the ethical principles in relation to their own use cases, drawing upon either domestic instruments touching upon the ethics of AI, both soft law and legislation or by drawing upon international instruments. In turn, the selection of ethical principles in relation to the research areas was conducted through a synthesis of the main ethical concerns identified in relation to how those research areas were used in the specific use cases. The deliverable will inform the values that will form part of the co-creation process to operationalise AI ethics guidelines in work package 3.

The deliverable begins with a literature review on the main ethical principles and values in relation to AI ethics, drawing from the rich literature on this space. The report then elaborates on the use cases and research areas. The review drew upon legislative instruments, ethics documents and research articles on AI ethics in general as well as in relation to ethical issues identified within specific use cases and research areas. The selection of the most relevant ethical principles and values in relation to EU partners' use cases is then presented, followed by the most relevant ethical principles and values on the specific research areas. The selection was also complemented and informed by the findings in two surveys – one sent to AIOLIA partners (internal survey) and one sent to networks and the AIOLIA Scientific Advisory Board (external survey). The last part elaborates on the ethical principles and values chosen by the international partners in relation to their use cases.

CONTENTS

1. Introduction

a. AIOLIA background and overview

Artificial intelligence (AI) is increasingly diffused and deployed in society, in private, professional and public domains alike. Alongside this, ethical principles and values in relation to AI are also proliferating around the world as an attempt to guide the proper design, testing and deployment of AI systems. Certain ethical principles and values are also being concretised through legislation in different parts of the world, such as within the European Union, China, India and South Korea. At the same time, even with increased proliferation, operationalisation of these ethical principles and values remain challenging – as companies, public authorities and individuals designing and deploying AI fail to adequately understand or translate the principles into practice.

AIOLIA brings together both European and international groups of AI practitioners working on AI systems that impact human cognition and behaviour to engage in a co-creation process to operationalise AI ethics in practice. Academic partners who research on different aspects of AI ethics are brought together with industrial partners who deploy AI technologies, in this co-creation process, ensuring that the theory and practice of AI ethics move in lock-step. This process involves the co-creation of AI ethics guidelines in context, moving bottom-up from specific use cases, defined as the use of AI within specific domains, such as AI for recruitment, to selected AI research areas – defined here as AI sub-fields, such as decision-support systems or image recognition (see section 2.c).

As AIOLIA focuses on the impacts of AI on human cognition and behaviour, the use cases selected will reflect on the challenges that AI introduces in private and professional behaviours. The ethical principles that are selected will help to navigate these challenges.

CEPS is tasked under task 2.2 to select the most relevant ethical principles in relation to the use cases chosen by the European industrial and academic partners. In addition, international partners in AIOLIA are tasked to select their own use cases, in consultation with domestic industrial partners, and the most relevant ethical principles in relation to each. The use cases chosen by European partners for AIOLIA are:

i. Medical doctors (radiologists, surgeons) using AI tools in diagnostics and treatment
ii. Safety engineers using AI tools to speed up software release approvals
iii. Recruiters using AI tools in hiring processes
iv. Security professionals using AI tools to detect hate speech
v. AI systems as personal and family virtual assistants
vi. Deepfake AI-based psychotherapy for processing trauma and grief

The use cases chosen by the international partners for AIOLIA are as follows:

vii. Workplaces equipped with AI tools for behavioural analysis
viii. AI systems for smart elderly care
ix. AI systems as personal companions assisting senior citizens
x. Generative Ghosts and the Grieving Process

In addition, AIOLIA will also both examine and operationalise ethical principles and values relating to research areas. While there are many different AI sub-fields, the research areas in focus for AIOLIA are image recognition, general-purpose AI (multi-modal and including multi-agent systems), emotional AI (biometric recognition and also emulated empathy) and AI decision-support systems.

b. Methodology for selection of ethical principles and values

This task employed a rigorous methodology, combining both qualitative and quantitative aspects, in selecting the most relevant ethical principles and values pertaining to the AIOLIA use cases and research areas. The methodology consists of six key elements and while each bears a different weight, depending on the use case or research area, the selection takes all six elements into account in the selection of each named area. The methodology consists of the following: scientific, ethical and procedural soundness, comparability, comprehensiveness and relevance. The overarching consideration that informs all six elements of the methodology is the impact upon human cognition and behaviour. Each element will be explained in turn.

1. Scientifically sound

This element pertains to the existence of scientific research and literature on the relevant principles and values in relation to AI in general and within specific use cases and research areas in particular. This takes the form of scientific studies conducted in the field of AI ethics, law, moral philosophy, human-computer interaction, social robots research and other adjacent fields. In addition, the rich field of AI ethics and research has helped to concretise certain key values into ethical guidelines that are adopted by both the public and private sectors. These guidelines differ in its granularity, coverage and selection of values but share the commonality of lack of enforceability. As its name implies, these values and principles act as guidelines for ethical behaviour and choices when it comes to designing and deploying AI. On the other hand, some ethical principles have been adopted by law, expressed as legal requirements to adhere to certain standards. This is notably the case of the recommendations of the High-Level Expert Group on AI on the Ethical Guidelines for Trustworthy AI in 2019 which has now been reflected in the European Union's comprehensive legislation on AI, the Artificial Intelligence Act (AI Act 2024). The requirement to embed ethical principles as an obligation upon developers and deployers of AI systems can found within legislative provisions, such as the requirement for transparency (Art. 50 AI Act) or be found as part of technical standards to adhere to, such as the requirements in the AI Act for Chapter III, Section 2 relating to high-risk AI.

Another element that informs scientific soundness is that, unlike other fields of scientific enquiry such as climate change, scientific consensus is not required as an indicator of scientific soundness. Ethical principles and values are by nature contestable and subject to varying and contextual, interpretations. The standard adopted here in judging upon scientific soundness, is instead the level of engagement by the epistemic community rather than consensus. Certain use cases, such as decision support systems within recruitment or medicine, have been the subject of many research studies, allowing the project to draw upon the rich scientific literature in this space. However, AI research is also rapidly developing, thus requiring the project to also engage with and refer to emerging literature in relation to ethical concerns in emerging, albeit lesser researched areas, such as AI companions. To this end, the selection of the most relevant ethical principles and values for each use case and research area is informed by the depth of engagement by the epistemic community within each vertical (research area and use cases).

2. Ethically sound

The second element of the methodology pertains to gauging how ethically sound – in this case, how future proof – the chosen principle or value would be for the particular use case or research area. Ethical soundness, in this sense, is assessed from the level of abstraction in connection to its alignment with European values, such as values in relation to European digital regulation. The future proof element is mentioned in Recital 138 of the AI Act in relation to legislative measures and in the drafts of the Code of Practice in relation to General Purpose AI. Future proofing in relation to governance measures, including regulation or ethical principles, means that the target of governance should not tether to the technology itself, as technologies such as AI evolve at rapid speed. Being future-proof relates to the sustainability of the regulation or ethical principle even in light of technological change. An example of a legislative measure that set out to be future proof is the GDPR as the regulation relates to protection of personal data regardless of its method of transfer or processing (e.g. online data or physical drives) and is thus technologically neutral. While ethical values may intuitively seem to be future-proof, the benefits brought about by the diffusion of AI in society might challenge this. For example, explainability might be sidelined if AI becomes increasingly reliable on a large macro, societal level scale (e.g. healthcare benefits, economic productivity etc.) and bring about overall utilitarian benefits. Further, even as the GDPR is heralded as being future-proof, generative AI, notably through using data from the internet as training data and the difficulties of measuring output accuracy of generative AI models, challenges its future-proof quality. The rapid development of AI is challenging key conceptual norms which we assumed would hold the test of time. We acknowledge that the debate on AI ethics might look very different 5-10 years from now, especially as the relentless pursuit of 'artificial general intelligence', understood as AI systems that match human levels of intelligence across various domains, is well-underway. As such, thinking of the term future-proof in an absolute sense is perhaps unwise. AIOLIA instead adopts a more time-bound approach to future-proofness, limiting our conceptualisation, analysis and operationalisation of the chosen ethical principles to the time-bound horizon of 5 years.

3. Procedurally sound

The third element of the methodology used in the selection of the relevant ethical principles and values for the use cases and research areas in AIOLIA relate to the feasibility and practicality of technically implementing the chosen principles in practice. This includes asking whether there is technical capacity or capability to operationalise the ethical principle, whether there are financial blockages that could hinder implementation – for example, where operationalisation of an ethical principle is prohibitively expensive, whether there are organisational roadblocks – for example, certain levels of secrecy is part of operational necessity within law enforcement agencies, which may prevent the full operationalisation of the ethical principle of transparency. However, challenges in operationalising (certain) ethical principles does not automatically discount the principle from being a relevant consideration. This concern has to be weighed against the other elements in the methodology when it comes to choosing a relevant principle or value for a given use case or research area. Procedural soundness is compatible with the existence of contestability over the value or ethical principle. As mentioned before, ethical principles are not settled and the use and development of AI will continue to raise new ethical concerns, reinvigorate old concerns or make intra-value contestability more salient. Procedural soundness focuses on the feasibility of implementing an ethical principle or value in practice – implementation is entirely consistent with the existence of ethical contestability.

4. Comprehensiveness

The fourth element of the methodology touches upon comprehensiveness of the ethical principles or values that are considered and those eventually selected. This requires that we examine a wide range of ethical principles and values – including those that have been adopted as part of legislative requirements and well as the consideration of principles and values from different jurisdictions, belief systems and cultural backgrounds (e.g. European versus non-European). A comprehensive examination of the array of ethical principles and values ensures representation and guards against the dominance of certain ethical values over others. While the inclusion of international partners in AIOLIA is one way to ensure comprehensiveness, this methodological approach is similarly applied to values chosen by the European partners in the project. Where relevant, the project has referred to research that provides for comprehensive overviews – geographically, as well as in relation to depth and breadth - of the landscape of ethical principles and values relevant to AI.

5. Comparability

The fifth methodological element to motivate the selection of ethical principles and values for AIOLIA is the ability to compare and contrast the content, interpretation and operationalisation of ethical principles and values. In considering a comprehensive set of ethical principles, it is also pertinent to note that the acceptance, interpretation or adoption of ethical principles and values are not necessarily uniform or cohere across geographic areas as these differ according to values, priorities and realities – culturally, politically, economically. As such, the selection process is guided by intra-value comparability (between the same sets of values) by different actors, jurisdictions and use cases. An example of applying the comparability methodology is the consideration of ethical principles around real-time remote biometric facial recognition. In the AI Act, such a use case is prohibited, barring certain exceptions of its use by law enforcement (Art 5 AI Act). However, real-time remote biometric facial recognition can be deployed in other jurisdictions, including on account of ensuring the safety and security of persons, be it in public or professional (workplace) settings. An interesting aspect of comparability arises here – wherein, jurisdictions can compare the interpretations, uptake, applicability and pushback against the deployment of specific AI technologies as well as its ethical justifications. The comparative aspect is in itself a notable research subject but can also critically inform policy and governance of AI.

6. Relevance

The sixth methodological element is that of relevance. Relevance can be interpreted from different perspectives – such as how it relates to the use case(s) in question or through how much regulatory or policy attention it garners or should garner. At the same time, there can be lack of clarity over how particular ethical principles can be operationalised in a given use case or research area (even while admitting its importance). An example of the latter is the increased attention on the ethical challenges posed by companion AI, especially the case where chatbots are designed or used for therapy, friendship or intimate companion purposes. The principle of doing no harm to users of such services might be obvious, but its operationalisation can be complicated due to the lack of predictability in relation to content generated by large language models. While the context of such uses is informed by heightened layers of vulnerability (Luna, 2009; Malgieri, 2023), individuals have different psychological make-up and while some may benefit from sycophantic interactions with chatbots, others can be detrimentally harmed by them. Navigating these uncertainties is a key relevant concern, especially as these chatbots are seeing rapid adoption and popularity. During this research and selection stage, the project does not take one singular interpretation but leaves this question open – hence the selection of the most relevant ethical principles and values can draw from these different ways of interpreting relevance.

Impact on human cognition and behaviour

The final methodological consideration is a cross-cutting one, namely that the selection of use cases and the ethical principles and values pertaining to those use cases should engage with its impact upon human cognition and behaviour. When it comes to selecting the most relevant ethical principles and values in relation to each use case, the impact on cognition and behaviour will be addressed from two perspectives, namely its impact on professional behaviour, where AI is deployed in professional contexts (e.g. recruitment, medicine, security, safety engineering) and its impact on private behaviour (e.g. personal companions deployed in individual or family settings).

c. Method

Having laid out the methodological considerations, the method used for selecting the most relevant ethical principles and values consists of the following steps:

1. Comprehensive literature review of ethical principles and values in relation to AI (general) and specific to use cases and research areas.
2. Initial selection by CEPS based on the literature review, a shortlist of most relevant ethical principles and values provided by partners and based on consultations with the consortium work package lead.
3. Design of two surveys, taking into account the methodological considerations outlined.
4. Survey sent to AIOLIA partners (internal survey) in April 2025.
5. A methodology questionnaire sent to international partners on how and why they selected the ethical principles for their respective use cases.
6. Survey sent to AIOLIA Scientific Advisory Board and networks through AIOLIA partners, ERCIM, ADRA and EUREC (external survey) in May 2025.
7. Synthesis of overall responses from both surveys, noting overall convergence on values and specific concerns raised.
8. Feedback from the use case partners on the selection and number of principles.
9. Finalisation of the selection of the most relevant ethical principles and values for the use cases and research areas to inform the operational stages (co-creation, training) of AIOLIA.

d. Overview of the landscape of European and international ethical guidelines and principles

The rapid global development and deployment of artificial intelligence (AI) technologies also ushered in a proliferation of AI ethics guidelines across academic, institutional, and governance domains (Corrêa et al., 2023; Hagendorff, 2020). Some of these ethical principles have evolved from guiding aspirations into binding requirements found within legislative frameworks. The European Union's Artificial Intelligence Act (AI Act) is a key example of this, whereby ethical principles put forth by the High Level Expert Group on AI in the Ethics Guidelines for Trustworthy AI in 2021 – foregrounding values such as transparency, human oversight, and risk mitigation, are now part of the EU AI Act. While the AI Act encourages voluntary compliance in low-risk contexts, it mandates stringent requirements for high-risk systems, including documentation, human oversight, and post-market surveillance. An incentive based (e.g. regulatory sandbox), co-governance (AI Office and the national supervisory authorities) and punitive approach, such as having high fines for violations, balances the twin ambitions of the Act - namely to encourage innovation while at the same time respecting fundamental rights.

Another notable milestone in the development of AI ethics is UNESCO's 2021 Recommendations on the Ethics of AI which was endorsed by all 193 member states. It is the first global normative instrument grounded in international law and centres on human rights, sustainability, diversity, and peace. The framework also promotes inclusion of underrepresented regions, especially in the Global Majority (UNESCO, 2021). Another clear example is the OECD's AI Principles, which aims to promote inclusive growth, transparency, and accountability. In turn, the Council of Europe's Framework Convention on AI, Democracy and the Rule of Law is the first transversal, legally binding international instrument which that aims to ensure that AI development and deployment respects human rights, democracy and the rule of law. These documents signal a clear normative direction internationally to take AI and ethics seriously and helps to shape policy and guide best practices, especially when regulation centred around AI remain nascent in many parts of the world.

Furthermore, international efforts are also being led by the UN. The UN Global Digital Compact (GDC) was adopted at the landmark Summit for the Future event in 2024. The UN GDC signalled the importance of inclusion and global buy-in towards AI governance, including through initiatives such as the Scientific Panel on AI and the Global Dialogue on AI Governance. The GDC highlights principles such as transparency, accountability, and human rights in developing and deploying AI, signalling a significant global development. AI ethics is thus moving beyond being a theoretical and academic discourse but is increasingly guiding concrete policy and regulatory action. In general, the shift from soft law principles to enforceable regulation reflects a positive trend in AI governance, where ethical principles are transformed into legal requirements and compliance mechanisms. In parallel, certain non-binding frameworks have also gained global legitimacy and influence.

This surge is evident in the meta-analysis of 200 guidelines conducted by Corrêa et al. The document reviews ethical guidelines published by diverse actors, including governments, civil society groups, companies, and academic institutions, offering a detailed overview of the common ethical principles and institutional and global patterns currently shaping democratic AI governance. The study thus provides a critical snapshot of the field's current state, as of 2023 (Corrêa et al., 2023). One of Corrêa et al.'s core findings is the significant convergence on specific core values, such as transparency, accountability, fairness, privacy and safety within AI ethics guidelines, suggesting the emergence of a shared normative vocabulary as a foundation for AI ethics. It identifies five key areas of normative convergence that offer a foundation for harmonisation. The most cited values are transparency, explainability, and auditability, especially among government actors who advocate for public oversight and clarity in algorithmic systems. These principles ensure that AI systems are understandable to non-experts and subject to review. Other convergent core values that were identified include accountability (and liability), justice, fairness, and non-discrimination and finally, privacy. In turn, values of reliability, safety and trustworthiness are typically highlighted by private companies in order to foster public confidence and market adoption (Corrêa et al., 2023).

These values align with findings from other comprehensive studies on the state of AI ethics (Jobin et al., 2019; Hagendorff, 2020), even as differences appear in terms of how values such as reliability or trust are framed. Semantic divergences are also present: key principles such as fairness are defined differently across guidelines, creating gaps in implementation (Corrêa et al., 2023). Further, fewer than 2% of guidelines include concrete mechanisms, such as audits or controls, to enforce ethical principles and values. This suggests that unless transformed into legislative requirements, the field as a whole remains principle-driven but underdeveloped in relation to governance infrastructures (Corrêa et al., 2023).

Despite these general trends of convergence, there are critical asymmetries that expose structural inequalities in the global AI ethics discourse.

Key asymmetries highlighted include geopolitical, typological, institutional and temporal asymmetries. First, geopolitical asymmetries are present as most AI ethics guidelines originate in the Global North.

Despite the emergence of AI powers such as China and emergent actors such as Brazil, their contributions remain vastly underrepresented. This limits participation in shaping future global norms and frameworks, cements power imbalances in AI development and perpetuates a form of 'normative colonialism', where ethical frameworks are based only on Western liberal values, such as individual privacy and procedural fairness, while overlooking local governance traditions, collective rights, and socio-political realities from the Global Majority.

Another key asymmetry is typological. Most guidelines offer general principles without specifying tools for the implementation and only 2% include assessment mechanisms or metrics. Moreover, 4.5% suggest legally binding measures, mainly because only governmental institutions, representing just 24% of the dataset, can propose enforceable rules. This gap leads to “ethics washing,” where guidelines function as soft law serving to protect and advance reputational goals rather than accountability (Corrêa et al., 2023). Most ethical guidelines remain non-binding. The overwhelming reliance on soft law undermines the transformative potential of AI ethics. Instead of serving as enforceable governance tools, many guidelines function as symbolic declarations, lending legitimacy to organisations without holding them accountable. This asymmetry aligns allows public and private actors to signal virtue while avoiding legal scrutiny (Wagner, 2018). Ethics reinforce reputational capital without disrupting existing deep power structures. Hagendorff also warned that many frameworks act as public relations (PR) tools rather than regulatory instruments (Hagendorff, 2020). Further, critical concerns such as ecological sustainability, labour rights, and democratic governance are often neglected.

Another key asymmetry highlighted was institutional asymmetry, wherein private sector stakeholders dominate the production of AI ethics guidelines, thus increasing the risk of regulatory capture (Corrêa et al., 2023). Gender imbalances are also present, as most of the AI ethics documents are authored by men (Hagendorff, 2020).

Finally, temporal asymmetries also inform the space of AI ethics guidelines. Nearly two-thirds of the 200 guidelines were published during the ‘AI ethics boom’ between 2016 and 2018 (Corrêa et al., 2023). Since then, activity has slowed, suggesting that sustained monitoring, revision, or implementation has not matched early enthusiasm. While this may be the case for some actors and jurisdictions, the decline might also reflect a maturation of the field. Many of the previously aspirational ethical values are being codified into binding legislative instruments, as mentioned, most notably through the European Union’s AI Act. Similarly, technical and guiding standards (such as through ISO or NIST) are gaining traction among private actors, indicating a broader industry buy-in. At the same time, the decline of interest in the field could also reflect deeper tensions in the sphere of AI global governance, as part of an ongoing deregulatory agenda. The speech by the US Vice-President JD Vance in the Paris AI Summit in February 2025 clearly signalled the push for deregulating AI to pursue innovation and competitiveness, seemingly at the cost of fundamental rights and ethics.

In short, the terrain of AI governance is shifting - from aspirational proliferation to selective institutionalisation, alongside competing global visions for how ethics should be operationalised in law, policy, and practice. As ethical principles become more entrenched in standards and legislation, critical attention is needed to ensure these commitments remain adaptive, inclusive, and accountable.

In conclusion, while legal instruments like the EU AI Act and global frameworks such as UNESCO’s Recommendation mark significant steps forward in codifying ethical principles, critical gaps remain, particularly in their operationalisation. As highlighted in the literature, core values such as transparency, accountability, and fairness are frequently cited but rarely supported by concrete implementation tools or mechanisms. Political misuse, environmental risks, and structural inequalities, especially regarding gender and geographic representation, continue to be insufficiently addressed.

Funded by the European Union

This underscores the urgent need to move from aspirational ethics to applied, context-sensitive practices that are enforceable, inclusive, and adaptive to technological change. AIOLIA is directly positioned to address this need. By engaging academic and industrial actors in co-creating ethical guidelines grounded in real-world use cases and focusing on AI's behavioural and cognitive impacts, the project offers a practical model for embedding ethics into the design and deployment of AI systems. AIOLIA's methodology provides a robust framework for operationalising, adapting and drawing best practices from AI ethics. These practices will in turn will be disseminated to wider groups of stakeholders through planned trainings and workshops.

2. Use cases and research areas

The European academic and industrial partners chose 6 use cases (see table 1): four in the sphere of professional behaviour and two in the sphere of personal behaviour. The use of AI impacts upon cognition in both spheres, with the change in the professional sphere being concretised as a change in expertise. Each use case involves one or more AI research areas - namely AI decision-support systems, general-purpose AI (GPAI), image recognition, and emotional AI. An overview is provided in the table below.

#	Sphere of influence	Use case	AI research areas	Academic partners	Industrial partners
1	Change in human expertise and professional behaviour	Medical doctors (radiologists, surgeons) using AI tools in diagnostics and treatment	Decision-support systems; Image recognition	AUMC	Oxipit; Afliant
2		Safety engineers using AI tools to speed up software release approvals	Decision-support systems; General-purpose AI (GPAI)	CEA	NIT
3		Recruiters using AI tools in hiring processes	Decision-support systems	CENTRIC	Eticas + Barcelona Activa
4		Security professionals using AI tools to detect hate speech	Decision-support systems; General-purpose AI (GPAI)	CENTRIC	NEBRC; South Yorkshire Police
5	Change in human cognition and private behaviour	AI systems as individual and family-level virtual assistants	General-purpose AI (GPAI); Emotional AI	THWS	Immerstream; Aurora First
6		Deepfake therapy for processing trauma and grief	Multi-modal general-purpose AI (GPAI); emotional AI	AUMC	ARQ Centrum '45

a. Use cases that impact upon professional behaviour

Use case 1: Medical doctors (radiologists, surgeons) using AI tools in diagnostics and treatment

There are two industrial partners working on two different scenarios for this use case: Oxipit and Afliant. Technology-wise, both partners' tools can be categorised as image recognition and decision-support systems.

Oxipit's chest x-ray suite looks for 75 most common radiological findings and identifies high confidence normal chest x-rays. This allows for three applications. The first one is the automation of 40% of chest x-ray reports, given that the product is CE certified (per MDR - EU Medical Device Regulation) for autonomous reporting of high-confidence normals. The second application is speeding up the radiological workflow through computer-assisted diagnosis (CAD) functionality. It identifies findings and provides initial indications along with heatmaps (overlay on chest x-ray image) to locate the suggested findings, which the radiologist can consult. Additionally, the CAD functionality can allow triage of chest x-rays based on the severity of the findings. The third application of the chest x-ray suite is screening the final radiological reports for potential misalignment with AI image interpretation. This serves as a safety net, allowing potentially missed findings to be caught.

Afliant's AI solution, on the other hand, supports the treatment of abdominal aortic aneurysms (AAAs) at three key decision points: pre-operative planning, intra-operative execution, and post-operative follow-up. The AI agent's scope includes automatically extracting morphological features from CT scans to populate a planning worksheet (e.g., vessel diameters and lengths, aneurysm characteristics), suggesting procedural details such as the choice of prosthesis and surgical access site, and highlighting critical vessels in real-time during surgery (via integration in a simulator environment). Post-operatively, the AI agent aims to predict the likelihood of complications such as endoleaks, guiding clinicians in determining patient follow-up schedules.

This integrated decision-support system is designed to reduce planning time, enhance procedural accuracy, and rationalize patient monitoring. Rather than replacing surgeons, the AI tool offers real-time, context-aware insights that can help reduce cognitive load and improve surgical outcomes. The technology is also integrated with a surgical simulator (ANGIO Mentor) which provides beneficial training for trainee surgeons.

Use case 2: Safety engineers using AI tools to speed up software release approvals

The academic partner CEA has partnered with NIT to study the AI use in safety management in the automotive industry. NIT is conducting a series of consultancy projects for the automotive industry, in which they are tasked to conduct safety analyses for the automotive products. Many automotive products include hardware and algorithms for autonomous and assisted driving. Safety analyses of these products must follow strict guidelines from standards such as ISO 26262 and ISO 21448. In relation to ISO 21448 – SOTIF (Safety of the Intended Functionality), potential hazards and risks are assessed using the System Theoretic Process Analysis (STPA). The STPA has the goal to analyze control actions between the system components, identify unsafe control actions and identify whether they can lead to loss scenarios (which can cause harm to road users). This analysis is essential since its accuracy influences the safety-related architecture and safety measures which are needed to make vehicles safe for use.

The automation existing for general software development in CI/CD pipelines is now starting to become employed for other connected tasks, including for safety analyses and rechecks. NIT is developing AI agents which aim to facilitate and speed up the development process involved in STPA. This tool would help by giving initial suggestions on how to conduct the analysis. It would also kick-off the analysis based on the general description of the system architecture and by providing the initial list of control actions and failure modes. It would also help to assess nearly completed analyses in order to identify remaining gaps and suggest fixes. NIT is beginning to deploy this in preliminary studies alongside an independent parallel manual process.

Use case 3: Recruiters using AI tools in hiring processes

Eticas has worked on auditing AI-driven hiring systems, focusing on bias detection and fairness in algorithmic recruitment. One of its key projects involved Barcelona Activa's hiring process, where the goal was to analyse how AI decision support systems assisted in screening and selecting candidates. Some of these tasks include processing large volumes of candidate data: extracting and analysing structured information from CVs, application forms, and interview records to assess qualifications, skills, and work experience. This involves parsing unstructured text, matching candidate profiles to job descriptions, and scoring applicants based on criteria such as educational background, work history, and specific skills.

Eticas' approach included evaluating fairness metrics in both the screening and final hiring stages, identifying protected attributes that could inadvertently act as indicators of gender or ethnicity, and comparing candidate distributions to broader population data to assess representation disparities.

Use case 4: Security professionals using AI tools to detect hate speech

Law enforcement authorities are considering using AI tools that can scan social media content over multiple platforms in real time to identify and flag online hate speech which incites violence or hatred against vulnerable and minority communities. Flagged content is raised to a law enforcement officer who decides whether to pursue an investigation into the individual based on specific criteria (e.g. incitement to violence or harm, likelihood of harm, past content).

The AI tools use natural language processing and machine learning algorithms. They are trained on large public datasets where content was manually labelled as "hate" or "not hate". Their capabilities are intended to improve over time: for example, if an officer deems content as hate and requiring investigation, this classification can be fed back into the system and used for re-training and fine-tuning.

b. Use cases that impact upon private behaviour

Use case 5: AI systems as personal and family virtual assistants

THWS partners have proposed two distinct personal companion scenarios. The first involves a one-to-one virtual assistant, geared toward immersive 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 large language model. The second focuses on a family-oriented assistant that aims to engage with multiple individuals, such as parents and children, within a shared chat environment. This service aims to help users to trace their preferences and account for them when planning joined activities, shopping or managing family calendars. The latter service is still at an ideation stage as of June 2025.

Use case 6: Deepfake AI-based psychotherapy for processing trauma and grief

In the healthcare context, deepfake technology can be applied within psychotherapy, where AI-generated hyper-realistic video footage can be used to simulate conversations with, for instance, 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.

The implementation of a deepfake therapeutic protocol for such trauma treatment is currently being further studied. 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 potentially pose numerous ethical challenges.

c. Research areas

General-purpose AI

The AI Act defines a general-purpose AI as an AI model, usually trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks, and that can be integrated into a variety of downstream systems or applications (Art 3(63) AI Act). The large generative AI models behind applications such as ChatGPT and Claude are examples of general-purpose AI because they allow for flexible generation of content, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks. Single modality and multi-modal models can equally be considered general-purpose if they are flexible enough. Currently, generative AI technologies are the technologies which exhibit enough generality in their capabilities to be considered general-purpose, and while we use these two terms interchangeably in this report, not all generative AI is general-purpose and vice-versa, not all GPAI must necessarily be based on generative AI technology (Triguero et al., 2024).

An AI agent is a software system that executes tasks in an automated way and that uses general-purpose AI for the decision-making over the control flow. This means that an AI agent is like a virtual assistant to a human user: the human user sets a goal and a possible set of actions and the assistant then goes on to execute tasks autonomously including taking the decision for what task to take in a given moment (Gabriel & Manzini, 2024). Another way to describe an AI agent is that it is a system on top of a GPAI model that has access to tools, such that it can not only generate action plans but also execute various actions.

Emotional AI

Emotional AI (also known as affective computing) refers to both emotion recognition and emotion emulation: the capacity to identify, quantify, respond to, or simulate affective states such as emotions and cognitive states (IEEE Std. 7014-2024). Techniques used to do emotion recognition include sentiment analysis of online language; facial coding of expressions and eye-tracking via image recognition algorithms; voice analytics; analysis of biometric data such as muscle activity, heart activity and respiration activity via wearables, and others. Emotion emulation refers to the ability to imitate emotions such that their functions are closely replicated. For example, a chatbot may be able to console users such as by showing care, without itself experiencing any emotive or inner states.

While emotional recognition, in terms of using AI to infer emotions of natural persons, is prohibited within education and workplace institutions in the EU AI Act (Art. 5 AI Act), its deployment in other environments and use cases are considered as part of high-risk AI in the AI Act, thus requiring both the provider and deployer of such systems to comply with a plethora of legal requirements. Furthermore, while the prohibitions are in place within educational and workplace settings, AI systems that detect physical states such as tiredness and stress are not prohibited (Recital 18 AI Act). Despite the clear stance on emotional AI taken in the EU, this technology is not similarly prohibited or considered high risk in other jurisdictions and can hence be an interesting point of comparison for AIOLIA.

AI decision-support systems

AI decision-support systems broadly refer to tools or platforms which use AI, including machine learning and natural language processing, to assist humans in making decisions. Cognitive work applications include providing information in the form of analysis, predictions, recommendations, alerts or others; industrial applications include supply chain optimization, energy management, predictive maintenance, quality control, and production planning (Soori et al., 2024). Under the GDPR, one has the right not to be subject to a decision based solely on automated processing, including profiling, barring some exceptions (Art. 22 GDPR). In a similar vein, Art. 14 AI Act requires that those deploying AI systems exercise human oversight over high-risk systems within the Act. AI decision-support systems can feature in multi-level decision making and can have different levels of personalisation and explanatory capabilities (C-634/21 SCHUFA Holding (Scoring)).

Image recognition

Image recognition is the process of identifying or extracting objects or features from a digital image. The process involves training a machine learning model to learn relevant features from sample images to be able to identify these features in new images. It often uses techniques such as convolutional neural networks (CNNs) to detect patterns like shapes, textures, or colours. Image recognition is widely used in applications such as facial recognition, medical imaging, autonomous vehicles, and quality inspection in manufacturing.

3. Selection of ethical principles and values

This section will broadly define the values that were considered across the use cases. The relevant principles will then be raised and discussed pertinent to each use case. We will first motivate an initial selection of values based on the literature review, concretise their implications, present the survey responses, and explain how the values were integrated into a final selection. Following this, we will discuss the selection of principles for the research areas. The numbers below do not imply a hierarchy, it is sorted in this manner only for purposes of presentational coherence. As mentioned, the judgement of the suitability of the principles cannot be evaluated on its own but will always be tied to a given context.

a. Considered ethical principles and values (definitions and tensions)

1. Autonomy and non-manipulation

The principle of autonomy, derived from moral and political philosophy, relates to a ‘capacity to be one’s own person, to live one’s life according to reasons and motives that are taken as one’s own and not the product of manipulative or distorting external forces.’ (Christman, 2020). Personal autonomy thus relates to the ability to be a self-governing agent, one that is able to take informed decisions and make free choices (Buss & Westlund, 2018). An AI system that respects human autonomy should support and complement the exercise of this autonomy, including being designed to support human decision-making and not to take decisions on behalf of the individual.

Manipulation and other forms of deception can infringe this autonomy, whereby one's actions are no longer 'self-governed', but controlled, shaped or dictated by others. AI-enabled influence that amounts to deception or manipulation directly threatens autonomy when it causes an individual to take an action they would otherwise not have taken (Susser et al., 2019). AI-enabled manipulation can be hidden and therefore difficult for individuals to detect or resist, especially when content is personalised according to individual relevance and interests. It is also a threat to well-being if said forms of manipulation lead to harms – in the physical, psychological or economic senses. A respect for autonomy also means that individual vulnerabilities – such as based on conditions of age, disability or socio-economic conditions, should not be exploited for the benefit of others. (Art 5.1.b AI Act, Recital 29 AI Act).

Manipulation can be intentional, when a specific AI model is designed to be manipulative, or unintentional, where it ends up having manipulative effects upon users (Art. 5.1.a AI Act). The latter includes unintended consequences due to the inherent limitations of the training paradigm, the unpredictability of outputs due to statistical probability and the difficulty of aligning the system with human values (Amodei et al., 2017). AI that is manipulative, deceptive and exploitative (of certain grounds) are prohibited under the EU AI Act (Article 5.1.a; Article 5.1.b)). However, the line between acceptable influence that complements and enhances autonomy versus influence through nudging, persuasion, or through the deployment of increasingly anthropomorphic AI that causes (over) dependence and therein potentially manipulative, are far from clear (Commission Guidelines on Prohibited AI, 2025) and should be assessed on a case to case basis.

2. Human oversight

Human oversight is an important ethical principle in relation to the development and deployment of AI systems. The EU High-Level Expert Group on Trustworthy AI listed human oversight as one of the key requirements for trustworthy AI, ensuring meaningful human agency and oversight to avoid automation bias and to safeguard fundamental rights. The principle of human oversight is also foregrounded on account of the varying degrees of autonomy that an AI system can display (Art. 3 AI Act): while it is designed to have some degree of independence of actions from human involvement and of capabilities to operate without human intervention (Recital 23 AI Act), these should be constrained in proportion to the risks. Additionally, oversight is important due to the so-called 'black-box' nature of certain AI systems that use deep learning techniques, whereby how decisions or recommendations are generated can be unclear to both designers and end-users.

The principle ensures that AI systems work as intended and is also broadly related to the principle of non-maleficence: ensuring AI use does not do harm. Both these elements can be operationalised in practice by having robust risk management practices in place (Art 9 AI Act) and ensuring a human is present and involved in different levels and/or different stages of the AI life cycle, e.g. it can relate to model development, system use and monitoring. While having a human-in-the-loop—the capability to intervene in every decision cycle of the system (European Commission, 2020), is a well-known measure of ensuring human oversight, oversight practices are much wider and can include tests, monitoring, audits and assessments by internal units, customers, users, independent third parties or governmental entities (Jobin et al., 2019). Oversight ensures that accountability and responsibility can always be attributed to humans or legal entities, including in terms of enabling meaningful access to remedy (UNESCO, 2021).

3. Human dignity

Human dignity is a foundational value that informs the international human rights law framework and one of the key values in the EU. In addition to this foundational understanding of human dignity that is traced to the inherent worth of the human being, human dignity is also both a fundamental right under the EU Charter of Fundamental Rights and an ethical value (McCrudden, 2008). Since it traces to the worth that is considered to be inherent in every human, respect for human dignity straightforwardly entails that one's right to life, physical and mental integrity, and the prohibition of torture or degrading treatment is respected (European Union, 2010). The respect for human dignity can also mean respect for the distinctive qualities of being human compared to non-human animals, including the capacities of reasoning, rationality and autonomous decision-making. However, the term human dignity has been subjected to stringent criticism on account of its lack of specificity (Pinker, 2008), including seemingly enabling contrasting positions to appeal to the same human dignity argument (Rueda et. al., 2025). Its wide definitional ambit also means that it might be difficult for the principle to be operationalised. At the same time, human dignity remains a compelling principle, typically invoked when some notion of human uniqueness or worth is deemed to be under threat. While admitting the amorphous nature of the term, its embeddedness in legal instruments mean that some degree of specificity has been introduced through case law and legal interpretation (McCrudden, 2008). This is also the case in relation how human dignity can be interpreted in relation to AI (Teo, 2023).

Human dignity has been named as one of the key factors behind the prohibited category of AI systems in the EU AI Act (Commission Guidelines, 2025). In relation to operationalising the protection of human dignity, we can think of, among others, that the development and deployment of AI should take into account the fact that human life and experience are inherently complex, subjective and at times irreducible to data. When datafying certain aspects of human life, especially emotions, care should be taken to ensure that this principle is respected, including enabling the individual to challenge the inference, recommendation or decision of the AI system (IEEE Std. 7014-2024). Certain uses of emotion recognition systems have been deemed to be such an affront to human dignity and fundamental rights that these are prohibited in the EU (Art 5.1.f AI Act).

4. Non-maleficence

Non-maleficence is a core principle derived from the field of medical ethics and is about the avoidance of harm. However, a key distinction in relation to perspectives is essential when distinguishing its use in medical ethics and in relation to AI ethics. From the perspective of medical ethics, the term non-maleficence relates to a patient that has departed from the norm of good health. Operationalising non-maleficence from this perspective thus entails returning the patient to this norm. This background assumption is however not present in AI ethics and neither is the goal of AI ethics concerned with the attainment of an ideal state. Thus, non-maleficence in relation to AI ethics should be informed by this distinction.

One way to operationalise the concept in relation to AI is to ensure that AI systems should not cause bodily harms or emotional harms to humans (Ryan & Stahl, 2021). Emotional harms include psychological harms, for example in the context of companion AI applications, as well as bodily harm, such as being encouraged to carry out forms of self-harm or inflicting harm to others by or through the use of AI. The principle does not only apply to AI in deployment but also in development, especially the process of labelling training data or training with human feedback.

While this principle can be more directly operationalised in the medical context and in the context of predictive processing tools, in the sense that a physician should do no harm to patients, the rise of generative AI development and adoption, underlines the generality of the principle across applications and present new challenges such as cyber-security or alignment (Hagendorff, 2024). One key manner this principle can be operationalised includes the adoption of risk management and safety practices throughout the AI lifecycle: in design, development and deployment, such as what is mandated by the AI Act for high-risk AI systems (Art. 9 AI Act) and for general purpose AI posing systemic risks. This includes ensuring appropriate red-teaming measures and putting guard rails in place, such as for large language models to refuse to engage in or encouraging harmful behaviours.

5. Robustness, reliability and safety

In the Ethics Guidelines for Trustworthy AI, the EU High Level Expert Group on AI emphasised that technical robustness and safety are not merely technical measures but also feature as ethical considerations. The Group noted that that this should include resilience measures against attacks, putting in place fallback plans and general safety considerations. A key element of the ethical principles on robustness, reliability and safety relate to the requirement for the outputs of AI systems to be accurate and work reliably and as intended, not only during training or testing but throughout the entire life cycle. However, the principles can be differentially applied depending on the AI system. An image recognition system for medical usage needs to be robust, safe and reliable, but a chatbot prompted to generate poems or other creative outputs need not be robust in the same manner.

The principles relate to harm minimisation and prevention, where possible, via risk management practices including continuous assessment and monitoring (European Commission, 2020; OECD, 2024, NIST, 2023). Safety also requires that the AI system used should be resilient against adversarial actions such as cybersecurity attacks. A proper consideration and operationalisation of the ethical principles means to ensure not only individual physical safety for the users of the system, for example, in self-driving cars, but also the cybersecurity risks that may be present, such as possible risks of hacking and the resilience of inter-connected systems that can surface vulnerabilities.

The link between reliability and accuracy is equally important. The ethical significance of accuracy is two-fold. Firstly, the developers and deployers should acknowledge the probabilistic nature of AI outputs and the inherent impossibility for AI outcomes to be error-free. This means setting the right expectations for the users and subjects of the technology. Secondly, the limitations in accuracy and performance—not as development flaws but in terms of the state-of-the-art—must inform cost-benefit analysis and acceptable risk thresholds in risk management frameworks. Consequently, the ethical consideration of the accuracy, and therein, reliability, of a technology can lead to a decision to not deploy the system.

6. Privacy and data protection

Privacy is one of EU's fundamental rights, and since the GDPR came into force, it has been central in discourses on societal impacts of technology. This is the case not only in Europe but also increasingly worldwide, with the GDPR principles broadly being accepted as a ‘gold standard’ (Lubin, 2022) and is a famous beneficiary of the so-called ‘Brussels effect.’ (Bradford, 2020). Core requirements of GDPR is that users must be able to know and control, e.g. be able to give informed consent about, what data about them is collected, processed and shared, and for what purpose; data processing should be in line with data protection principles (data minimisation, purpose limitation, fairness, transparency etc.) and respect data subject rights (e.g. to rectification, erasure, access, objection etc).

The principles of privacy, consent and data protection create significant tension with AI systems' hunger for data. This has intensified with the recent surge in generative AI development, as well as with new data and AI legislation in the EU, namely the EU AI Act and the data spaces. The new legislations raise direct contradictions with the GDPR, for example, contrary to GDPR, the AI Act allows for the processing of sensitive data since this can be used to combat bias (Art 10.5 AI Act; van Bekkum, 2025). It also raises indirect contradictions, such as the technical difficulties of exercising rectification and erasure rights of the data subjects in relation to generative AI. More broadly, the new opportunities for personalisation and research benefits are forcing society to revisit definitions of privacy, not only in a legal sense but also normatively. Clarity on the appropriate legal basis for data processing, especially in relation to the training of GPAI models, remains elusive and contested.

7. Freedom of expression and risk of censorship

The freedom of expression is a cornerstone human right and is protected under various international human rights instruments, such as the UDHR (Article 19), International Covenant on Civil and Political Rights (Article 19) and the European Convention on Human Rights (Article 10). In the EU, this right is also protected under Article 11 of the Charter of Fundamental Rights. The freedom of expression can be found together with the freedom of information and both these freedoms reinforce each other – where information is restricted, the freedom of expression suffers. Where freedom of expression is restricted, the flow will be negatively impacted, hence curtailing access to information.

Freedom of speech is an essential element that contributes to a thriving democracy. A population that can freely exercise these freedoms for democratic deliberation can bring forth debate, raise concerns, challenge power and demand for accountability and rights. At the same time, the freedom of expression and information are not absolute. They can be subjected to restrictions to protect the interests of national security, public safety and other legally justified grounds, provided that the measures taken to restrict speech are legally justified and proportionate. Authorities are thus able to take action to restrict hate speech and incitement to violence. A balance needs to be struck between restricting speech and information on the one hand and the prevention of harm to minority communities and violence on the other hand. Unjustified censorship and mass surveillance typically amount to human rights violations as they breach not only the freedom of expression and information but also key European values and laws on privacy and non-discrimination.

In short, actors taking action to monitor and restrict speech, such as law enforcement authorities, must respect the freedom of expression and avoid practices resembling unjustified censorship and mass surveillance.

8. Non-bias, non-discrimination and fairness

This set of principles upholds the idea that an individual must be treated fairly by AI systems (Corrêa et al., 2023), regardless of the different sensitive attributes that may characterise them, including gender (Buolamwini & Gebru, 2018), age, race, ethnic origin, and specific socio-economic situation. Bias is one of the biggest challenges to mitigate in AI systems. Bias is not an original technical fault – biased AI systems reflect the bias present in society (Wachter et al., 2021). However, AI systems are involved in a negative feedback loop that can perpetuate it: when bias is embedded in the training data, AI outputs can lead to discriminatory decisions and outcomes, that, in turn, feed into new training data, perpetuating even more biased systems. Moreover, a source of discrimination can be developer bias, which gets embedded in the AI system design (Mehrabi et al., 2021). A key suggestion on combatting bias to the inclusion of more representative datasets, namely one that reflects captures the diverse (racial, age, ethnicity, gender etc) composition of different segments of society. However, this might not always be possible due to trust deficits, inaccessibility or unwillingness. The use of synthetic data has been proposed as a measure to address bias and to promote the design of fairer AI systems. Synthetic data can augment training datasets that lack minority representation and thus improve the overall accuracy of the algorithmic model (Juwara et. al., 2024). In general, upholding fairness requires that AI be purpose-built, include stakeholder participation, include impact assessments on various societal dimensions in order to minimise and manage bias, and be developed to enhance human well-being rather than pose risks to it (Ryan & Stahl, 2021).

9. Transparency and explainability

Transparency is one of the most prevalent AI principles, but its definitions and operationalisation can vary greatly (Jobin et al., 2019). Under the EU AI Act, transparency means that AI systems are 'developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their rights.' It is evident from this definition that the principle of transparency is multi-faceted and it involves information sharing at different levels, and in different directions. One way to dissect transparency is to differentiate between horizontal transparency—between the AI providers and the users; and vertical transparency—between the AI providers and regulators (Söderlund et al., 2024).

Transparency and explainability go hand in hand in the literature, yet explainability has a distinctive focus which merits comparable attention, especially in legal terms. Normatively, explainability refers to the ability to explain why and how a certain AI-generated or recommended output has taken place (European Commission, 2020). Its legal form, the right to explanation, has shifted from being briefly mentioned in the Recitals of the GDPR and where the 'right' itself was subjected to interpretive contestations (Selbst & Powles, 2017; see however Wachter et al., 2017), to a stage where it is now more clarified in case law (C-203/22 Dun & Bradstreet Austria). The requirement is now being properly incorporated in the AI Act, as a right in relation to high-risk systems (Metikoš & Ausloos, 2025). The two legislations broadly align on certain dimensions, such as the kinds of explanations that are required, but they also differ in some aspects (Kaminski & Malgieri, 2025). For example, while GDPR targets automated decision-making, the AI Act does not specify this as a criterion, but rather places importance on the level of impact that the system has (Metikoš & Ausloos, 2025), which is seen as a positive improvement.

The role of transparency and explainability is to allow for addressing risks by facilitating human oversight (Panigutti et al., 2023), and to foster public trust.

10. Accountability and responsibility

This principle pertains to blameworthiness and responsibility. Who takes the blame in case of harm– the professional, the AI provider, the AI developer, or other parties (Matthias, 2004)? Should there be joint responsibility or new forms of responsibility that is more relational in nature (Liu & Zawieska, 2020)? This principle is about ensuring there are accessible mechanisms and procedures in place that allow for questioning, contestability and redress of in case of harm or adverse impacts (European Commission, 2020). This includes the ability to audit a given system, exercise oversight, and perform impact assessments, as well as mechanisms such as whistle-blowers' protection (UNESCO, 2023).

Closely related to the idea of responsibility is liability, but this is a strictly legal term: if a person is liable for an act, "they can be convicted, punished, or obliged to pay compensation solely by virtue of having performed a certain act, regardless of their state of mind at the time" (Kiener, 2024). As a matter of liability, AI presents many challenges to existing legal systems due to its complicated value chain and the "many hands" problem—where multiple actors contribute to an outcome, and no single individual can be held solely responsible (Coeckelbergh, 2020). This is especially complicated if one is trying to establish particular technical sources of fault; instead, framing AI harm as a socio-technical problem can help focus accountability measures onto the people, who have the knowledge and control over the various socio-technical risk factors, such as human-computer interaction, organisational culture, and procedural arrangements around development and deployment (Fraser & Suzor, 2025).

11. Risk of over-reliance and de-skilling

The ethical principle on the possible risk of over-reliance on AI systems comes with the attendant risks of the lack of meaningful human oversight and eventual deskilling. This principle concerns risks of automating processes and decisions in such ways that it leads to over-trusting them. This can create a negative cycle: the false sense of reliability can potentially feed back into the way the system is deployed and used, leading to higher levels of reliance. This is important not only due to the risk of harm but because it can lead to cognitive offloading in the judgement of different scenarios and skill attrition due to lack of practice (Crowston et al., 2025). Research demonstrates the existence of automation bias in the use of AI systems, where humans tend to defer to recommendations provided by such systems on account of its perceived neutrality and objectivity (Alon-Barkat & Busuioc, 2023). The risk of over-reliance and de-skilling is linked to the idea of conferring epistemic authority to intelligent systems (Abdelwanis et al., 2024) and also potential longer-term impacts on labour.

12. Labour impacts

Labour impacts can refer to job displacement, changes in job quality, shifts in the demand of certain skills, changes in job specifications, and other similar effects of AI use. All these aspects, in one way or another, impact upon the labour market, including changing labour force demands and desired skillsets. These include both direct effects, such as automation of tasks that reduce the need for human workers, and indirect effects, such as changes in organisational dynamics, coordination and hiring (CEPS, 2024). Relevant factors for these impacts include productivity increases due to the partial or full automation (Kelly, 2025) of certain tasks, reduction in economic costs, and shifting power to the technology developers and deployers. These labour impacts may also not be equally distributed as disaggregated impacts may be seen on already disadvantaged or vulnerable groups according to age or gender. Labour market effects relate to labour rights which require commitment to ensure appropriate upskilling, re-training and compensation for workers affected by adoption of AI.

13. Societal well-being and environmental sustainability

AI should contribute to societal well-being. It should not lead to negative impacts on public health and safety, democratic processes, economic security and the flow of information (European Commission, 2020). Additionally, social AI (Shevlin, 2025) use might have negative effects on social interactions and relationships. To consider societal well-being in AI development means to plan for and mitigate the indirect effects technology can bring, looking beyond merely its immediate impacts.

Ensuring societal well-being includes trying to gain foresight on how AI systems can interact with each other in the same deployment context as well as across different applications. Given increasing adoption and diffusion of AI in different areas of human life, risks from AI can become systemic: it can become increasingly difficult to isolate sources of fault to single systems or events. Negative effects could result in chain reactions and disruptions in critical sectors which are harmful to entire cities, domain activities or communities (AI Act, 2024). The societal level impacts are considered as systemic risks under the AI Act and various efforts, including having in place a Code of Practice, are underway as of June 2025 to catalyse private sector commitment towards addressing these risks.

Additionally, environmental sustainability is also a key concern, especially in relation to the environmental impacts of generative AI (Zewe, 2025). The amount of compute required for the training and deployment of increasingly larger models take up massive amounts of energy and water resources, sometimes to the detriment of local populations who rely on these supplies for their daily needs. Companies developing these large models do not in turn openly share data on energy consumption, thus hindering transparency and accountability of their environmental footprint. In addition, the mining of critical resources, including rare earth minerals that are essential for components such as batteries and chips that power AI systems and other AI driven technologies such as self-driving cars, are leading to exploitation and human rights abuses in different parts of the world (Ren & Wierman, 2024). At the same time, proponents assert that the pursuit of AI development is essential as AI can help to solve the climate change and other environmental issues through better modelling and predictions 'across the entire spectrum of industries and social activities' (Recital 4, AI Act). This leads us to a chicken and egg problem – should we trust and rely upon the (future) potential of AI to combat climate change while sacrificing short term environment sustainability or is this a false promise? In any case, increased transparency is essential in order to monitor the environmental impacts of AI (Shrishak & Warso, 2024). Such forms of transparency is also part of the reporting requirements for GPAI providers with systemic risk in the AI Act.

b. Ethical values, principles and challenges for selected use cases

Table 1. Selection of ethical principles for the European partners' use cases

#	Sphere of influence	Use case	AI research areas	Selected ethical principles
1	Change in human expertise and professional behaviour	Medical doctors (radiologists, surgeons) using AI tools in diagnostics and treatment	Decision-support systems; Image recognition	1. Non-maleficence 2. Accountability and responsibility 3. Transparency and explainability
2		Safety engineers using AI tools to speed up software release approvals	Decision-support systems; General-purpose AI (GPAI)	1. Robustness, safety and reliability 2. Oversight and autonomy 3. Risk of over-reliance and deskilling
3		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 de-skilling
4		Security professionals using AI tools to detect hate speech	Decision-support systems; General-purpose AI (GPAI)	1. Freedom of expression and non-censorship 2. Non-bias, fairness and non-discrimination 3. Accountability and responsibility

5	Change in human cognition and private behaviour	AI systems as individual and family-level virtual assistants	Conversational general-purpose AI (GPAI); Emotional AI	1. Non-manipulation and non-maleficence 2. Privacy, consent and data protection
6		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

Use case 1: Medical doctors (radiologists, surgeons) using AI in diagnostics and treatment

#	Sphere of influence	Use case	AI research areas	Selected ethical principles
1	Change in human expertise and professional behaviour	Medical doctors (radiologists, surgeons) using AI tools in diagnostics and treatment	Decision-support systems; Image recognition	1. Non-maleficence 2. Accountability and responsibility 3. Explainability and transparency

For this medical use case, an initial selection of the following ethical principles was made: **explainability; bias and fairness;** and **accountability and responsibility**.

Explainability is essential for engaging medical professionals' expertise and oversight; understanding the rationale behind an AI system's recommendation enables clinicians to critically assess and appropriately integrate AI outputs into their decision-making (Högberg et al., 2024). **Bias and fairness** address the risks of unequal health outcomes due to overrepresentation of certain populations in the training data and underrepresentation or lack of representation of minorities (Obermeyer et al., 2019). These are gaps that can exacerbate existing disparities in access to care and diagnostic accuracy. **Accountability and responsibility** are central to maintaining trust: when errors or harms occur, clear attribution of responsibility, whether to the tool, the healthcare provider, or the institution, is necessary for redress (Najjar, 2023).

The selection was informed by foundational principles in bioethics: beneficence, non-maleficence, autonomy, and justice (Varkey, 2020), aligning with our criterion for scientific soundness. In our selection, we also focused on the impact on professional expertise, comprehensiveness and procedural soundness. **Explainability** closely supports autonomy by enabling clinicians to make informed decisions and to engage in oversight, aligning with the professional obligation to act in patients' best interests. **Explainability and transparency** have been identified as key elements to assess trustworthiness within the field of radiology (Högberg et al., 2024). As such, this principle was selected over autonomy for its higher relevance for the professional and to avoid repetitiveness (and ensure comprehensiveness). Explainability is also one of the principles for use-case in recruitment, which allows for cross-case comparison: how do the definition and operationalisation of explainability differ across these two fields?

Bias and fairness in the medical context relate to the pursuit of representational and access equality and justice, the latter entailing equitable treatment in certain justified circumstances. The existence of bias within AI systems can further perpetrate social injustice and cement societal inequalities going forward. In some cases, biased outcomes can result in discrimination towards protected groups and other forms of tangible harm (Obermeyer et. al., 2019). This principle was also chosen on account of its procedural and scientific soundness, namely the fact that scholarly research has both highlighted its importance but also that theory is being translated into practice through the exploration and testing of various technical solutions. Similarly, while **accountability and responsibility** reflect justice in their demand for assignable liability, they were prioritized for their practical importance in determining how responsibility is shared or assigned in complex, high stakes and multi-agent settings such as hospitals.

The responses in the internal survey with our AIOLIA partners resulted in alignment on **accountability and responsibility** (top 1) and **transparency and explainability** (top 2). The principle of **do no harm (non-maleficence)** was preferred (same number of responses as for explainability) over **bias and fairness.**

The survey was also sent out to externals – namely the Scientific Advisory Board and networks through AIOLIA partners, EUREC, ADRA and ERCIM. The survey results selected **non-maleficence** as top 1 (21 responses), **accountability and responsibility** as top 2 (20) and both **oversight** and **transparency and explainability** as top 3 (tied at 19 each). As additional considerations, one respondent adds beneficence, namely to "maximise patient benefit" while another respondent added humility/modesty as "admitting uncertainty, not projecting false confidence", which can be linked to the principle of transparency—transparency about the limitations of the technology.

Both surveys underline the importance of non-maleficence and both downgraded the principle of bias. The second survey surfaced the value of oversight. We incorporated these findings into our final selection, by dropping bias, including non-maleficence, and adding oversight as a fourth optional choice. We adopted the ordering from the results in the external survey. The selection is as follows:

1. Non-maleficence;
2. Accountability and responsibility;
3. Explainability and transparency;
4. [shortlisted] Oversight and autonomy

Use case 2: Safety engineers using AI tools to speed up software release approvals

#	Sphere of influence	Use case	AI research areas	Selected ethical principles
2	Change in human expertise and professional behaviour	Safety engineers using AI tools to speed up software release approvals	Decision-support systems; General-purpose AI (GPAI)	1. Robustness, safety and reliability 2. Oversight and autonomy 3. Risk of over-reliance and deskilling

CEPS initially selected two ethical principles as being the most relevant to the context of AI assistance in software release approvals for automotive safety: **robustness, reliability and safety;** and the **risk of over-reliance and de-skilling**.

Robustness, reliability and safety are critical in this context due to the severe consequences of software malfunction, whether through cybersecurity vulnerabilities (Kidmose, 2025), unintended interactions in complex systems, or road safety affecting human life and limb. These elements all underscore the need for rigorous quality control and resilience in design, particularly where software interacts with other subsystems in interconnected environments. Reliability and safety, as in other cases, is a self-referential challenge—we require AI to be reliable and at the same time, we use AI as a tool for system reliability and safety, such as for fault detection and diagnosis, anomaly detection, and risk assessment of engineering-related systems across the industry (Tamascelli et al., 2024).

The **risk of over-reliance and de-skilling** highlights the change in professional cognitive behaviour: automation bias may lead professionals to overly trust the recommendations of the AI system. At the same time, over-reliance on such systems may result in skill attrition due to the reduced exercise of professional judgement. This can diminish essential skillsets in scenario development and planning, and critical risk anticipation (Pulignano & Doellgast, 2020). This principle also relates to the principle of human oversight as a certain level of professional competence needs to be exercised and maintained over time in order to ensure that human oversight can remain effective.

For this use case, we picked two rather than three principles because we thought these were distinctively fit for the use case, while others, such as explainability or responsibility, are present in other use cases. This ensures we avoid over-repetitiveness, as per the comprehensiveness criterion. The principle of **robustness** aligns closely with procedural soundness criterion, which mandates safety, cybersecurity and risk mitigation as core requirements. The inclusion of **de-skilling and over-reliance** responds to our criterion of impact on professional behaviour as well as ethical soundness in the sense of future proofing. Over-trusting AI outputs can lead to gradual erosion of human expertise in safety-critical tasks and undermine long-term effectiveness of human oversight.

The responses from the AIOLIA internal survey surfaced three key values. Similar to our initial selection, the first chosen ethical principle is the need to ensure **robustness, safety and reliability** (14 responses) of the AI system. The second chosen value was **accountability and responsibility** (12) and the third one was the possible **risk of over-reliance and deskilling** (10). The external survey results were as follows: **robustness, safety and reliability** as top 1 (33), **human oversight** (24) as top 2, and **risk of over-reliance and deskilling** as top 3 (19), closely followed by **accountability and responsibility** (18).

Since accountability and responsibility was not in our selection but was ahead of our second most relevant value, we added it as an optional value it in the final selection, making the total principles being 3, uniform to the other use cases.

Due of the high score of oversight and upon consultation with the consortium lead CEA, we chose to include it in our final selection. The ethical principle of accountability was downgraded to an optional consideration. The final selection is as follows:

1. Robustness, safety and reliability
2. Human oversight
3. Risk of over-reliance and deskilling
4. [optional] Accountability and responsibility

Use case 3: Recruiters using AI tools in hiring processes

#	Sphere of influence	Use case	AI research areas	Selected ethical principles
3	Change in human expertise and professional behaviour	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 de-skilling

In the recruitment context, CEPS initially identified three key ethical principles: **bias, fairness and non-discrimination**; **transparency and explainability**; and **oversight and the risk of over- and under-reliance**. The principle of **bias, fairness and non-discrimination** is particularly salient given the risk of AI systems reproducing or amplifying social inequalities embedded in training data. Although non-discrimination standards are well-established in law, the ideals and ambitions of fairness, remain contested and may vary across sectors or jurisdictions (Rigotti & Fosch-Villaronga, 2024). While designers of AI systems wish to avoid bias, deciding when a system is considered fair enough remains a challenge as these thresholds are conscious choices that must be justified. Thus, the operationalisation of fairness, at a scale afforded by AI, also remains contested.

Human oversight is critical in addressing **risks of over-reliance**: over-trusting AI systems can result in harmful impacts such as inaccuracies, errors, and can inadvertently perpetuate discrimination, for example via the use of proxies or inferences that correlate with protected characteristics (Hunkenschroer & Luetge, 2022). At the same time, there is a risk of under-utilising AI and missing opportunities to improve objectivity and equal opportunity; AI has been described as a potential equaliser in hiring processes, if implemented thoughtfully (Hunkenschroer & Kriebitz, 2022). **Transparency and explainability** are essential not only for enabling meaningful review of automated recommendations but also to meet legal requirements under non-discrimination law and ensure public-facing accountability.

These principles were selected for their strong grounding in ethical theory as well as the comprehensive literature on the ethics of AI in recruitment, fulfilling the criterion of scientific soundness. **Non-bias, non-discrimination and fairness** are supported by the justice perspective (Hagendorff, 2020; Mori et al., 2024), particularly procedural justice, which emphasises fairness in decision-making processes. **Transparency and explainability** reflect both the justice perspective, specifically informational justice, and a rights-based approach (Hagendorff, 2020; Chen, 2023; Mori et al., 2024; Hunkenschroer & Kriebitz, 2022), centred on individuals' rights to understand and contest decisions that affect them. **Over-reliance and the need for human oversight** are a risk related to the utilitarian perspective (Hagendorff, 2020) that AI offers efficiency gains. Unchecked reliance can not only undermine the efficiency gains but also lead to harmful outcomes.

Apart from scientific soundness, we put emphasis on procedural soundness, relevance for professionals, and ethical soundness in the form of future-proof values. Non-bias, fairness and non-discrimination was prioritised due to its high procedural soundness, including the availability of many proposed technical solutions, even if still imperfect, and its salience in the legal and regulatory frameworks (Rigotti & Fosch-Villaronga, 2024). The latter is an important factor for transparency as well. The impact on professional behaviour is evident in the risk of over-reliance and the need for oversight. Finally, there is potential for comparability between hiring and hate speech detection, where we selected non-bias and fairness: how does operationalisation differ, e.g. with respect to group-based metrics and individualised treatment?

The responses from the AIOLIA partner survey showed alignment in two of the initially selected principles: **non-bias, fairness and non-discrimination** as top 1 (16); and **transparency and explainability** as top 2 (13). Interestingly, very few respondents chose over-reliance. The top 3 selected principle was **consent, privacy and data protection** (8). The external survey surfaced the same principles and in the same order.

A reason why the surveys did not surface oversight and over-reliance is the fact that they were presented as separate options across the use cases, and in the specific instance of recruitment, over-reliance was the only available option. Having also consulted with the consortium lead CEA, we decided to keep over-reliance in the final selection, due to its high relevance for professional expertise and behaviour, and have oversight as a separate, optional value. We added privacy as a separate optional value due to the high number of responses on privacy in the surveys. The final selection is as follows:

1. Non-bias, fairness and non-discrimination
1. Explainability and transparency
2. Over-reliance and de-skilling
3. [shortlisted] Oversight
4. [shortlisted] Privacy

Use case 4: Security professionals using AI tools to detect hate speech

#	Sphere of influence	Use case	AI research areas	Selected ethical principles
4	Change in human expertise and professional behaviour	Security professionals using AI tools to detect hate speech	Decision-support systems; General-purpose AI (GPAI)	1. Freedom of expression and non-censorship 2. Non-bias, fairness and non-discrimination 3. Accountability and responsibility

Hate speech is considered as 'any expression of discriminatory hate towards people' and can consist of both lawful and unlawful speech (Article 19, 2020), hence presenting a key challenge for detection, possible removal and accountability, including legal accountability for such speech. Hate speech is one of the most prominent examples of online harms—defined as illegal or unacceptable activity that can put a person or a group of people at risk, often with consequences for the offline world (Cortiz & Zubiaga, 2021). While recent focus has been on the duties of private platforms such as Facebook to tackle hate speech, states are the primary duty bearers when it comes to protecting the human right to freedom of expression and information (Pentney & McGonagle, 2020). This can take the form of possible positive measures such as requiring companies to ensure a hate-speech free environment and in pursuing justice and accountability for hate speech. This use case, involving law enforcement, is an example of how AI can aid them in doing so.

The most relevant ethical concern in relation to the use case that we have identified is about **freedom of expression and censorship**. An overarching concern is that of possible over-capture due to automatic removal of flagged content—examples include the removal of 'counter-speech', political speech and satire, and documentation of human rights violations (Pentney & McGonagle, 2020). In this use case, every flagged content is presented to a law enforcement officer who gets to decide whether to escalate the expression for further action or otherwise. However, false positives are not unavoidable due to automation. At the same time, given that there is no consensus on the definition of hate speech (Cortiz & Zubiaga, 2021), the risk of the law officer approving a false positive remains a risk. Additionally, this oversight opens doors to intentional misuse and abuse, including removal of content critical to the state or content of political opponents (Pentney & McGonagle, 2020).

Non-bias, fairness and non-discrimination are another group of principles critical to the use of AI for detecting hate speech. Language understanding is critical for the efficacy of an automated detection system. A risk of bias can be present where hateful content in non-majority languages is less likely to be detected, and even if the content is in a majority language such as English, nuances, contextualised and localised uses of the language may mean that content is either over or under-detected (Udupa et al., 2023). Other bias-related risks include the disproportionate removal of minority groups' content, the prioritisation of certain cultures or groups at the expense of others (Pentney & McGonagle, 2020), the insensitivity of data annotators to certain cultural nuances leading to incorrect labelling (Cortiz & Zubiaga, 2021), as well as the inherent bias in foundation models, which is propagated downstream in hate speech detection applications (ibid). Finally, there is a normative aspect for consideration in relation to detection as well: what counts as hate speech? For example, there is evidence that racist and homophobic tweets are more likely to be categorised by human moderators as hate speech than sexist ones (Pentney & McGonagle, 2020).

The third principle in our shortlist is **accountability and responsibility**. It is concerned with the crucial role of human oversight in the process of hate speech detection by law enforcement agencies, and the ability to justify and audit the decision-making process. An important problem is the inability for the professional to catch false negatives: there should be a clear differentiation of responsibility between them and the AI developers/providers.

The internal survey aligned with these principles, while the external survey differed in the third choice: it was **transparency and explainability** (17), closely followed by **privacy** (16). **Accountability and responsibility** was downgraded to 5th place (17).

In our final selection we kept our initial prioritisation of accountability and responsibility over explainability. While both are relevant, explainability is present in two other use cases in the professional sphere. We chose to focus on accountability to avoid repetitiveness, and present transparency and explainability as optional. Additionally, accountability is one of the principles for the medical profession, which is a very different context and will make an interesting comparison. The final selection is as follows:

1. Freedom of expression and non-censorship
2. Non-bias, fairness and non-discrimination
3. Accountability and responsibility
4. [shortlisted] Transparency and explainability

Use case 5: AI systems as personal and family virtual assistants

#	Sphere of influence	Use case	AI research areas	Selected ethical principles
5	Change in human cognition and private behaviour	AI systems as individual and family-level virtual assistants	Conversational general-purpose AI (GPAI); Emotional AI	1. Non-manipulation and non-maleficence 2. Privacy, consent and data protection

In the context of AI personal assistants, CEPS initially selected the following ethical principles: **non-maleficence; consent, privacy and data protection**; and **autonomy and non-manipulation**. **Non-maleficence** addresses a broad range of potential harms: from direct effects, such as users receiving inaccurate, harmful advice, and showing unhealthy emotional dependency (Dewitte, 2024; Boine, 2023; Coghlan et al., 2023) or social isolation (Ciriello et al., 2024), to indirect effects such as the erosion of social relationships or the reinforcement of social biases (Dewitte, 2024; Gao, 2024). The elderly and children can be especially vulnerable to harmful impacts (Portacolone et al., 2020; Australian Communications and Media Authority, 2025).

Privacy, data protection and consent are critical, given the constant collection of sensitive user input which raises risks of data breaches and profiling, and questions over meaningful user control over their data (Hasal et al., 2021; Murtarelli et al., 2021; Gumusel, 2025). This is especially pertinent seeing that private companies are processing and possessing user data gleaned from very private conversations and environment. Privacy is also especially interesting in the planned scenario with the shared family environment: the assistant must be able to handle sensitive information shared by individual users, while effectively supporting intra-family dynamics.

The principle of **non-manipulation** concerns the exploitative potential of chatbots that have anthropomorphic design and use personalisation to influence user behaviour, especially under business models driven by engagement metrics (Ciriello et al., 2024; Murtarelli et al., 2021). Additionally, the risk speaks to unintentional influence, bearing in mind that steering a social AI system is non-trivial, partially because the training data can contain dubious interactions and partially because what constitutes ethical social behaviour is highly contextual and subjective.

These principles were chosen according to our methodological criteria, with a specific focus on procedural soundness for **non-maleficence** and **non-manipulation**. Implementation remains procedurally challenging across technical feasibility, e.g. there are no robust technical solutions for preventing hallucinations (Barros, 2025) or in designing for well-being, as the latter is individual and context dependent. In legal terms, chatbots are in a grey regulatory area: they can have an effect on physical and mental health, yet they are not medical devices (Corfmat, 2025). Organisational readiness also raises concerns because economic incentives may conflict with ethical use. Chatbots might incentivise staying on the platform for longer, even in problematic situations where the user would benefit from being redirected to a mental health service (Ciriello et al., 2024). In the context of ethical soundness, non-maleficence and non-manipulation involve a tension: it is non-trivial to identify how much influence, which can include persuasion or manipulation, is reasonable (Bakir et al., 2024), given that some forms of manipulative behaviours are tolerated in human-human interactions.

Other considerations about the inclusion criteria include choosing non-maleficence over non-bias and non-discrimination because it is an overarching concern surfaced through the literature review. We chose to do so as well as to ensure comprehensiveness, given that non-bias has been selected in a few of the other use cases. From the standpoint of opportunities for comparability, non-maleficence is also selected for the medical use case: it will be interesting to contrast the operationalisation of the two, with this particular use case centring on emotional and behavioural risks more than physical harm (Dewitte, 2024; Gao, 2024). Finally, the selected principles score highly on normative relevance and align with the future-proof element expressed in EU digital regulation.

The AIOLIA partners' responses in the internal survey coincided with all three values. The principle that received the most votes was autonomy and non-manipulation, the second most selected was non-maleficence, and the third was consent, privacy and data protection. Similarly, the external survey sent to SAB and external networks chose the same top 3 principles and values. The initial selection of values included: autonomy and non-manipulation, non-maleficience and consent, privacy and data protection. However, upon discussion with the academic and industrial partners, it was felt that both non-manipulation and non-maleficience could be grouped together, as respect for autonomy is its underlying purpose. At the same time, the partners also welcomed the inclusion of consent, privacy and data protection.

The final selection is as follows:

1. Non-manipulation and non-maleficence
2. Consent, privacy and data protection

Use case 6: Deepfake AI-based psychotherapy for processing trauma and grief

#	Sphere of influence	Use case	AI research areas	Selected ethical principles
6	Change in human cognition and private behaviour	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 (over-attachment and dependency)

The academic partner for this use case, AUMC, identified numerous challenges posed by deepfake therapy. Amongst others, it can risk impacting the boundaries between reality and simulation and thus negatively impact the psychological safety of the patient. In particular, this form of therapy can introduce risks of re-traumatization, either through the therapy itself or through the process of collecting data (images, videos, audio recordings) to create the deepfake.

In light of this, in the context of using deepfake technologies as a therapy tool for overcoming trauma and grief, CEPS initially selected **non-maleficence; consent, privacy, and data protection;** and **transparency.**

We assign risks grouped around emotional and behavioural harm to the principle of **non-maleficence.** These include emotional harms in the form of unhealthy emotional attachment and emotional distress (Harbinja et al., 2023), arising from the process of data collection or during therapy and cognitive effects such as confusion between reality and simulation and the distortion of authentic memory and relationship with the deceased (Hoek et al., 2024). It is key to underline that deepfake use in therapy is currently only carried out in research and development settings, and that its effectiveness, benefits and risks are not yet clear. This makes the operational feasibility criterion challenging—foundational research will be needed to understand how to make this AI use non-harmful, including whether to apply the technology at all.

The second key value we have identified is around **consent, data protection and privacy**. This relates differently to the personification of the deceased or perpetrators, in cases involving trauma. In the case of the former, the concern is related to the principle of dignity rather than privacy. It is about the potential damage to the integrity of the deceased by instrumentalising them for grief management (Kraaijeveld et al., 2024). Having said that, privacy is still a relevant consideration to be taken into account. Recital 27 in GDPR leaves it open to individual member states to provide post-mortem privacy protection if they so wish. Some countries, such as Italy, have delegated GDPR rights over the deceased's data to their heirs (Harbinja et al., 2023). Other countries, such as the Netherlands, protect "portrait rights" which may also grant limited postmortem protection rights to the relatives of the deceased (Hoek et al., 2024). The operationalisation of this principle can engage with these different legal interpretations.

In the case of trauma therapy, the principle is concerned with the trade-off between the privacy of the perpetrator and the expected benefit for the patient (ibid). While the act of creating a deepfake of the perpetrator may be justified by the patient's legitimate interest in the therapy, one must take balance these against the privacy and data protection interests of the perpetrator as well.

Transparency refers to 1) the disclosure and explanation of the use of the deepfake technology, such as by making the patient aware that the image or avatar is an imitation/simulation, and 2) clearly marking the deepfakes as such, such as with a watermark. Norm (1) is concerned with minimising the risk of confusion in the patient by providing sufficient information for a safe and beneficial interaction. Norm (2) relates to the mitigation of risks associated with data breaches, i.e. leaking of the deepfakes outside of the clinical context, which may damage the integrity of the deceased in the grief scenario and the reputation of the impersonated perpetrator in the trauma scenario. Disclosure is now expressly required in the AI Act (Article 50). This principle similarly relates to the potential risk of confusion between reality and simulation.

All three selected principles are relevant to the effects of the technology on cognition and behaviour, to various degrees. Some are more operationally feasible than others: there are existing mechanisms for ensuring privacy-by-design, while it is less clear how harm can be avoided, notably in the context of gravely sensitive scenarios such as grieving the dead or engaging with traumatic experiences. These concerns are compounded by the very early stage of the research in this area. At the same time, selecting non-maleficence as a principle ensures scientific and ethical soundness, given that the therapeutic benefit of the technology is of central importance to this use case.

Since this use case was included in the project at a later stage, we did not include it in the internal survey. The external survey aligned in the top 1 value being non-maleficence (25) but it uncovered **autonomy and non-manipulation** as top 2 (22) and **over-reliance** as top 3 (18), closely followed by our chosen value of **privacy** (17).

It is interesting that over-reliance, which we frame as a professional risk, surfaces as relevant for a use case that focuses on AI's impact on the person, i.e. the therapy patient. An explanation might be that this distinction was not considered by the respondents. Alternatively, over-reliance might have been perceived as the state of being unduly reliant on this particular method over other technology with established effectiveness, or traditional therapy. This links to the principle of subsidiarity—a criterion for "good care" in bioethics–which requires that the therapy involves the least intrusive alternative (Hoek et al., 2024). Finally, over-reliance might have been understood as a patient's state of attachment and possible risk of addiction and over-dependency to the deep fake. We presented these interpretations to the academic partner and upon consultation, a joint decision was made to consider over-reliance from the personal, i.e. the therapy patient's, perspective. We thus find that the principle of over-reliance as possible over-attachment and unhealthy dependency as relevant for this use case, and we have therefore included over-reliance with this meaning in mind in the final selection. Privacy is kept as an optional principle.

The final selection is as follows:

1. Non-maleficence
2. Autonomy and non-manipulation
3. Risk of over-reliance (over-attachment and dependency)
4. [shortlisted] Privacy, consent and data protection

c. Ethical values, principles and challenges for selected research areas for AIOLIA European partners

In addition to principles and values for the use cases, ethical values and principles are also chosen for the research areas. However, the choices made are not arbitrary. Instead, the principles chosen for the research areas are lifted from the commonalities identified across the use cases. While the academic and industrial partners will operationalise the chosen principles in the next work package (WP3) of the project, CEPS is tasked to design non-technical context enriched ethical guidelines for the research areas. Principles chosen here are both informed by the context in which the research area is deployed in and the sphere of influence in which the research area impacts upon (see table below).

Table 2. Selection of ethical principles for each AI research area in the European partners' use cases

AI research area	Sphere of influence	Use cases	Selected ethical principles
General-purpose AI (GPAI)	Change in human expertise and professional behaviour	2: Safety engineers using AI tools to speed up software release approvals 4: Security professionals using AI tools to detect hate speech	1. Oversight and autonomy 2. Labour impacts
	Change in human cognition and private behaviour	5: AI systems as personal and family virtual assistants 6: Deepfake AI-based psychotherapy for processing trauma and grief	1. Non-manipulation and autonomy 2. Societal well-being and environmental sustainability
Emotional AI	Change in human cognition and private behaviour	5: AI systems as personal and family virtual assistants 6: Deepfake AI-based psychotherapy for processing trauma and grief	1. Human dignity 2. Non-manipulation and autonomy 3. Reliability and accuracy

Decision-support	Change in human expertise and professional behaviour	1: Medical doctors (radiologists, surgeons) using AI in diagnostics and treatment 2: Safety engineers using AI tools to speed up software release approvals 3: Recruiters using AI tools in hiring processes 4: Security professionals using AI tools to detect hate speech	1. Non-bias, fairness and non-discrimination 2. Over-reliance and de-skilling 3. Explainability and transparency
Image recognition	Change in human expertise and professional behaviour	1: Medical doctors (radiologists, surgeons) using AI in diagnostics and treatment	1. Robustness, safety and reliability 2. Transparency and explainability 3. Privacy and data protection

While an overview table is provided above, the following sections will reproduce the tables relevant for the specific research areas, followed by an analysis of the chosen principles and values for each.

General-purpose AI

AI research area	Sphere of influence	Use cases	Selected ethical principles
General-purpose AI (GPAI)	Change in human expertise and professional behaviour	2: Safety engineers using AI tools to speed up software release approvals 4: Security professionals using AI tools to detect hate speech	1. Oversight and autonomy 2. Labour impacts
	Change in human cognition and private behaviour	5: AI systems as personal and family virtual assistants 6: Deepfake AI-based psychotherapy for processing trauma and grief	1. Non-manipulation and autonomy 2. Societal well-being and environmental sustainability

General-purpose AI is used in four use cases by EU partners, two in the professional sphere and two in the private sphere. Our initial selection, not specified to applicable spheres, included the principles of **non-bias, fairness and non-discrimination; labour impacts;** and **societal well-being**.

Bias is one of the most discussed ethical concerns in the development and deployment of generative AI systems (Hagendorff, 2024) and it is selected as an ethical principle in two of the use cases listed above. Similarly to other types of AI systems, it refers to immediate effects such as stereotyping, marginalisation, and discriminatory outputs that can emerge from models trained on imbalanced or ideologically skewed datasets—manifesting in forms such as racism, sexism, and cultural erasure. Distinctive to GPAI are long-term and large-scale effects, due to the wide reach of GPAI and the nature of the value chain. Bias can lead to epistemic injustice (Helm et al., 2024) which includes not only representational harm but epistemic inequities (Kay et al., 2024)—the normative imposition of the cultures and values outside of the regions where the AI systems are developed, especially in multilingual contexts. Other related concerns are the monopolization or centralization of power in large AI labs (Hagendorff, 2024), and LLM-solutionism, specifically favouring LLMs for their versatility with no appropriate account of some tasks' requirements for precise and unbiased analysis (Hajikhani & Cole, 2024).

Labour impacts encompass the range of economic and professional consequences brought about by the integration of generative AI into workplaces. These include not only job displacement and automation but also large-scale deskilling (Hagendorff, 2024) and shifts in role specifications, including increased demand and complexity in positions that involve human-AI collaboration (Chen et al., 2025). Generative AI is bringing about a transformation of creative and knowledge-based work (Hagendorff, 2024), recognising both the threat to these skills and capacities but also the potential for new forms of human-machine collaboration.

Societal impacts refer to the broader, often systemic effects of generative AI on collective safety, public discourse, and human values. These include the spread of harmful content and disinformation, the risks posed by human-level or superhuman models, and alignment failures where AI systems behave unpredictably or against human interests (Hagendorff, 2024). This principle also relates to the values of **non-maleficence** and **safety and robustness**. From a scientific soundness standpoint, societal impacts bring together both technical challenges (how to ensure the systems act in our interest) and normative complexity (whose values should AI align to, how they are chosen, and which values to choose in case of conflict). It is also vital under the future-proofing criterion: we want to ensure preparedness for the long-term risks of more capable systems, that are ever more integrated into society, especially within critical infrastructure.

> Reflection by a survey respondent:
>
> "Due to the magnitude of potential impact General Purpose AI could have, it seems a priority to ensure it is used for the benefit of humanity, addressing societal well-being, non-manipulation and labour impacts first. The other values seem important, but of lower immediate impact."

What we consider as societal impacts is also informed by the definition of "systemic risk" in the EU AI Act and the examples of systemic risk in the third (and penultimate) draft of the GPAI Codes of Practice (European Commission, 2024). Companies developing GPAI models with systemic risk are obliged to have a risk mitigation framework according to which they identify, assess, and mitigate such risks. The EU AI Act defines systemic risk as being specific to the currently most advanced capabilities. As per the third draft of the GPAI Codes of Practice, GPAI model providers are required to include the following risks in their risk mitigation framework: cyber-offence; chemical, biological, radiological and nuclear risks (CBRN); harmful manipulation; and loss of control. Not all of these concerns are directly relevant to AIOLIA. Risks to public health, safety, or public security and risks to fundamental rights are part of the additional list for consideration.

In our internal survey, we asked our partners about the most relevant principles for GPAI which resulted in the autonomy and **non-manipulation** as top choice, **accountability and responsibility** as the second choice, and an equal number of votes for **privacy, labour impacts**, and **societal well-being**.

The external survey had a different design: we asked for principles belonging to GPAI in relation to each case with GPAI, to investigate how these may apply according to the context. In all use cases with the exception of hate speech, there is an overlap of at least 2 of the 3 selected principles for question 1) the use case regardless of the research area, and question 2) the use case in the context of using GPAI. When it came to the hate speech use case, the single overarching principle chosen was **non-bias, fairness and non-discrimination**, while the other chosen values in the context of GPAI were **autonomy and oversight**, and **societal well-being**—prioritised over free speech and transparency.

The aggregate selections revealed highest preference for **oversight** for the professional sphere, and highest preference for **non-manipulation** in the personal sphere. These were followed by equal number of votes for **privacy**, **bias**, and **non-maleficence**. **Labour impacts** and **societal well-being** came in after these.

For our final selection, we decided to separate the principles between the professional and private sphere, selecting two for each. We respected the high scores leaning towards respect for autonomy, namely **oversight** and **non-manipulation**, and included them as the top principle in each respective sphere. These principles are not only important in aggregate, but they are amongst the top 3 voted values for each GPAI-specific question per use case. Furthermore, for the professional sphere, oversight can be linked to principles that were chosen to be operationalised in their respective use cases – namely: over-reliance and deskilling for the safety engineering use case (use case 2) and accountability and responsibility for the hate speech detection use case (use case 4).

As for the second principle, we gave precedence to **labour impacts** for the professional sphere and **societal well-being** for the personal sphere, since they distinctively address the wide reach and advanced capabilities of general-purpose AI, that are unmatched by the other research areas. The inclusion of labour impacts ensures that our comprehensiveness criteria is respected, since they are not selected elsewhere, and it ensures focus on impact on professional behaviour, touching on patterns of resistance, realignment, and resignation among workers. Labour impacts also include the risks of de-skilling and over-reliance.

In the consideration of societal well-being, operational feasibility plays a significant role, with the concept of systemic risk being central for the Codes of Practice for GPAI models. It is worth noting that the Codes of Practice reflect a tight link between the chosen principle of **non-manipulation** with societal well-being: loss of control and harmful manipulation are selected types of systemic risk. At the same time, companies deploying GPAI models are similarly working on red-teaming, transparency reporting, auditing and model evaluation, amongst their existing efforts at promoting AI safety. Legal relevance is another key selection criterion for societal well-being in that ensuring society benefits from GPAI requires regulation and norms, as pointed out by a few respondents: " optimising individual goals of a user within the societal regulatory framework."; "It is necessary to establish social norms to prevent the use of AI from amplifying social conflict or reinforcing the polarization of resources". At the same time, due to the increasing concern over the environmental impacts of GPAI, notably the energy and water demands surrounding the training runs of large models, we also shortlisted this principle to be considered as an overarching principle across both private and professional spheres. These are mainly concerns around actions taken by companies developing large language and other generative AI models but at the same token, it can also be an ethical choice made by partners in AIOLIA when choosing whether or not to use such models or to rely on smaller models that serves a similar purpose.

The final selection of GPAI is as follows:

- Professional sphere:
    - Oversight and autonomy
    - Labour impacts

- Personal sphere:
  - Non-manipulation and autonomy
  - Societal well-being
- Cross-cutting:
  - Environmental sustainability

Emotional AI

AI research area	Sphere of influence	Use cases	Selected ethical principles
Emotional AI	Change in human cognition and private behaviour	5: AI systems as personal and family virtual assistants 6: Deepfake AI-based psychotherapy for processing trauma and grief	1. Human dignity; 2. Non-manipulation and autonomy; 3. Reliability and accuracy

Emotional AI is a research area in both European use cases in the personal sphere of influence: AI systems as personal and family virtual assistants; and deepfake AI-based psychotherapy for processing trauma and grief. Both these use cases also include GPAI. In our initial selection, we kept the principles different from these for GPAI, to take advantage of the opportunity to operationalise a more comprehensive set of values.

Our initial shortlist consisted of the principles of human dignity, **non-manipulation,** and **reliability and accuracy**.

> Reflection by a survey respondent:
>
> "Dignity is paramount in psychological well-being. I see no connection with all other principles as long as the dignity of the user is preserved and respected."

Human dignity is closely linked to the use of emotional data: there is a concern that using emotion recognition leads to the objectification of experiences (Gremsl & Hödl, 2022), which are inherently subjective. The IEEE Standard on emulated empathy (IEEE Std 7014-2024) points out that the first-person declaration of state must always take precedence over the emulated state. The use of emotional data for emotion recognition is prohibited under the AI Act in the context of the workplace and education. More generally, prohibitions in the Act have been justified on several grounds, with respect for human dignity being cited as a key consideration (Commission Guidelines, 2025). Apart from these extreme effects, the feeling of being surveyed can lead to chilling effects and gradual, yet pervasive changes in behaviour (Gremsl & Hödl, 2022).

Non-manipulation is the second principle in our initial selection. Lack of emotional privacy can create conditions that increases the risk of manipulation and exploitation of emotional states, especially in vulnerable individuals. Manipulation can be for political, economic, and marketing purposes (Ghotbi, 2023). The risk is not unique to emotional AI, but it is categorically different: influencing through an affective state could be more powerful, pervasive or subliminal, compared to influencing via reason.

Manipulation, especially as a form of deception, is not considered universally unethical, e.g. some forms of deception are acceptable in a social context, which makes operationalisation challenging

(Bakir et al., 2024). The combination of emotional AI with GPAI in the use cases, given that current GPAI models have been documented to exhibit deception capabilities, makes the concern for harmful manipulation increasingly more relevant (Park et. Al., 2025). Moreover, the principle is chosen because non-manipulation has received significant legal attention. The AI Act prohibits the use of AI for manipulation or deception, if it leads to significant harm (Art. 5.1.a AI Act), and harmful manipulation is one of the selected systemic risks in the third draft of the GPAI Codes of Practice (see Measure 11.3.1 and Appendix 1.1).

Reliability and accuracy is an important concern with respect to emotional AI due to the state of the theoretical foundations of emotion: there is no consensus on the conceptual model to explain what emotions are (Stark & Hoey, 2021; Mohammad, 2022). Most affective computing solutions which involve tracking bio-signals are grounded in Basic Emotion theory (Ekman, 2000) which, some argue, does not lend sufficient scientific evidence describing emotions as causal or constructed phenomena (Barrett, 2017; Stark & Hoey, 2021). Using inferred states to predict behaviour can thus be dangerous. Moreover, emotions are irreducible to data and require social context, which are fundamental barriers to accurately representing emotion (Stark & Hoey, 2021). Due to these reasons, the reliability of emotional AI should be considered categorically different from other deployment contexts (IEEE Std 7014-2024).

Reflection by a survey respondent:

"There is a possibility that errors may occur in the analysis of human emotional states due to the application of overly schematic emotional understanding patterns".

As for the selection elements, we draw attention to the challenges for procedural soundness of reliability and non-manipulation, such as the lack of clarity in methods for determining acceptable risk levels. These challenges are non-trivial but at the same time, they have to be weighed against the other methodological elements: both reliability and non-manipulation are scientifically and ethically sound, they are of high legal relevance, and, importantly, they score highly on the relevance for human cognition and behaviour.

Both surveys coincide in the top 2 principles being human dignity and non-manipulation. With respect to dignity, a respondent shares that "dignity is paramount in psychological well-being. I see no connection with all other principles as long as the dignity of the user is preserved and respected.". With regard to non-manipulation, a respondent feels that "manipulation seems to be the biggest concern here. We are afraid the primary interest into the use would be for mass manipulation. "

On the contrary, the surveys surface privacy rather than reliability as the third concern. While privacy is admittedly a significant consideration in the context of processing sensitive data, we find that it is partially covered by the principle of dignity, under which we discuss surveillance. More importantly, we find that the high level of contention about the reliability of the capabilities to infer emotive states, especially with the objective to understand and steer behaviour, makes the principle of accuracy very distinctive for this research area. This concern has been raised by a few respondents in the optional free-text boxes for giving justification: "emotions are not always readable by tech"; "There is a possibility that errors may occur in the analysis of human emotional states due to the application of overly schematic emotional understanding patterns".

Further, since the operationalisation of the principles belonging to the research areas are not meant to be technical in nature, it will be of value to unravel the problem from the perspective of applied ethics. The methodological element of comprehensiveness also features here. As privacy is a selected value under the image recognition research area, it is prudent to ensure a fair representation of values across the different research areas. Motivated by the reasons given above, we therefore select reliability and accuracy as our third principle for emotional AI, and the final selection is as follows:

1. Human dignity
2. Non-manipulation
3. Reliability and accuracy

AI decision-support systems

AI research area	Sphere of influence	Use cases	Selected ethical principles
Decision-support	Change in human expertise and professional behaviour	1: Medical doctors (radiologists, surgeons) using AI in diagnostics and treatment 2: Safety engineers using AI tools to speed up software release approvals 3: Recruiters using AI tools in hiring processes 4: Security professionals using AI tools to detect hate speech	1. Explainability and transparency 2. Over-reliance and de-skilling 3. Non-bias, fairness and non-discrimination

AI decision-support is a research area in all European use cases within the professional sphere. In this context, the initial selection of ethical principles by CEPS focused on **transparency and explainability**, and **over-reliance and de-skilling.**

Explainability is crucial in this context, since an effective mutual decision-making is grounded in understanding how a given output or recommendation was generated, hence impacting justifiability and trust. The inability to understand reasons behind a recommendation could lead to dismissal, or an uncritical adoption of the output or recommendation, thus leading to the question as to the added value of the system. Explanations are also necessary to ensure compliance with relevant laws, for risk management, and for model validation, amongst others (Guttmann & Ge, 2024).

Related to the problem of interpretable AI outputs is the problem of **over-reliance**. Humans are known to have automation bias—the inclination to trust a machine more than a human, since it is perceived as being more logical, accurate, objective or otherwise error-free. Due to known limitations of AI, humans must remain vigilant and critical of the AI-proposed decisions, to avoid harms and to ensure accountability. **De-skilling** is a possible outcome of over-reliance**:** over-trusting the system can lead to skill attrition and may disengage the professional, over time, from critically exercising their judgement and expertise. The principle also scores highly on the selection element of impact on professional behaviour.

> Reflections from the survey in relation to over-reliance:
>
> "automation complacency and lack of human re-checks are the most problematic here"
> "linked to the possible changes in professional skills and behaviors, which hugely relevant to the project aims"

The internal survey highlighted **explainability and transparency** as top 1 principle (14 responses), followed by **over-reliance** as top 2 (12 responses). **Robustness** (9); **accountability** (8) and **bias** (8) scored very closely.

With regard to over-reliance, notable comments by respondents include that "automation complacency and lack of human re-checks are the most problematic here", and that over-reliance is "linked to the possible changes in professional skills and behaviours, which hugely relevant to the project aims".

The external survey results surfaced **bias** and **robustness, safety and reliability** as more important than **over-reliance and de-skilling** and **transparency and explainability**. In our final selection, we have included **over-reliance and de-skilling**; and **transparency and explainability,** where our initial selection and the survey results converge. We prioritise **non-bias, fairness, and non-discrimination** over **robustness** as the third selected principle, recognising that bias is one of the most highlighted concerns in the literature when it comes to decision-support systems. Notably, bias has received much operational attention, yet continues to remain a persistent problem due to the presence of societal bias. This means that its operationalisation needs to meaningfully articulate what the pursuit of fairness means in contextually informed fields. Operationalising protection against AI bias is thus less about its eradication but more about management and minimisation and a justification of the social policy choice that is made to pursue a particular vision of fairness. As a respondent reflects, " since simple technical accuracy alone cannot guarantee socially just decision-making, efforts are needed to minimize bias".

The selection of principles for AI decision-support is:

1. Transparency and explainability
2. Over-reliance and de-skilling
3. Non-bias, fairness and non-discrimination

Image recognition

AI research area	Sphere of influence	Use cases	Selected ethical principles
Image recognition	Change in human expertise and professional behaviour	1: Medical doctors (radiologists, surgeons) using AI in diagnostics and treatment	1. Robustness, safety and reliability 2. Transparency and explainability 3. Privacy and data protection

Image recognition belongs to a single European use case—medical doctors (radiologists, surgeons) using AI tools in diagnostics and treatment. We have therefore limited the scope of ethical considerations for image recognition to applications in the domain of healthcare.

We have initially identified **robustness, reliability and safety; bias, fairness, non-discrimination;** and **privacy and data protection** as the values to focus on. Robustness, safety, and reliability are important from the point of view of bioethics, specifically the principles of non-maleficence and subsidiarity. To justify the use of AI in the analysis of medical images, the technology must be reliable, and, if not superior, at least on par with human judgement. Similarly, the problem of bias shares a strong linkage to principles from bioethics, namely non-maleficence and justice, since biased AI systems can lead to harmful and discriminatory outcomes for patients. The third principle, privacy, concerns patient privacy and patient consent for the use of their data.

The internal survey resulted in many values having a close number of votes. The second, external survey was more informative: it surfaced **robustness, reliability and safety** as top 1 (24 responses), and had **transparency and explainability** and **non-bias, fairness and non-discrimination** as second most relevant (19 responses each).

> Reflection by a survey respondent:
>
> "For general Computer Vision applications, consent, privacy and data protection and Robustness seem as practical things to address, while Transparency, do no harm and bias need not apply for every use case. Priorities would change if we were to discuss Computer Vision applications for identifying, tracking and classifying people."

The results surface the salience of transparency and explainability and we adjust our selection to reflect this. Transparency and explainability also score highly on the methodological element of relevance for professional expertise: medical professionals must be able to understand the systems in order to effectively use them to benefit the patients, as well as to assess the benefits of using them. As to the other chosen principles, we keep **robustness, reliability and safety** and **privacy,** as in our initial selection**.** The reason to favour privacy over bias, which was in the top 3 principles from the survey, is due to two reasons. First, for coverage purposes, non-bias and fairness are selected for AI decision-support systems while privacy has not been selected in any other research area. Selection of this value will ensure that no values are left out. Secondly, in relation to medical imaging, these consist of health data which are considered as sensitive data. The latter is subjected to more stringent requirements of data protection (Art. 9 GDPR). Moreover, the value of privacy that is coupled together with data protection is also pertinent in that each piece of individual health data can contribute towards health research for the benefit of society. Using data for this purpose needs to balance individual privacy (including patient dignity) and societal and public health benefits (Nass et al., 2009).

The selection of principles for image recognition is:

1. Robustness, safety and reliability;
2. Transparency and explainability;
3. Privacy and data protection

d. International partners' selection of use cases and ethical principles and values

Introduction

International partners in AIOLIA were tasked to select the use cases in which AI would impact human cognition and behaviour, as well as the relevant ethical principles applicable therein to the use cases. Below are a list of use cases and ethical principles and values selected. The content in relation to the use case is taken from the selection list in the report for D2.1 and have only been adjusted for grammatical, stylistic or clarificatory purposes. Similarly, the numbering of the use cases follows the order of AIOLIA D2.1. Table summaries are provided for the selected ethical principles and values for use cases and research areas respectively are provided below, followed by an elaboration of the methodology and values.

Methodology and method for selection of ethical principles and values

While international partners are tasked to select the use cases and associated ethical principles and values themselves, the partners similarly followed a methodology in identifying the most relevant ethical principles that pertains to each use case. Partners carried out a horizon scanning in relation to the possible applicability of relevant domestic legislation, soft law or international guidelines on AI ethics (e.g. UNESCO Recommendations on the Ethics of AI) on their specific use case, alongside various levels of consultations undertaken with partners or relevant stakeholders, such as industry actors. Where the key documents referred to are not in English, translated versions have been provided by the partner to AIOLIA. Detailed methodological steps are elaborated under each use case.

Selection of ethical principles and values for international use cases

Table 3. Selection of ethical principles for the international partners' use cases

#	AI research	Sphere of influence	Use cases	Ethical principles selected by international partner
7	Emotional AI Image recognition Video analysis	Change in human expertise and professional behaviour	Workplaces equipped with AI tools for behavioral analysis	1. Proportionality 2. Fairness and non-discrimination 3. Transparency and explainability
8	GPAI Emotional AI Image recognition Decision-support systems		AI systems for smart elderly care in Wuxi city	1. Privacy and data security 2. Emotional dependency and risk of deception 3. Algorithmic bias 4. Accountability
9		Change in human cognition and private behaviour	AI systems as ‘personal companions’ to assist senior citizens	1. Prevention of psychological manipulation 2. Diminished autonomy 3. Accountability (and the requirement of explainability)
10	GPAI Emotional AI		AI systems as grief-alleviating personal assistants (griefbots, chatbots designed to imitate a deceased loved one)	1. Dignity of the deceased 2. Patient well-being (i.e. protecting the long-term interests of grievers) 3. Multiculturalism

Selection of ethical principles for AI research areas for international partners

The selection of the ethical principles for AI research areas for international partners are the same for those selected for European partners. This uniformity is intentional as these principles relate to research areas – GPAI, image recognition, AI decision support systems and emotional AI, which are common for both international and European partners, even as use cases differ. These principles will form the basis of the creation of non-technical, context informed guidelines in the next phase of the project. Even though the international partners were not initially envisioned as being directly involved as part of the creation of the non-technical guidelines, we believe that the inclusion of international partners brings an important comparative element, notably in seeing whether there are region or culture specific considerations that can inform and enrich the non-technical guidelines.

Table 4. Selection of ethical principles for the AI research areas in the international partners' use cases

AI research area	Sphere of influence	Use cases	Selected ethical principles
General purpose AI (GPAI)	Change in human cognition and private behaviour	8. AI systems for smart elderly care in Wuxi city 9. AI systems as 'personal companions' to assist senior citizens 10. AI systems as grief-alleviating personal assistants (griefbots, chatbots designed to imitate a deceased loved one)	1. Non-manipulation and autonomy 2. Societal well-being and environmental sustainability
Emotional AI	Change in human expertise and professional behaviour	7. Workplaces equipped with AI tools for behavioral analysis	1. Fairness and non-discrimination 2. Transparency and explainability
	Change in human cognition and private behaviour	8. AI systems for smart elderly care in Wuxi city 9. AI systems as 'personal companions' to assist senior citizens 10. AI systems as grief-alleviating personal assistants (griefbots, chatbots designed to imitate a deceased loved one)	1. Human dignity 2. Non-manipulation and autonomy 3. Reliability and accuracy
Decision-support	Change in human expertise and professional behaviour	8. AI systems for smart elderly care in Wuxi city 9. AI systems as 'personal companions' to assist senior citizens	1. Over-reliance and de-skilling; 2. Explainability and transparency

			3. Non-bias, fairness and non-discrimination
Image recognition (video analysis)	Change in human expertise and professional behaviour	7. Workplaces equipped with AI tools for behavioral analysis 8. AI systems for smart elderly care in Wuxi	1. Robustness, safety and reliability 2. Transparency and explainability 3. Privacy and data protection

Use case 7: Workplaces equipped with AI tools for behavioural analysis

The University of Osaka will be analyzing ethical and social aspects of workplace-video–analysis systems. These systems couple fixed cameras and wearable sensors with machine-learning pipelines trained on annotated footage of real production lines. The core computer-vision model classifies each frame into fine-grained task steps, detects anomalies (e.g., slips, tool misuse), and recognises safety infringements such as workers entering exclusion zones or failing to wear helmets. Down-stream modules transform these detections into several application tiers:

- Productivity optimisation: footage is segmented into elemental tasks so that differences between novice and expert technique can be visualised; one company reported cycle-time reductions of up to 99 % after redistributing best practices.
- Automated reporting: An AI system that converts daily camera streams into structured work reports, eliminating manual logs.
- Real-time safety: vision models trigger alarms for missing Personal Protection Equipment or boundary violations and can combine with facial recognition to identify the individual at risk.
- Well-being analytics: by fusing facial cues and heart-rate telemetry, an "emotion engine" flags stress or fatigue episodes that may degrade quality of work.

Methodology and selection of the most relevant ethical principles and values in relation to AI tools for behavioural use cases in workplaces:

The methodology employed for selecting the ethical principles and values by University of Osaka involved identification of these values through prior research conducted between members of the research team and the industrial partners including NEC and RICOH. This led to a project synthesizing recent literature on the ethics of workplace surveillance, as well as one on the ethics of emotion recognition systems (emotional AI). A part of this research had involved a qualitative analysis of recommendations within the academic literature for ethical principles to guide the use of emotion recognition (Katirai, 2023). As the Osaka team will take up a spectrum of technologies related to behavioural analysis, including emotion recognition, this prior research was used as a starting point, and then generalized to other technologies within the use case. The principles were then compared with those in the UNESCO Recommendations on the Ethics of Artificial Intelligence, to refine the selection and framing. The UNESCO Recommendations were used as the primary reference point in shortlisting the ethical principles and guidelines. Although domestic legislation such as Japan's Personal

Information Protection Law were also analysed, the partner did not find any principles in the legislation which were useful or applicable.

The ethical principles and values were chosen based on its importance, as highlighted in the literature on the research area, and through previous reviews and analyses. The multi-year collaboration (mentioned above) by the parties also reinforced the fact that these values were key points of concern and relevance in the Japanese context.

The ethical principles and values chosen by The University of Osaka for this particular use case are:

1. Proportionality
2. Fairness and non-discrimination
3. Transparency and explanability

The principles are elaborated below:

- Proportionality

This principle consists of ensuring that there are sufficient merits when using AI for the purposes of behavioural analysis in the workplace in order to balance against the potential risks or disadvantages that it could pose, especially for workers. In ensuring the operationalisation of the principle of proportionality, there needs to be an awareness of power imbalances which may lead to greater benefits for system implementers (employers) over data subjects (workers). Research has highlighted that there is a (commercially-motivated) tendency to focus on efficiency, cost-cutting, and higher profits (Crawford, 2021; Mantello et al., 2021), potentially sidelining the interests and well-being of workers. Proportionality ensures that the aims and applications for the technology are appropriate to the particular use case, considering a balancing of commercial, safety and worker interests and rights.

- Fairness and non-discrimination

One consideration in this ethical principle is the necessity of avoiding “function creep” – where a system designed for one function is used to inform another function. At the same time, one should avoid incidental findings in biometric work monitoring. This can be due to the fact that some of these AI systems are ‘trained on datasets which are unsuitable for deployment in occupational settings.’ (Awumey et al., 2024; 7). Another issue is possible misuse of sensitive personal data, through for example, gauging the emotions of workers, to their detriment. This principle also ensures that AI systems do not lead to disadvantages for particular groups of workers (e.g., minorities, vulnerable populations) and to prevent disadvantages or harms for workers who opt-out or may reject the systems (Stark et al., 2020; Awumey et al., 2024; Katirai, 2024).

- Transparency and explainability

This principle serves to ensure that data subjects (workers) are adequately informed about the AI system, its use, and its implications. This is especially the case when there is a stark power imbalance between the employer and employee, making it more pertinent for the employee to know how AI tools are being used in the workplace as ”harms are exacerbated by worker uncertainty about how these systems hold them accountable for certain behaviours” (Awumey et al., 2024; 7). A more contextualised explanation in foregrounding transparency and explainability is also relevant in this case – notably due to the strong social value placed on social trust in the Japanese context (McStay, 2021).

Use case 8: AI systems for smart elderly care in Wuxi city

AIOLIA's partner from China, CASTED, selected the use case on the use of AI for smart elderly care. The Da Tou A Liang (大头阿亮) is an elderly care robot, developed by Jiangsu Aiyu Wencheng Elderly Care Robot Co., Ltd. The industrial partners for this project are: Hongdou Group, iFlyTek, DeepSeek, Wuhan University, Shanghai Jiao Tong University, Wuxi Municipal Government, Wuxi Huishan Nursing Home, Aging Care Alliance.

The elderly care robot is an AI-powered companion designed to enhance the quality of life for seniors, particularly those living alone. Launched in 2025, this robot integrates artificial intelligence (AI), Internet of Things (IoT), big data analytics, and sensor technologies to provide comprehensive support in health monitoring, emotional companionship, and daily assistance.

The key features of this AI system include:

- Health Monitoring: Equipped with sensors to track vital signs (heart rate, blood pressure, blood glucose) and detect emergencies, including falls (triggering alerts within 3 seconds).
- Emotional Interaction: Uses natural language processing (NLP) and dual AI engines (iFlyTek and DeepSeek) to engage in conversations, play music, and adapt responses based on emotional cues (e.g., detecting loneliness or distress).
- Daily Assistance: Reminds users to take medication, controls smart home devices (lights, appliances), and schedules activities.
- Remote Connectivity: Allows family members to monitor seniors via video calls and receive real-time health updates through a mobile app.
- Autonomous Navigation: Uses computer vision and obstacle-avoidance algorithms to patrol homes and ensure safety.

The robot operates within Wuxi's smart elderly care ecosystem, which includes cloud platforms, IoT-enabled devices, and partnerships with local care institutions.

Methodology and selection of the most relevant ethical principles and values in relation to AI tools for smart elderly care:

CASTED identified the relevant ethical principles for this particular use case through field research and multi-stakeholder discussions (e.g., developers, enterprises, elderly, and their families). A preliminary 'co-creation' methodology involved governments, academia, enterprises, and care institutions to embed ethical governance rules in technology development. Case studies were also conducted (e.g., Wuxi's "Datoualiang大头阿亮" Robot) which surfaced ethical issues such as data security, privacy protection, and employment impacts on caregivers. It was clarified that while no international ethical guidelines were named in the meetings, the project generally aligns with EU ethical governance standards. The primary sources relied upon were domestic sources, such as the Chinese Ministry of Civil Affairs' elderly care policies, National Data Standardization Committee's data security standards, Governance Principles for the New Generation Artificial Intelligence Developing Responsible Artificial Intelligence and the China AI Ethics and Security White Paper. For example, Wuxi's emergency response services is under a legal obligation to comply with local disability federation regulations.

The process of identifying the most relevant ethical principles and values also included the following stakeholders:

i. Enterprise stakeholder: Jiangsu AIYU Wencheng Elderly Care Robot Co., Ltd., which served to highlight ethical conflicts in product design

ii. Academia: Renmin University, Guangxi University of Science and Technology, which examined theoretical ethical frameworks
iii. Public authorities: Ministry of Civil Affairs, Wuxi Elderly Care Association, to examine policy-practice alignment

In turn, the selection criteria employed to identify the most relevant ethical principles and values included the following:

i. Practical challenges: Data safety due to the fact that the robots are considered as home-use products, and potential caregiver unemployment risks, since these are also institutional products
ii. Policy demands: the National Data Standardization Committee's requirements for industry standards were examined, alongside the Ministry of Civil Affairs' push for tech-supported elderly care
iii. Global Influence: How the consideration of this use case and the selection of the relevant principles can showcase China's contribution to the global ethical discourse and governance frameworks

As a result, the ethical principles and values chosen for this particular use case are:

1. Privacy and data security
2. Emotional dependency and risk of deception
3. Algorithmic Bias
4. Accountability

Each of these are elaborated below:

- Privacy and data security

This principle is pertinent as users of the elderly care robot will likely be monitored continuously as part of fully utilising the functionalities of the robot. These include tracking users for vital signs and tracking facial expressions during interactions. Further, the tracking will include ambient surrounding elements such as home appliances and obstacle detection in the home but also enabling remote tracking through access by family members who wish to receive updates from the robot. These continuous forms of tracking can inhibit the user from expressing or undertaking certain activities, thus limiting the user's autonomy, but also pose increased risk to possible data leaks, including through hacking and misuse of user data. CASTED has proposed to address this risk through encryption, anonymisation and user consent in order to ensure that the design of such elderly care robots is done in an ethical way.

- Emotional dependency and risk of deception

The ethical principle chosen here relates to the possible risk of users forming an unhealthy emotional connection to the elderly care robot, including through perceiving the robot as a human-like companion. The latter can lead to possible over-reliance on the advice or recommendations provided by the robot but also possible risks of deception, when the user is emotionally manipulated to or manipulated to take certain actions.

To this, CASTED has identified that setting proper boundaries for AI-driven interactions with the users are key in order to avoid undermining human relationships. This is based on the assumption that human relationships are still important for its own sake. This may include distinguishing human relationships and interactions with robots, the latter which arguably undertakes (only) a functional role of assisting the user. However, the boundaries between human and robot interactions are blurring, such as through fulfilling a user's emotional needs, and insisting merely on an ontological distinction may not be enough when engaging with this ethical question.

CASTED has identified that the priority for the co-creation phase of the project will be on questions of privacy and emotional dependency, as these ethical concerns will directly impact user trust towards the robot and user safety.

- Algorithmic bias

CASTED identified that there is a risk that the health recommendations provided by the algorithm may end up favouring certain demographic groups, whether advertently or inadvertently, due to biased training data. As mentioned in section 3.a, bias arising from lack of representation within the training data is a key concern. However, it is not the only concern in relation to algorithmic bias. Bias can also arise from proxy data, where a given datapoint indirectly relates to a legally protected characteristic. New forms of bias may also arise that do not relate to characteristics subjected to legal protection as AI is able to discern new characteristics or behaviours that end up unjustly favouring or disfavouring certain groups. CASTED has identified the necessity of auditing datasets in order to ensure inclusivity within AI model training data. In addition, an iterative audit may also be necessary to pick up on any algorithmic bias shown by the system after it has been deployed.

- Accountability

The fourth value identified by CASTED is that of accountability. This is where ambiguity arises as to who is responsible and therein liable, when errors occur, such as when the user is harmed as a result of the AI system missing a fall alert. While questions surrounding liability may be addressed through current or upcoming domestic legislation, the issue of moral responsibility is a separate though concurrent concern that relate to accountability. To this end, CASTED has identified that an important focus is to clarify the role of developers, caregivers and institutions from the very beginning.

Use case 9: AI systems as personal companions assisting senior citizens

STEPI, our academic partner from South Korea, will be partnering with Naver and KIST to examine the design and development of AI systems that can act as personal companions to assist senior citizens. Senior care chatbots are multimodal companion robots aimed at home-based care for senior citizens. The platform delivers entertainment, companionship and ambient monitoring, while simultaneously collecting continuous streams of biometric, behavioural and emotional data. Each device couples far-field microphones and RGB-D cameras with an on-board AI stack. This enables the following capacities:

1. Conversational intelligence: cloud speech-to-text / text-to-speech, intent recognition that draws on an ontology of daily-living concepts, and a long-term memory layer that stores user data so the agent can maintain context over weeks or months.
2. Affective sensing: LLM APIs are used to classify sentiment in spoken utterances, while face- and pose-recognition modules cross-check visual cues for stress, fatigue or immobility.
3. Context-aware dialogue and scheduling: the robot blends object recognition with knowledge of the user's calendar to suggest taking medication, physiotherapy or leisure activities in the user's own dialect and preferred speech style.

4. Autonomous mobility and safety: in the Dasom and Mybom versions of senior care chatbots, a mobile base patrols the home, issues emergency calls if falls are detected, and can locate household items on request.

There are a few similarities between this use case and use case selected by CASTED on the elderly care robot Da Tou A Liang (大头阿亮). Both partners can thus compare and contrast the ethical principles in question in order to unpack the potential but also complexities of such technologies.

Methodology and selection of the most relevant ethical principles and values in relation to AI systems as personal companions assisting senior citizens

STEPI conducted both a literature review and interviews with key stakeholders in selecting the most relevant ethical principles and values pertaining to this use case. The literature review focused on three types of data, namely:

1. The 'Ethics Self-Check List' for AI engineers provided by the Korean Government (and which was based on the research of KISDI), the Robot Ethics Guideline by KIRIA and finally, referring to academic papers on ethical issues of chatbots (covering papers published domestically & in international journals). Where relevant, the EU AI Act and the UNESCO Recommendations on the Ethics of AI were also referenced.
2. Reports by local governments on projects related senior care chatbots in Korea
3. News media articles related to senior care chatbots that were reported in Korean media

In turn, interviews were also conducted to explore the selection of the most relevant ethical principles and values. These interviews included different actors, namely, chatbots developers, AI developers, service providers and bodies which drafted the ethics guidelines. The list is as follows:

1. Chatbot developers: Hyodol, ROAIGEN(KIST), ETRI, Korea University, Dong-A University
2. AI developers: Electronics and Telecommunication Research Institute (ETRI), Portal301
3. AI service provider: SK CareCall
4. Drafters of AI ethics guidelines: Korea Information Society Development Institute (KISDI), Korea Institute for Robot Industry Advancement (KIRIA)

The interviewees were asked to share ethical issues related to the design, development and deployment of the technology and the technical measures technical measures to address the concerns raised. STEPI also raised ethical concerns as identified through the literature review and invited the interviewees to reflect on these topics. Based on these steps, the ethical principles and values eventually chosen for this particular use case are:

1. Prevention of psychological manipulation
2. Diminished autonomy
3. Accountability (and requirement of explainability)

Each principle is elaborated below:

- Prevention of psychological manipulation

Similar to the European use case impacting private behaviour, the child and family AI companion, the ethical principle that is foregrounded relates to manipulation. STEPI has also identified an adjacent concern, namely emotional dependency, which can lead to manipulation. Psychological manipulation concerns the risk of the personal companion, through persuasive language or its human-like speech and expression, potentially influencing, persuading or encouraging the user to undertake an action against her will or to take actions that may endanger her mental or physical well-being. The risk of psychological manipulation can be heightened in this particular use case due to the affective sensing capabilities that is offered part of the personal companion.

- Diminished autonomy

The ethical principle of autonomy recognises the self-governance, rationality and decision-making capacities of individuals, as one who can give reasons and justifications for actions. Diminished autonomy, which STEPI highlighted as one of the key ethical concerns, pertain to the inability to freely exercise autonomy due to potentially widened incursion of the personal companion into the everyday life of the elderly person, impacting rational decision-making skills and behaviours. While a scenario of complete freedom enabling the unrestrained exercise of autonomy is perhaps illusory, as individuals are variously influenced by society, culture and the environment, a more direct impact of autonomy through the use of the personal companion rightly raises a serious ethical concern. Increased trust and reliance towards this companion may displace or negatively impact an individual's own ability to exercise judgement or take actions.

- Accountability (and requirement of explainability)

The ethical principle of accountability, in this case, which STEPI tied to explainability, is also a dominant concern in relation to personal companions. This is due to the fact that, unlike other technologically mediated social relationships, such as within social media, interactions with the personal companion does not have a human behind the screen. While responses and interactions are based on the underlying large language model, these models are usually inscrutable, including in relation to how it generates responses to the user. Furthermore, responses and interactions are not informed by moral norms or social norms of communication, unless these values are expressly included in the model design. These aspects, alongside the inclusion of affective sensing capacities, foreground the need to operationalise the ethical principle of accountability, enabled through explainability of how the technology works and its impacts upon its users.

Use case 10: Generative Ghosts and the Grieving Process

McGill University, together with industrial partners - Canadian Psychological Association, *Association des psychologues du Québec*, Canadian Counselling and Psychotherapy Association, will be examining the design and development of AI chatbots that assist in the grieving process.

These griefbots (sometimes also called deathbots, deadbots, thanabots, or generative ghosts) are AI chatbots designed to simulate the personality and speech of a deceased loved one. They range from basic chatbots interacting with users through text-based interface to more advanced virtual avatars or robots aiming to mimic how a person looked, sounded, and moved.

These chatbots are based on Large Language Models (LLMs). They are also likely to become widely available as it becomes easier to develop personalized chatbots through Retrieval-augmented generation (RAG) systems, which are relatively easy to deploy and cheap. RAG systems allow conversational AI to deliver more precise and contextually relevant responses by drawing on external information sources in addition to a basic model's internal knowledge. In other words, a user or a company could easily use a LLM to draw on the digital data of a recently deceased person – which can reflect a person's speech pattern, sense of humour, preferences, typical activities, etc. – to create a chatbot supposed to behave like they would. These griefbots challenge how we think about death and how we should protect a person's interests and rights over time and can risk impacting the wellbeing of grievers and change how we conceptualize the grieving process.

Methodology and selection of the most relevant ethical principles and values in relation to AI systems as grief-alleviating personal assistants:

The relevant ethical principles were identified on the basis of discussions between ethicists and legal theorists at both the University McGill (through the involvement of Prof. Jocelyn Maclure, Hugo Cossette-Lefebvre (Ph.D.), and Christophe Facal), and the University of Toronto (Prof. Karina Vold). International documents that were considered as part of these discussions include: UN Report on the Protection of the Dead; the EU Artificial Intelligence Act; the Montréal Declaration for Responsible AI Development; and the § 50-f addition to the New York Civil Rights Law entitled 'Right of Publicity' which concerns an individual's inherent right to control the commercial use of one's personal characteristics (Brandon, 2021). In addition, domestic instruments such as different federal and provincial legislation were considered to examine how the Canadian Federal Government and the different Canadian provinces approach griefbots and their regulation. The exercise served to examine and identify urgent gaps that exist within Canadian legislation.

At the federal level, the following documents were considered:

i. Personal Information Protection and Electronic Documents Act (PIPEDA)
ii. Copyright Act
iii. Criminal Code of Canada

At the provincial level, the partner examined Ontario's Succession Law Reform Act.

While structured and formal consultations have not yet begun with non-academic partners in time for this Report, McGill will be consulting with the following non-academic partners to identify the main ethical principles in play. These include, amongst others:

i. MILA (Quebec Artifical Intelligence Institute)
ii. EmoScienS (private company developing emotional AI software)
iii. Canadian Psychological Association
iv. Association des Psychologues du Québec
v. Canadian Counselling and Psychotherapy Association

The relevant ethical principles (see below) were chosen on the basis of a consensus between the main investigators of the project as they are considered the most urgent to address given the lack of existing Canadian legislation and legal tools regulating these aspects of LLMs and chatbots. It would also be interesting to compare the ethical principles and its operationalisation for this use case with the European use case 6 which involves deepfake therapy.

The ethical principles and values chosen for this particular use case are:

1. Dignity of the deceased
2. Patient well-being
3. Multiculturalism

These principles are elaborated in turn:

- Dignity of the deceased

The first ethical principle chosen pertains to the question of how one should engage with the dignity of those who are deceased. This includes considering the interests (imputed, implied or expressed) and rights of the deceased over their data and personal information. While griefbots can assist the living in the process of grieving, this should not come at the expense of those who did not wish for their data, image or voice to be used posthumously in this manner. Even if expressed wishes cannot be ascertained, this principle foregrounds the need to consider the dignity of deceased persons in general, as data can be misused, leaked or hacked, and the eventual griefbot that is designed using LLM may output interactions that may be distressing or inappropriate.

- Patient well-being, namely centring around protection of the long-term interests of the grievers

This ethical principle centres on whether patient well-being is served through the design and use of the griefbot(s) in question. While the grieving process differs from person to person, this period is typically marked by increased vulnerability. This vulnerability can potentially be compounded and exploited by the griefbot, whether in intended or unintended ways. Further, chatbots have in general been demonstrated to lead to emotional dependency and reliance. Risks of unhealthy emotional dependency can be heightened through the use of griefbots as it relates to a loved one who has passed on. In operationalising the ethical principle of centring patient well-being, McGill and its industrial partners will design and monitor the testing process to guard against emotional abuse and unhealthy over-reliance on the bot.

- Multiculturalism

The third ethical principle on multiculturalism relates to the inclusion of the different cultures in Canada (including indigenous nations), specifically pertaining to how they approach and conceptualise griefbots. It is pertinent to examine if cultural elements can inform the resistance or reception towards the use of such chatbots for grieving purposes.

4. Coverage and limitations

This report covered the selection of the most relevant ethical principles and values pertaining to use cases and AI research areas that impact upon human cognition and behaviour. As such, one key limitation of the report is that while ethical principles and values might be relevant in relation to other reasons, such as for innovation, progress and scientific advancement, the parameters informing the selection process is limited to the former, namely its impacts on human cognition and behaviour. This report also does not decisively take a cost-benefit approach in analysing the applicable ethical principles to use cases and research areas. While the introduction and operationalisation of ethical principles and values in AI use cases can mean more added costs for industrial actors, this is only one out of the six elements taken into consideration while shortlisting the ethical principles.

Finally, advancements in the development and deployment of AI may mean that weight and importance attached to certain values may change or decrease over time. This is for example where industries or individuals may prefer to use AI due to its general reliability despite its lack of transparency. Hence, the value of reliability may come to replace the value of transparency (Danaher & Sætra, 2022). While such value evaluation and eventual substitution is an interesting and evolving area of AI ethical research as the use of AI is increasingly dispersed in society, such an examination is beyond the scope of this report.

Finally, as already noted in different parts of this report, the ethical principles selected for the use cases relate to its potential for operationalisation, including through technical means. The principles selected for the research areas, on the other hand, lifts the ethical concerns raised in use cases and will inform the context enriched non-technical AI ethics guidelines for the research areas in the next phase of the project. In this way, the selected principles for research areas also reflect a longer temporal aspect, taking a longer-term perspective in relation to the research area, but also accommodating the aspect of scale – namely the societal-level impacts of the research area as these technologies get more diffused and deployed in the future.

5. Conclusion

This comprehensive report reasoned and justified the selection of the most relevant ethical principles pertaining to the use cases selected by European partners, namely the use of AI in medicine, vehicle safety testing, recruitment, law enforcement, AI as personal assistants and deepfaked therapy for grief and trauma. The report also presents the ethical principles of the use cases selected by international partners. The findings in this report, specifically, the selected ethical principles, will inform a co-creation process between academic and industrial partners in AIOLIA to operationalise these principles in practice.

The report also analysed the ethical principles for the research areas which are common across European and international use cases, namely in the areas of image recognition, AI decision-support systems, general purpose AI and emotional AI. The selected ethical principles for the research areas will subsequently inform a bottom-up co-creation of context-enriched, non-technical AI ethics guidelines. This report contributes towards a better understanding of the ethical AI landscape, aiding towards its operationalisation and eventual dispersion of this knowledge through trainings and workshops for various stakeholders.

6. Selection tables

Table 1. Selection of ethical principles for the European partners’ use cases

#	Sphere of influence	Use case	AI research areas	Selected ethical principles
1	Change in human expertise and professional behaviour	Medical doctors (radiologists, surgeons) using AI tools in diagnostics and treatment	Decision-support systems Image recognition	1. Non-maleficence 2. Accountability and responsibility 3. Transparency and explainability
2		Safety engineers using AI tools to speed up software release approvals	Decision-support systems General-purpose AI (GPAI)	1. Robustness, safety and reliability 2. Oversight and autonomy 3. Risk of over-reliance and deskilling
3		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 de-skilling
4		Security professionals using AI tools to detect hate speech	Decision-support systems General-purpose AI (GPAI)	1. Freedom of expression and non-censorship 2. Non-bias, fairness and non-discrimination 3. Accountability and responsibility
5	Change in human cognition and private behaviour	AI systems as individual and family-level virtual assistants	Conversational general-purpose AI (GPAI) Emotional AI	1. Non-manipulation and non-maleficence 2. Privacy, consent and data protection
6		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

Table 2. Selection of ethical principles for AI research areas in the European partners' use cases

AI research area	Sphere of influence	Use cases	Selected ethical principles
General-purpose AI (GPAI)	Change in human expertise and professional behaviour	2: Safety engineers using AI tools to speed up software release approvals 4: Security professionals using AI tools to detect hate speech	1. Oversight and autonomy 2. Labour impacts
	Change in human cognition and private behaviour	5: AI systems as personal and family virtual assistants 6: Deepfake AI-based psychotherapy for processing trauma and grief	1. Non-manipulation and autonomy 2. Societal well-being and environmental sustainability
Emotional AI	Change in human cognition and private behaviour	5: AI systems as personal and family virtual assistants 6: Deepfake AI-based psychotherapy for processing trauma and grief	1. Human dignity 2. Non-manipulation and autonomy 3. Reliability and accuracy
Decision-support	Change in human expertise and professional behaviour	1: Medical doctors (radiologists, surgeons) using AI in diagnostics and treatment 2: Safety engineers using AI tools to speed up software release approvals 3: Recruiters using AI tools in hiring processes 4: Security professionals using AI tools to detect hate speech	1. Non-bias, fairness and non-discrimination 2. Over-reliance and de-skilling 3. Explainability and transparency
Image recognition	Change in human expertise and professional behaviour	1: Medical doctors (radiologists, surgeons) using AI in diagnostics and treatment	1. Robustness, safety and reliability 2. Transparency and explainability 3. Privacy and data protection

Table 3. Selection of ethical principles for the international partners' use cases

#	AI research	Sphere of influence	Use cases	Ethical principles selected by international partner
7	Emotional AI Image recognition Video analysis	Change in human expertise and professional behaviour	Workplaces equipped with AI tools for behavioral analysis	1. Proportionality 2. Fairness and non-discrimination 3. Transparency and explainability
8	GPAI Emotional AI Image recognition Decision-support systems	Change in human cognition and private behaviour	AI systems for smart elderly care in Wuxi city	1. Privacy and data security 2. Emotional dependency and risk of deception 3. Algorithmic bias 4. Accountability
9	GPAI Emotional AI		AI systems as 'personal companions' to assist senior citizens	1. Prevention of psychological manipulation 2. Diminished autonomy 3. Accountability (and the requirement of explainability)
10	GPAI Emotional AI		AI systems as grief-alleviating personal assistants (griefbots, chatbots designed to imitate a deceased loved one)	1. Dignity of the deceased 2. Patient well-being (i.e. protecting the long-term interests of grievers) 3. Multiculturalism

Table 4. Selection of ethical principles for AI research areas in the international partners' use cases

AI research area	Sphere of influence	Use cases	Selected ethical principles
General purpose AI (GPAI)	Change in human cognition and private behaviour	11. AI systems for smart elderly care in Wuxi city 12. AI systems as 'personal companions' to assist senior citizens 13. AI systems as grief-alleviating personal assistants (griefbots, chatbots designed to imitate a deceased loved one)	1. Non-manipulation and autonomy 2. Societal well-being and environmental sustainability
Emotional AI	Change in human expertise and professional behaviour	7. Workplaces equipped with AI tools for behavioral analysis	1. Fairness and non-discrimination 2. Transparency and explainability
	Change in human cognition and private behaviour	8. AI systems for smart elderly care in Wuxi city 9. AI systems as 'personal companions' to assist senior citizens 10. AI systems as grief-alleviating personal assistants (griefbots, chatbots designed to imitate a deceased loved one)	1. Human dignity 2. Non-manipulation and autonomy 3. Reliability and accuracy
Decision-support	Change in human expertise and professional behaviour	8. AI systems for smart elderly care in Wuxi city 9. AI systems as 'personal companions' to assist senior citizens	1. Over-reliance and de-skilling 2. Explainability and transparency 3. Non-bias, fairness and non-discrimination
Image recognition (video analysis)	Change in human expertise and professional behaviour	9. Workplaces equipped with AI tools for behavioral analysis 10. AI systems for smart elderly care in Wuxi	1. Robustness, safety and reliability 2. Transparency and explainability 3. Privacy and data protection

References

Abdelwanis, M., Alarafati, H. K., Tammam, M. M. S., & Simsekler, M. C. E. (2024). Exploring the risks of automation bias in healthcare artificial intelligence applications: A Bowtie analysis. *Journal of Safety Science and Resilience*, *5*(4), 460–469. https://doi.org/10.1016/j.jnlssr.2024.06.001

AI Risk Management Framework, National Institute for Standards and Technology, https://csrc.nist.gov/projects/risk-management/about-rmf

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. https://doi.org/10.1093/jopart/muac007

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565. https://doi.org/10.48550/arXiv.1606.06565

Awumey, E., Das, S., & Forlizzi, J. (2024). A systematic review of biometric monitoring in the workplace: Analyzing socio-technical harms in development, deployment and use. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 920-932).

Article 19. Hate Speech Explained: A Summary. (2020). https://www.article19.org/resources/hate-speech-explained-a-summary/

Bakir, V., Bennet, K., Bland, B., Laffer, A., Li, P., & McStay, A. (2024). When is Deception OK? Developing the IEEE Recommended Practice for Ethical Considerations of Emulated Empathy in Partner-based General-Purpose Artificial Intelligence Systems (IEEE P7014. 1). In 2024 IEEE International Symposium on Technology and Society (ISTAS) (pp. 1-6). IEEE

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. Social cognitive and affective neuroscience, 12(1), 1-23.

Barros, S. (2025). I Think, Therefore I Hallucinate: Minds, Machines, and the Art of Being Wrong. arXiv preprint arXiv:2503.05806.

Boine, C. (2023). Emotional attachment to AI companions and European law. MIT Schwarzman College of Computing Social and Ethical Responsibilities of Computing. https://mit-serc.pubpub.org/pub/ai-companions-eu-law/release/2

Bradford, A. (2020). The Brussels effect: How the European Union rules the world. Oxford University Press.

Brandon. (2021). New York's New Right of Publicity Law: Protecting Performers and Producers. New York State Bar Association. https://nysba.org/new-yorks-new-right-of-publicity-law-protecting-performers-and-producers/

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

Buss, S., & Westlund, A. (2018). Personal Autonomy. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Spring 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

C-203/22 Dun & Bradstreet Austria, ECLI:EU:C:2024:745.

C-634/21 SCHUFA Holding (Scoring), ECLI:EU:C:2023:957.

Center for European Policy Studies. (2024) AI at work: Why there's more to it than task automation. Available at: https://www.ceps.eu/ceps-publications/ai-at-work/

Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. Humanities and Social Sciences Communications, 10(1), 1-12.

Chen, W. X., Srinivasan, S., & Zakerinia, S. (2025). Displacement or Complementarity? The Labor Market Impact of Generative AI. Harvard Business School Working Paper 25-039.

Christman, J. (2020). Autonomy in Moral and Political Philosophy. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/

Ciriello, R., Hannon, O., Chen, A. Y., & Vaast, E. (2024). Ethical tensions in human-AI companionship: A dialectical inquiry into Replika. In Proceedings of the 57th Hawaii International Conference on System Sciences (Paper 5). Hilton Hawaiian Village, Honolulu, Hawaii. https://aisel.aisnet.org/hicss-57/cl/ethics/5

Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, *26*(4), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., & D'Alfonso, S. (2023). To chat or bot to chat: Ethical issues with using chatbots in mental health. Digital Health, 9. https://doi.org/10.1177/20552076231183542

Cohen, H., & Aulbach, L. (2024). Beyond artificial intelligence ethics: Exploring empathetic ethical outcomes for artificial intelligence. In Ethics in Online AI-based Systems: Risks and Opportunities in Current Technological Trends (pp. 279-295). Intelligent Data-Centric Systems. https://doi.org/10.1016/B978-0-443-18851-0.00017-2

Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & 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

Cortiz, D., & Zubiaga, A. (2021). Ethical and technical challenges of AI in tackling hate speech. The International Review of Information Ethics, 29, 1–10. https://doi.org/10.29173/irie416

Crowston, K., Bolici, F., & e del Lazio Meridionale, C. (2025). Deskilling and upskilling with generative AI systems. Proceedings of the iConference.

Danaher, J., & Sætra, H. S. (2022). Technology and moral change: The transformation of truth and trust. *Ethics and Information Technology*, *24*(3), 35. https://doi.org/10.1007/s10676-022-09661-y

Dewitte, P. (2024). Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. Computer Law & Security Review, 54, 106019. https://doi.org/10.1016/j.clsr.2024.106019

Ekman, P. (2000). Basic emotions. In T. Dalgleish & M. Power (Eds.), Handbook of cognition and emotion (pp. 45–60). John Wiley & Sons.

EU High Level Expert Group on AI, Assessment List for Trustworthy AI, 2021 https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

European Commission. (2024). Third draft of General-Purpose AI Code of Practice published by independent experts. Publications Office of the European Union. https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center Research Publication No. 2020-1, Available at SSRN: https://ssrn.com/abstract=3518482 or http://dx.doi.org/10.2139/ssrn.3518482

Fraser, H. L., & Suzor, N. P. (2025). Locating fault for AI harms: a systems theory of foreseeability, reasonable care and causal responsibility in the AI value chain. Law, Innovation and Technology, 17(1), 103-138.

Gabriel, I., & Manzini, A. (2024). The ethics of advanced AI assistants. Google Deepmind. https://deepmind.google/discover/blog/the-ethics-of-advanced-ai-assistants/

Gao, Ziwei. (2024). “Why Does AI Companionship Go Wrong?”. The International Review of Information Ethics 34 (1). Edmonton, Canada. https://doi.org/10.29173/irie526.

Ghotbi, N. (2023). The ethics of emotional artificial intelligence: a mixed method analysis. Asian Bioethics Review, 15(4), 417-430.

Gremsl, T., & Hödl, E. (2022). Emotional AI: legal and ethical challenges. Information Polity, 27(2), 163-174.

Gumusel, E. (2025). A literature review of user privacy concerns in conversational chatbots: A social informatics approach: An Annual Review of Information Science and Technology (ARIST) paper. Journal of the Association for Information Science and Technology, 76(1), 121-154. https://doi.org/10.1002/asi.24898

Guttmann, M., & Ge, M. (2024). Research Agenda of Ethical Recommender Systems based on Explainable AI. Procedia Computer Science, 238, 328-335.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hagendorff, T. (2024). Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. Minds & Machines 34, 39. https://doi.org/10.1007/s11023-024-09694-w

Hajikhani, A., & Cole, C. (2024). A critical review of large language models: Sensitivity, bias, and the path toward specialized ai. Quantitative Science Studies, 5(3), 736-756.

Harbinja, E., Edwards, L., & McVey, M. (2023). Governing ghostbots. Computer Law & Security Review, 48, 105789. https://doi.org/10.1016/j.clsr.2023.105789

Hasal, M., Nowaková, J., Saghair, K. A., Abdulla, H., Snášel, V., & Ogiela, L. (2021). Chatbots: Security, privacy, data protection, and social aspects. Concurrency and Computation: Practice and Experience, 33(19). https://doi.org/10.1002/cpe.6426

Helm, P., Bella, G., Koch, G., & Giunchiglia, F. (2024). Diversity and language technology: how language modeling bias causes epistemic injustice. Ethics and Information Technology, 26(1), 8.

Hoek, S., Metselaar, S., Ploem, C., & Bak, M. A. R. (2024). Promising for patients or deeply disturbing? The ethical and legal aspects of deepfake therapy. *Journal of Medical Ethics*. Advance online publication. https://doi.org/10.1136/jme-2024-109985

Högberg, C., Larsson, S., & Lång, K. (2024). Engaging with artificial intelligence in mammography screening: Swedish breast radiologists’ views on trust, information and expertise. DIGITAL HEALTH, 10, 20552076241287958. https://doi.org/10.1177/20552076241287958

Hunkenschroer, A.L., Luetge, C. (2022). Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda. J Bus Ethics 178, 977–1007. https://doi.org/10.1007/s10551-022-05049-6

Hunkenschroer, A.L., Kriebitz, A. (2023). Is AI recruiting (un)ethical? A human rights perspective on the use of AI for hiring. AI Ethics 3, 199–213. https://doi.org/10.1007/s43681-022-00166-4

IEEE Standard for Ethical Considerations in Emulated Empathy in Autonomous and Intelligent Systems, in IEEE Std 7014-2024 , vol., no., pp.1-51, 28 June 2024, doi: 10.1109/IEEESTD.2024.10576666.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), Article 9. https://doi.org/10.1038/s42256-019-0088-2

Juwara, L., El-Hussuna, A., & El Emam, K. (2024). An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns*, *5*(4), 100946. https://doi.org/10.1016/j.patter.2024.100946

Katirai, A. (2024). Ethical considerations in emotion recognition technologies: A review of the literature. *AI and Ethics*, *4*(4), 927–948. https://doi.org/10.1007/s43681-023-00307-3

Kay, J., Kasirzadeh, A., & Mohamed, S. (2024, October). Epistemic injustice in generative ai. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (Vol. 7, pp. 684-697).

Kelly, J. (2025, May 4). It's time to get concerned: Klarna, UPS, Duolingo, Cisco and many other companies are replacing workers with AI. Forbes. https://www.forbes.com/sites/jackkelly/2025/05/04/its-time-to-get-concerned-klarna-ups-duolingo-cisco-and-many-other-companies-are-replacing-workers-with-ai/

Kaminski, M. E., & Malgieri, G. (2025). The Right to Explanation in the AI Act. Available at SSRN 5194301.

Kidmose, B. (2025). A review of smart vehicles in smart cities: Dangers, impacts, and the threat landscape. Vehicular Communications, 51, Article 100871. https://doi.org/10.1016/j.vehcom.2024.100871

Kiener, M. (2025). AI and responsibility: No gap, but abundance. Journal of Applied Philosophy, 42(1), 357-374.

Kraaijeveld, S., Ivanova, D., & Bak, M. (2024). Het nieuwe rouwen? Over deepfakes en relaties met overleden dierbaren. Podium voor Bio-ethiek, 31(4), 21–26.

Liu, H.-Y., & Zawieska, K. (2020). From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence. *Ethics and Information Technology*, *22*(4), 321–333. https://doi.org/10.1007/s10676-017-9443-3.

Lubin, A. (2022). The rights to privacy and data protection under international humanitarian law and human rights law. In Research Handbook on Human Rights and Humanitarian Law (pp. 462-491). Edward Elgar Publishing.

Luna, F. (2009). Elucidating the Concept of Vulnerability: Layers Not Labels. International Journal of Feminist Approaches to Bioethics, 2(1), 121–139.

Malgieri, G. (2023). Vulnerability and Data Protection Law. Oxford University Press.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6, 175–183. https://doi.org/10.1007/s10676-004-3422-1.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35.

Metikoš, L., & Ausloos, J. (2025). The right to an explanation in practice: insights from case law for the GDPR and the AI Act. Law, Innovation and Technology, 17(1), 205–240. https://doi.org/10.1080/17579961.2025.2469349

McCrudden, C. (2008). Human dignity and judicial interpretation of human rights. european Journal of international Law, 19(4), 655-724.

McStay, A. Emotional AI, Ethics, and Japanese Spice: Contributing Community, Wholeness, Sincerity, and Heart. Philos. Technol. 34, 1781– 1802 (2021). https://doi.org/10.1007/s13347-021-00487-y

Mohammad, S. M. (2022). Ethics sheet for automatic emotion recognition and sentiment analysis. Computational Linguistics, 48(2), 239-278.

Mori, M., Sassetti, S., Cavaliere, V., & Bonti, M. (2024). A systematic literature review on artificial intelligence in recruiting and selection: A matter of ethics. Personnel Review. Advance online publication. https://doi.org/10.1108/pr-03-2023-0257

Murtarelli, G., Gregory, A., & Romenti, S. (2021). A conversation-based perspective for shaping ethical human–machine interactions: The particular challenge of chatbots. Journal of Business Research, 129, 927–935. https://doi.org/10.1016/j.jbusres.2020.11.004

Najjar, R. (2023). Redefining radiology: a review of artificial intelligence integration in medical imaging. Diagnostics, 13(17), 2760.

Nass, S. J., Levit, L. A., Gostin, L. O., & Rule, I. of M. (US) C. on H. R. and the P. of H. I. T. H. P. (2009). The Value and Importance of Health Information Privacy. In Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK9579/

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Panigutti, C., Hamon, R., Hupont, I., Llorca, D. F., Yela, D. F., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Garrido, J. S., & Gomez, E. (2023). The role of explainable AI in the context of the AI Act. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23) (pp. 1139-1150). Association for Computing Machinery. https://doi.org/10.1145/3593013.359406

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, *5*(5). https://doi.org/10.1016/j.patter.2024.100988

Pentney, K., & McGonagle, T. (2020, September). The opportunities and challenges of addressing hate speech with artificial intelligence: Submission to the OSCE Representative on Freedom of the Media #SAIFE Public Consultation. OSCE Representative on Freedom of the Media

Pinker, S. (2008). The Stupidity of Dignity. *New Republic (New York, NY)*. https://dash.harvard.edu/handle/1/38822030

Portacolone, E., Halpern, J., Luxenberg, J., Harrison, K. L., & Covinsky, K. E. (2020). Ethical issues raised by the introduction of artificial companions to older adults with cognitive impairment: A call for interdisciplinary collaborations. Journal of Alzheimer's Disease, 76(2), 445-455. https://doi.org/10.3233/JAD-190952

Pulignano, V., & Doellgast, V. (2020). The challenge of digital transformation in the automotive industry (Chapter 7). European Trade Union Institute. https://www.etui.org/sites/default/files/2020-09/The%20challenge%20of%20digital%20transformation%20in%20the%20automotive%20industry-2020.pdf

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), Pub. L. No. 32016R0679, 119 (2016). http://data.europa.eu/eli/reg/2016/679/oj/eng

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (2024). http://data.europa.eu/eli/reg/2024/1689/oj/eng

Ren, S., & Wierman, A. (2024, July 15). The Uneven Distribution of AI's Environmental Impacts. *Harvard Business Review*. https://hbr.org/2024/07/the-uneven-distribution-of-ais-environmental-impacts

Rigotti, C., & Fosch-Villaronga, E. (2024). Fairness, AI & recruitment. *Computer Law & Security Review*, *53*, 105966. https://doi.org/10.1016/j.clsr.2024.105966SSRN+5Economic Research Journals+5OUCI+5

Rueda, J., Ausín, T., Coeckelbergh, M., del Valle, J. I., Lara, F., Liedo, B., Llorca Albareda, J., Mertes, H., Ranisch, R., Raposo, V. L., Stahl, B. C., Vilaça, M., & de Miguel, Í. (2025). Why dignity is a troubling concept for AI ethics. *Patterns*, *6*(3), 101207. https://doi.org/10.1016/j.patter.2025.101207

Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. Journal of Information, Communication and Ethics in Society, 19(1), 61-86.

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. International Data Privacy Law, 7(4), 233–242. https://doi.org/10.1093/idpl/ipx022

Shevlin, H. Ethics at the frontier of human-AI relationships. Available at: https://philarchive.org/rec/SHEEAT-12

Shrishak, K., & Warso, Z. (2024, May 21). *Hope: The AI Act's Approach to Address the Environmental Impact of AI*. Tech Policy Press. https://techpolicy.press/hope-the-ai-acts-approach-to-address-the-environmental-impact-of-ai

Söderlund, K., Engström, E., Haresamudram, K., Larsson, S., & Strimling, P. (2024). Regulating high-reach AI: On transparency directions in the Digital Services Act. Internet Policy Review, 13(1). https://doi.org/10.14763/2024.1.1746

Soori, M., Jough, F. K. G., Dastres, R., & Arezoo, B. (2024). AI-based decision support systems in Industry 4.0, A review. Journal of Economy and Technology.

Stark, L., & Hoey, J. (2021, March). The ethics of emotion in artificial intelligence systems. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 782-793).

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. Internet policy review, 8(2), 1-22.

Teo, S. A. (2023). Human dignity and AI: Mapping the contours and utility of human dignity in addressing challenges presented by AI. *Law, Innovation and Technology*, *15*(1), 241–279. https://doi.org/10.1080/17579961.2023.2184132

Tamascelli, N., Campari, A., Parhizkar, T., & Paltrinieri, N. (2024). Artificial Intelligence for safety and reliability: A descriptive, bibliometric and interpretative review on machine learning. Journal of Loss Prevention in the Process Industries, 105343

Triguero, I., Molina, D., Poyatos, J., Del Ser, J., & Herrera, F. (2023). General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, societal implications and responsible governance. Information Fusion, 103, 102135. https://doi.org/10.1016/j.inffus.2023.102135

Udupa, S., Maronikolakis, A., & Wisiorek, A. (2023). Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence. Big Data & Society, 10(1), 1–15. https://doi.org/10.1177/20539517231172424

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000381137

van Bekkum, M. (2025). Using sensitive data to de-bias AI systems: Article 10 (5) of the EU AI act. Computer Law & Security Review, 56, 106115.

Varkey, B. (2021). Principles of clinical ethics and their application to practice. Medical Principles and Practice, 30(1), 17-28.

Wachter S., Mittelstadt B. & Russell C. (2021). Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law, 123 W. Va. L. Rev. 735. Available at: https://researchrepository.wvu.edu/wvlr/vol123/iss3/4

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wagner, B. (2018). Ethics as an Escape from Regulation.: From “Ethics-Washing” to Ethics-Shopping? In E. Bayamlioglu, I. Baraliuc, L. Janssens, & M. Hildebrandt (Eds.), *BEING PROFILED* (pp. 84–89). Amsterdam University Press; JSTOR. https://doi.org/10.2307/j.ctvhrd092.18

Zewe, A. (2025, January 17). *Explained: Generative AI’s environmental impact*. MIT News | Massachusetts Institute of Technology. https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117