EthicAI4CARE
Ethics by Design Framework for Trustworthy AI in Healthcare

The integration of artificial intelligence (AI) into healthcare holds immense potential to revolutionize the sector – enhancing diagnostic accuracy, enabling personalized treatment pathways, and increasing operational efficiency. However, with these benefits come significant ethical challenges, particularly concerning the protection of patient autonomy in AI-based clinical decision-making, the governance of sensitive health data, algorithmic bias that may lead to disparities in diagnosis or treatment across different patient groups, and accountability for AI-driven clinical outcomes.

As AI systems become pervasively used in healthcare and increasingly embedded in clinical practice, the need for a robust, context-sensitive ethical framework for their trustworthy development, implementation, and use becomes more and more critical. Despite the growing body of literature and ethical frameworks developed in the last few years aimed at fostering responsible and ethical AI in various sectors, there remains a lack of a comprehensive framework that translates the core principles and requirements underpinning the European (EU) strategy on AI (such as those proposed in the EU *Ethics Guidelines for Trustworthy AI*) into actionable guidance specifically tailored to the healthcare sector. Filling this gap is especially urgent given the pervasive use of AI in healthcare and the resulting need to train and upskill university professors, educators, and researchers in the field, who play a pivotal role in shaping future healthcare professionals (or healthcare providers: HCPs), so as to ensure they will be capable of engaging with AI technologies in a responsible, ethical, and critical manner.

Methodology

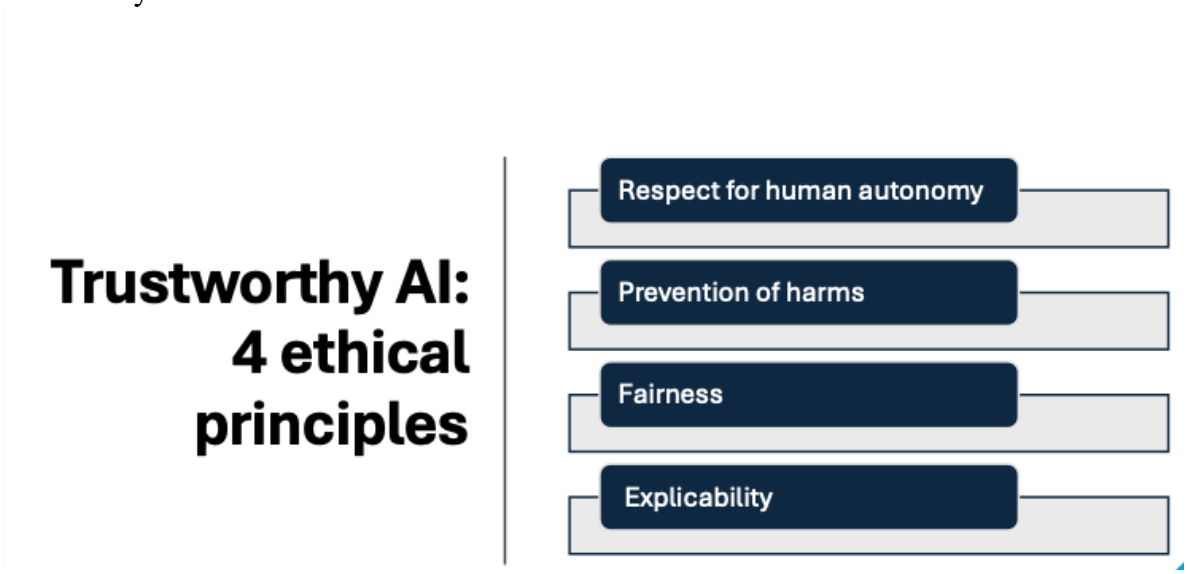
The elaboration of the ethics framework and the self-assessment tool for trustworthy AI in healthcare, as required in WP2 of the project, will be based on the benchmark AI ethics frameworks underpinning the EU digital strategy and approach for trustworthy AI: the Ethics Guidelines for Trustworthy AI, developed by the European Commission (EC)'s High-Level Expert Group on AI (HLEG AI), and ALTAI – the Assessment List for Trustworthy AI, the practical tool developed to help organizations implement the EU Ethics Guidelines. We also align our work with the EU AI Act, which classifies certain AI systems used in healthcare as high risk. The Act imposes obligations on providers and deployers, such as implementing risk management and data governance systems, maintaining technical documentation and logs, and ensuring human oversight and transparency.

However, while the EU Ethics Guidelines for Trustworthy AI and its operational tool ALTAI provide a comprehensive foundation for ethical and trustworthy AI, it is important to note they are designed as general-purpose frameworks and therefore are not specifically tailored for healthcare AI solutions. Our project seeks to bridge that gap, while also attempting to translate the AI Act's obligations into practical requirements for healthcare settings. (see WP2).

In particular, the goal of our project in WP2 is to develop an AI ethics framework and self-assessment tool for trustworthy AI explicitly adapted for the healthcare and clinical domains, building upon the EU's established principles of trustworthy (i.e., lawful, ethical, and technically robust) AI, while addressing the sector's particular demands for health equity, inclusiveness, transparency, and patient-centered care.

The elaboration of the framework will therefore integrate the EU ethical guidelines and tools for trustworthy AI with other relevant frameworks developed beyond the EU ones for ethical AI in healthcare, as those developed by international organizations (see, for example, WHO 2021 "Ethics and Governance of AI in Health", and WHO 2023 "Guidance on Large Multimodal Models") and academic scholarship. Building over the EU approach for trustworthy and ethical AI, the ethical framework and self-assessment tool for trustworthy AI in healthcare is based on the respect of four benchmark ethics principles (i.e., the respect of human autonomy, the prevention of harms, fairness, and explicability) and entails the operationalization of seven key ethical requirements that ought to

be respected and implemented in order to promote the design, development, deployment, and use of trustworthy AI in healthcare.



Foundations for Trustworthy AI:
AI Ethics principles for Trustworthy AI in Healthcare

1. Respect for Human Autonomy

The principle of respect for human autonomy – central to the EU fundamental rights – demands that trustworthy AI systems in healthcare support and uphold the self-determination and dignity of patients and healthcare professionals and providers (HCPs). Individuals interacting with AI in clinical contexts must retain full and effective control over their decisions, particularly in matters that affect their health, treatment options, and personal well-being. AI systems must not coerce, deceive, or manipulate users, whether clinicians or patients, but rather be designed to support human expertise, empower informed decision-making, and reinforce trust in the patient-HCPs relationship. The allocation of responsibilities between humans and AI should follow human-centric design principles, ensuring that final authority and accountability remain with qualified healthcare professionals. This includes embedding mechanisms for meaningful human oversight in all AI-supported clinical processes, from diagnostic assistance to healthcare and treatment recommendations. Furthermore, as AI systems increasingly reshape clinical roles and workflows, they should be introduced in ways that support meaningful professional work (not substitute or deskill it), promote collaborative practices, and enhance the clinical judgment and relational skills essential to high-quality care.

2. Prevention of Harm

AI systems in healthcare must be designed and deployed in ways that do not cause or exacerbate harm to patients, healthcare professionals, or society at large. This includes the protection of human dignity, as well as physical and mental integrity, which are foundational to medical ethics and bioethics. AI systems used in clinical environments must meet the highest standards of safety, security, and technical robustness, minimizing the risk of errors, adverse outcomes, or unintended consequences in diagnosis, treatment, or health and care management. In this regard, ensuring resilience to system

failures, cyberattacks, or misuse is essential, especially given the sensitive and high-stakes nature of healthcare tasks, activities, and decisions. Special attention must be given to vulnerable individuals and populations – such as the elderly, persons with disabilities, children, and those with complex or rare conditions, ensuring they are not exposed to disproportionate risk. Furthermore, specific attention must be devoted to avoiding power and information asymmetries that may arise when patients or even clinicians (and HCPs) do not fully understand or are unable to contest AI-generated outputs. Trustworthy AI deployment must also consider the potential environmental impact of healthcare AI, promoting sustainability of the human and environmental ecosystem.

3. Fairness

The development, deployment, and use of AI systems in healthcare must uphold the principle of fairness, which encompasses both substantive and procedural dimensions. Substantively, fairness requires that AI systems contribute to the equitable distribution of healthcare benefits and burdens, and actively prevent harm from unfair bias, discrimination, or stigmatization. This is particularly critical in healthcare, where disparities in diagnosis, treatment, and access to services can have life-changing consequences. AI must be designed to recognize and address systemic inequalities, ensuring that no patient group – such as those from marginalized ethnic, gender, socioeconomic, or geographic communities – is disproportionately misdiagnosed, undertreated, or excluded from the benefits that can stem from AI deployment and use. Fairness also demands that patients and HCPs are not deceived or manipulated, and that their freedom of clinical and personal choice is preserved. Equal access to AI-driven healthcare innovations, digital tools, and services must be fostered to avoid deepening existing healthcare inequalities. Developers and healthcare institutions should respect the principle of proportionality, carefully balancing efficiency gains with the need to protect patient rights, clinical judgment, and equitable outcomes. Procedurally, fairness requires that decisions made by or with the assistance of AI systems are contestable and accountable. Patients and healthcare professionals must have clear avenues to seek redress, challenge AI-supported outcomes, and understand the rationale behind clinical recommendations. This necessitates that the responsible parties are identifiable, and that decision-making processes are intelligible, well-documented, and aligned with medical ethics and legal standards.

4. Explicability

Explicability is essential for building and maintaining trust in AI systems within healthcare, where decisions often carry significant implications for patient health, safety, and well-being. This principle requires that AI systems be transparent in their design, purpose, and capabilities, and that their outputs are, to the extent possible, understandable and interpretable by those directly and indirectly affected – especially clinicians, patients, and caregivers. In healthcare contexts, the ability to explain why an AI system generated a particular diagnosis, risk prediction, or treatment recommendation is not merely a matter of user preference but a clinical and ethical necessity. Such explanations enable informed consent, support shared decision-making, and empower clinicians to critically evaluate AI outputs against their own expertise and medical guidelines. However, not all AI models are inherently interpretable – so-called “black box” algorithms may not provide meaningful insights or evidence about how decisions are made. In these cases, additional explicability measures must be implemented, such as the traceability of data inputs and decision pathways, the auditability of system performance and outcomes, and the transparent communication about system limitations, accuracy levels, and areas of uncertainty. The required level of explicability should be proportional to the clinical risk involved: the higher the potential for harm or misjudgment, the greater the demand for interpretability and oversight. Even when full technical explainability is not feasible, the overall system must still uphold patients’ fundamental rights, ensure human accountability, and provide clear mechanisms for contesting and reviewing AI-supported decisions.



Requirement #1: Human Agency and Oversight

Trustworthy healthcare AI systems should support HCPs' and patients' agency and decision-making, as prescribed by the principle of respect for human autonomy. This requires that such systems support (1) HCPs' and patients' agency, not replace or undermine it, and (2) uphold fundamental rights, which should be underpinned through mechanisms for meaningful human oversight on healthcare AI functioning and outputs, especially on critical healthcare decisions (as those concerning vulnerable populations or rare pathological conditions).

Assessment list for trustworthy AI in healthcare:

(1) Human agency and autonomy in healthcare AI

- Is the healthcare AI system designed to interact with, guide, or make decisions for healthcare professionals (HCPs) or patients that could affect individual health outcomes or broader public health?
  - Could the system create confusion for HCPs, patients, or caregivers about whether a diagnosis, recommendation, treatment plan, or outcome was generated by an algorithm rather than a human expert?
  - Are HCPs and patients clearly informed when clinical content, advice, or decisions are influenced or produced by an AI system?
- Could the system lead users (especially patients or vulnerable individuals) to believe they are interacting with a human when they are actually interacting with an AI?
  - Are users (e.g., patients, family members, or clinicians) explicitly informed when an AI system is acting as a conversational agent, triage assistant, or health recommender system?
- Could the AI system foster excessive reliance by HCPs or patients, leading to uncritical acceptance of its outputs?
  - Have safeguards been implemented – such as alerts, documentation, or training – to ensure that users retain critical oversight and do not defer blindly to AI-generated recommendations?

- Could the AI system inadvertently interfere with the clinical decision-making process (e.g., by nudging users toward specific treatments without justification or by obscuring alternative options)?
    - Have procedures been established to prevent unintended influence on the autonomy of HCPs or patients?
- Does the AI system simulate human-like social interaction (e.g., empathy, conversation, rapport) in a clinical or caregiving context?
    - If so, are mechanisms in place to ensure that this simulation supports rather than undermines trust, informed consent, and the autonomy of the patient?
    - If so, are mechanisms in place to ensure that this simulation does not undermine informed consent and the autonomy of the patient?
- Could the AI system – especially if it simulates empathy, conversational responses, or caregiving roles – lead to disproportionate emotional attachment from patients, caregivers, or healthcare professionals? If so, answer to the following questions:
    - Could the AI system's design or interaction model (e.g., constant availability, reward feedback loops, gamified elements) foster compulsive or addictive behaviors in patients or users?
    - Have features been audited or adjusted to minimize the risk of overuse or psychological dependence, particularly among vulnerable populations (e.g., those with mental health conditions or chronic loneliness)?
    - Are there any practices of monitoring and/or periodic re-evaluation of the tool based on data regarding attachment and psychological dependence?
    - Have measures been put in place to address potential negative consequences of such attachment (e.g., distress when the AI is unavailable or discontinued)?
    - Could the AI system unduly influence user behavior – such as nudging treatment compliance, lifestyle changes, or commercial choices – beyond medically justified boundaries?
    - Have mechanisms been implemented to prevent manipulation, including transparency about intent, clear opt-out options, and ethical boundaries for persuasive design?

(2) Human oversight in healthcare AI

- What type of oversight applies to the AI system in the healthcare context (select all that apply)?

  ☐ The system is *self-learning* or *autonomous*, adapting its outputs based on new data without ongoing human oversight.

  ☐ The system operates under *Human-in-the-Loop* oversight (human actively involved in real-time decision-making).

  ☐ The system operates under *Human-on-the-Loop* oversight (human supervises and can intervene post-hoc).

  ☐ The system operates under *Human-in-Command* oversight (human sets goals and can override or shut down the system)
- Have all responsible personnel (e.g., clinicians, operators, administrators) received role-specific training to understand and effectively exercise oversight over the AI system?
- Have mechanisms been implemented to detect and respond to adverse or undesirable effects (e.g., false positives, clinical misjudgments, ethical red flags) of the AI system on patients, HCPs, or care outcomes?
- Is there a clear and accessible procedure (a "stop button" or abort protocol) in place to safely interrupt or shut down the system when necessary (e.g., in case of malfunction or patient harm)?

- If the AI system is self-learning or autonomous, have you implemented additional control measures to address the risks associated with ongoing adaptation (e.g., unexpected behavior, performance drift, audit logging)?
- If the system operates under *Human-on-the-Loop* oversight, is the post-hoc intervention designed to prevent, and always capable of preventing, any substantial harm to the user or others?

Requirement #2: Technical Robustness and Safety

Healthcare AI systems must be secure, accurate, and resilient to errors or attacks across all stages of their lifecycle. They should be designed to resist and operate safely under technical faults, adversarial attacks, and inappropriate use, especially in high-stakes clinical environments. These systems must meet rigorous standards of accuracy, with robust clinical validation and safeguards to detect errors, prevent performance drift, and maintain reliability across diverse patient populations and care settings. Security protections – including threat assessments, red-teaming, and alignment with healthcare cybersecurity standards – are essential to protect patient safety and data integrity. Fail-safe mechanisms, fallback procedures, and transparency around risks and limitations must be in place to ensure that AI supports, rather than compromises, clinical care.

Assessment list for trustworthy AI in healthcare:

(1) Resilience to Attack and Security in healthcare AI

- Could the AI system pose risks to patient safety, public health, or clinical integrity in the event of design flaws, technical faults, outages, adversarial attacks, or inappropriate/malicious use?
- Is the system compliant with applicable healthcare cybersecurity regulations and standards, such as the EU Cybersecurity Act, HIPAA Security Rule, ISO/IEC 27001, or relevant FDA guidance on medical device cybersecurity?
- How exposed is the AI system to cyber-attacks, particularly in settings with high sensitivity like hospital networks, electronic health records (EHR), or diagnostic tools?
- Have you assessed potential attack vectors specific to healthcare applications?
- Have you evaluated the system's vulnerability to:
    - Data poisoning, e.g., manipulation of training datasets derived from clinical records.
    - Model evasion, e.g., the system being tricked into misdiagnosis or inaccurate risk assessments.
    - Model inversion, e.g., the potential to infer personal health data from model outputs.
- Have you implemented safeguards to ensure the integrity, robustness, and availability of the AI system throughout its lifecycle, especially during critical clinical use?
- Has the system been red-teamed or penetration-tested in a healthcare-relevant environment?
- Have healthcare providers and users been clearly informed about:
    - The duration and scope of security support and updates?
    - The procedures for reporting vulnerabilities or incidents?
    - The existing risks and the back-up systems set up- procedures?

(2) General Safety in Healthcare AI

- Have you defined the clinical and operational risks associated with the AI system for each specific healthcare use case (e.g., diagnostics, treatment recommendation, patient monitoring)?

    - Is there a process in place for continuous risk monitoring and reassessment, including during post-deployment use in clinical settings?
    - Have healthcare professionals, patients, and relevant stakeholders been informed of known or potential risks, such as false positives, false negatives, or data-driven biases?
- Have you identified potential threats to safety, including:
    - Design faults (e.g., inadequate consideration of medical edge cases),
    - Technical faults (e.g., model drift due to changing clinical protocols),
    - Environmental threats (e.g., degraded performance in low-resource or emergency settings)?
    - Have you assessed the risks of malicious, unintended, or inappropriate use, especially in high-stakes scenarios (e.g., automated triage, remote diagnostics)?
    - Have you defined safety-criticality levels, particularly in relation to harm to patient health, dignity, or well-being?
- Have you assessed the dependency of clinical decisions on the AI system's stability and reliability?
    - Are the system's reliability and testing protocols aligned with clinical safety requirements (e.g., consistent performance across patient subgroups, environments, and time)?
- Have you planned for fault tolerance through redundancies, such as:
    - A fallback to human review or clinician override,
    - A conventional system or alternative diagnostic method,
    - Or a parallel AI system for cross-validation in safety-critical cases
- Have you established mechanisms to re-evaluate safety and robustness whenever the AI system undergoes:
    - Significant updates (e.g., retraining, new input features),
    - Integration into a new clinical context,
    - Or observed changes in real-world performance?

(3) Accuracy in Healthcare AI

- Could a low level of accuracy in the AI system lead to harmful clinical consequences, such as misdiagnosis, inappropriate treatment recommendations, or failure to detect critical conditions?
- Have you ensured that the training and validation data:
    - Is current, medically relevant, and consistent with the latest clinical standards?
    - Is high-quality, complete, and representative of the diverse patient populations, care settings, and conditions in which the AI system will be deployed?
- Have you implemented a robust process to monitor and document the system's accuracy, including:
    - Pre-deployment clinical validation,
    - Ongoing performance tracking in real-world settings, and
    - Regular recalibration where appropriate?
- Have you considered how the deployment context might cause a mismatch with training assumptions – e.g., changes in clinical workflows, patient demographics, or data sources – and how this might lead to performance degradation or adversarial consequences?
- Have you established procedures to ensure that clinicians, patients, and healthcare institutions are clearly informed about:
    - The expected accuracy and limitations of the AI system,
    - Any performance thresholds or confidence scores to guide interpretation, and
    - The appropriate role of human oversight when acting on AI outputs?

(4) Reliability, Fall-back plans and Reproducibility in Healthcare AI

- Could low reliability or reproducibility of the AI system result in critical harm, such as incorrect clinical decisions, patient safety incidents, or erosion of trust in medical judgment?
  - Have you established a structured process to monitor whether the AI system is consistently achieving its intended clinical goals, including effectiveness, safety, and alignment with benchmark standards of care?
  - Have you evaluated whether specific clinical contexts (e.g., ICU vs outpatient care) or patient subpopulations affect the system's reproducibility and reliability?

- Have you implemented comprehensive verification and validation procedures, including:
  - Clinical validation studies under real-world conditions?
  - Systematic logging of inputs, outputs, and decision pathways for traceability and auditability?
  - Continuous monitoring for performance degradation or drift?

- Have you documented and operationalized testing protocols to ensure reliability and reproducibility over time, including across:
  - Different healthcare settings (urban/rural, high/low resource),
  - Devices or platforms (e.g., integration with EHR, mobile health tools),
  - Updates to clinical data or guidelines?

- Have you designed failsafe fallback mechanisms (e.g., escalation to human decision-makers, alert systems) and clearly defined governance protocols for activating them when errors, anomalies, or system failures occur?
- Is there a defined process for managing low-confidence outputs, such as:
  - Flagging results for human review,
  - Suppressing or qualifying uncertain recommendations,
  - Communicating uncertainty to clinicians or patients?

- If your AI system uses continual or online learning:
  - Have you assessed the risks of undesirable adaptation, such as the system learning from spurious correlations, feedback loops, or overfitting to recent or biased data?
  - Have you established controls and validation gates before newly learned behavior is deployed in clinical workflows?

Requirement #3: Privacy and Data Governance

Trustworthy healthcare AI must respect privacy and ensure data protection through secure data practices, transparency in data use, and proper control over personal data. Sensitive health data requires strict protection. AI systems must ensure compliance with GDPR, enable patient control over personal health data, and use techniques like data anonymization and secure data sharing. Consent processes should be clear and meaningful.

Assessment list for trustworthy AI in healthcare:

(1) Privacy in Healthcare AI

- Have you assessed how the AI system may impact patients' fundamental rights, particularly:

- The right to privacy, including the confidentiality of medical data and sensitive health information?
- The right to physical and mental integrity, especially in cases involving diagnostic or therapeutic interventions?
- The right to data protection, including compliance with regulations such as the GDPR, HIPAA, or local health data laws?

- Depending on the specific healthcare use case (e.g., mental health, reproductive care, genomic medicine), have you:
  - Implemented context-sensitive privacy protections, such as differential privacy, pseudonymization, or local data minimization?
  - Established mechanisms for patients or clinicians to flag privacy concerns, such as:
    - Inappropriate inferences or profiling,
    - Misuse of sensitive health attributes,
    - Overreach in data access or retention?

- Have you ensured transparency about how patient data is used, including:
  - What data is collected, for what purpose, and who has access?
  - Whether data is shared with third parties or used for further training?
  - How individuals can exercise their rights to access, correct, or delete their health data?

(2) Data Governance in Healthcare AI

- Was your AI system developed or trained using personal data, including special categories of personal data such as health records, genetic data, biometric identifiers, or mental health information?
- Have you implemented all required and recommended data protection measures, particularly those mandated under the General Data Protection Regulation (GDPR) or equivalent healthcare data protection frameworks (e.g., HIPAA, UK DPA 2018)? These may include:
  - Conducting a Data Protection Impact Assessment (DPIA) prior to development or deployment,
  - Appointing and involving a Data Protection Officer (DPO) at an early stage of the AI system's lifecycle (development, procurement, or deployment),
  - Establishing strict oversight mechanisms for data access and processing:
    - Access restricted to qualified personnel only,
    - Audit logs to monitor and trace data access and changes,
    - Role-based access control and regular permission reviews.
  - Implementing privacy-by-design and by-default measures, such as:
    - Data pseudonymization, encryption, and aggregation,
    - Secure storage and transmission protocols,
    - Anonymization where feasible, especially during model training.
  - Ensuring data minimisation, particularly when dealing with sensitive medical or diagnostic data.

- Have you incorporated mechanisms to uphold patient data subject rights, including:
  - The right to withdraw consent for data processing at any time,
  - The right to object to automated decision-making that affects clinical outcomes,
  - The right to erasure ("right to be forgotten"), where applicable?

- Have you considered the long-term implications of data privacy throughout the AI system's entire lifecycle, including:

  - Updates, retraining, integration with other systems, and
  - Data generated through usage (e.g., predictions, logs, new medical annotations)?

- Even when using non-personal data, have you:
  - Assessed whether it could indirectly lead to re-identification or bias (e.g., through rare disease patterns or small cohort datasets)?
  - Evaluated the data protection and ethical risks associated with these sources?

- Is your system aligned with relevant standards and best practices, such as:
  - ISO/IEC 27701 (privacy information management),
  - ISO/IEC 27001 (information security),
  - IEEE 7002 (data privacy process),
  - Or any national health data governance protocols?

Requirement #4: Transparency

Healthcare AI systems should be understandable and explainable. This includes traceability of processes, transparency of algorithms, and clear communication about system capabilities and limitations. Patients and HCPs must understand how and why an AI system reached a particular recommendation or diagnosis. This includes explainability of algorithms, traceable data sources, and clear communication about limitations and intended use cases of the AI system.

Assessment list for trustworthy AI in healthcare:

(1) Traceability in healthcare AI

- Have you implemented traceability mechanisms to track the AI system's behavior and evolution across its entire lifecycle, from development through clinical deployment and updates?

- Do you continuously assess the quality and integrity of input data, especially when it involves:
  - Electronic Health Records (EHRs), lab results, imaging data, or real-time monitoring feeds
  - Sources that may be subject to noise, variation, or inconsistent coding across care settings

- Can the system support backward traceability to:
  - Identify which specific input data was used in generating a clinical decision, diagnosis, or treatment recommendation
  - Determine which version of the AI model, algorithm, or rules were active when the output was generated

- Have you established ongoing processes to assess the quality and clinical relevance of outputs, including:
  - Checking for consistency with clinical guidelines,
  - Monitoring for unexpected or harmful recommendations, and
  - Comparing outputs against clinician judgment where feasible

- Have you implemented adequate logging practices that:
  - Capture and store decision-making steps, including data inputs, model versioning, and final outputs
  - Allow for auditable records in the event of adverse events, regulatory review, or patient inquiries

- Are traceability measures designed to be transparent to clinical staff, without disrupting workflows, while enabling:
  - Root cause analysis for errors or discrepancies,
  - Explainability of outputs for informed consent and shared decision-making?

(2) Explainability

- Have you ensured that the decisions or recommendations made by the AI system are explained in a way that is understandable and clinically meaningful to its intended users (e.g., physicians, nurses, medical technicians)?
  - Are explanations tailored to the clinical context, such as justifying a diagnosis, triage decision, or risk prediction based on identifiable features (e.g., lab values, symptoms)?
  - Where applicable, are patient-facing explanations also available in clear, non-technical language to support informed consent and shared decision-making?

- Do you continuously evaluate whether users (both clinical and, if applicable, patient users) understand the AI system's outputs and how to interpret them?
- Have you implemented feedback mechanisms or usability studies to assess:
  - Whether explanations aid decision-making or cause confusion,
  - Whether users can recognize when not to trust the AI,
  - Whether users are appropriately relying on or overriding AI recommendations?

(3) Communication

- In cases where the AI system interacts directly with users (e.g., medical chatbots, virtual health assistants, AI symptom checkers), have you clearly communicated that users are engaging with an AI system, not a human clinician?

- Have you established mechanisms to inform users – patients, caregivers, or healthcare professionals – about:
  - The purpose of the AI system (e.g., triage support, information provision, wellness coaching),
  - The criteria and logic the system uses to generate responses or recommendations (where explainable),
  - And the limitations of the AI system, including what it is not intended to do (e.g., it is not a substitute for professional medical advice)?

- Have you clearly communicated the benefits of using the AI system, such as improved access, faster responses, or reduced administrative burden?
- Have you also communicated the technical limitations and potential risks, such as:
  - Possible accuracy issues or error rates,
  - Limitations in understanding complex medical histories or nuanced symptoms,
  - And risks of misuse, overreliance, or misinterpretation?

- Have you provided:
  - Training materials or onboarding instructions for clinical staff and users on how to safely and effectively use the system,
  - Disclaimers that clarify the system's role (e.g., "not a substitute for medical diagnosis") and promote appropriate human oversight?

Requirement #5: Diversity, Non-Discrimination, and Fairness

Trustworthy healthcare AI systems must avoid unfair bias and be inclusive. This means their design, development, and use should ensure accessibility and the fair treatment of all individuals. To do so, it is crucial that AI in healthcare is trained on diverse population data to avoid biases that could lead to misdiagnosis or unequal treatment, with special regard to data on minorities and rare pathologies. Fairness in healthcare AI entails ensuring substantially equitable access to AI-based healthcare tools and related benefits with specific attention to already marginalized and underrepresented groups.

Assessment list for trustworthy AI in healthcare:

(1) Avoidance of Unfair Bias

- Have you developed a bias mitigation strategy that addresses the risk of unfair bias across the entire AI system lifecycle, including:
  - Input data (e.g., demographic imbalance, underrepresentation of rare diseases),
  - Algorithm design choices (e.g., model tuning that may inadvertently prioritize certain patient profiles),
  - And output interpretation (e.g., how recommendations affect different groups of patients)?

- Have you ensured that the training and validation data is:
  - Representative of diverse patient populations, including variations by age, sex, gender, ethnicity, socio-economic status, geography, and comorbidity?
  - Inclusive of historically marginalized, underserved, or clinically complex groups (e.g., persons with disabilities, elderly individuals, pregnant persons, people with rare or chronic conditions)?

- Have you tested the AI system for:
  - Performance disparities across specific subgroups or use cases, such as language minorities, rural populations, or non-binary individuals?
  - Clinical fairness issues (e.g., different false negative rates by race or sex in diagnostic tools)?

- Have you used state-of-the-art technical tools to:
  - Analyze model behavior and feature importance across demographic groups?
  - Detect proxy variables that may encode bias (e.g., zip code as a proxy for income)?
  - Monitor and quantitatively assess fairness metrics, such as equalized odds, calibration by group, or disparate impact?

- Have you embedded bias auditing and mitigation processes throughout the AI lifecycle, including retraining, post-deployment monitoring, and real-world validation?

- Have you implemented educational initiatives to raise awareness among developers, data scientists, and clinicians involved in the design and deployment of the AI system, regarding:
  - How bias can be embedded through data, model assumptions, or interface design?
  - The ethical and clinical consequences of biased decision-making?

- Have you created a clear process to flag and address issues of bias, including:
  - User-friendly channels for clinicians, patients, or affected communities to report concerns,
  - Clear protocols for reviewing flagged cases, and

- Corrective actions to adjust system behavior or training data as needed?

- Have you identified all stakeholder groups and patient populations that may be directly or indirectly affected by the AI system?

- Did you consider multiple definitions of fairness (e.g., procedural fairness vs outcome fairness) before selecting the most appropriate one for your healthcare use case?

- Did you consider consulting representative communities (e.g., disability advocates, elderly patient groups, gender-diverse individuals) in defining what fairness means in your clinical context?

- Did you establish mechanisms to ensure fairness in your healthcare AI system?
  - Have you operationalized fairness using quantitative fairness metrics, and integrated them into your model evaluation and deployment criteria?

(2) Accessibility and Universal Design

- Have you ensured that the AI system accommodates the variety of preferences, abilities, and access needs found across patients, caregivers, and healthcare professionals?

- Does the system account for variations in language, literacy, digital skills, and cognitive or physical abilities?

- Have you assessed whether the AI system's user interface is accessible to individuals with disabilities or those at risk of exclusion, including:
  - Users with visual, hearing, motor, or cognitive impairments?
  - Individuals who rely on assistive technologies, such as screen readers, alternative input devices, or closed captioning?

- Have you ensured that all information presented by the AI system (e.g., alerts, recommendations, explanations) is:
  - Available in accessible formats,
  - Designed for clarity, simplicity, and usability in high-stakes clinical environments?

- Have you involved or consulted with patients and clinicians with accessibility needs during the planning, design, and testing of the AI system?

- Have you applied Universal Design principles – designing for all users from the outset – throughout development and deployment phases?

- Have you evaluated the impact of the AI system on different end-users, particularly:
  - Whether any groups (e.g., elderly, rural populations, people with low digital literacy) may face barriers to use, understanding, or benefit?
  - Whether certain populations may be disproportionately affected by AI-driven decisions, such as those with rare diseases or chronic conditions often excluded from training datasets?
- Has the development team proactively engaged with target end-users, such as:
  - Healthcare providers across specialties and settings,
  - Patient advocacy groups,

- Representatives from marginalized or underserved communities?

- Have you assessed the potential risks of exclusion, discrimination, or downstream harm, especially where the AI system's outputs could:
  - Influence access to care,
  - Modify treatment plans,
  - Or inadvertently perpetuate inequities in healthcare delivery?

(3) Stakeholder Participation

- Did you establish a mechanism to ensure the meaningful participation of a broad and diverse range of stakeholders throughout the AI system's design, development, validation, and deployment?
- Have you engaged with relevant actors such as:
  - Clinicians, nurses, and allied health professionals from various disciplines and care settings,
  - Patients and caregivers, including those from vulnerable or underserved populations,
  - Hospital administrators, IT professionals, and compliance officers,
  - Medical ethicists, legal experts, and data protection officers,
  - Public health authorities or regulators,
  - And advocacy groups representing affected communities (e.g. disability rights organizations, elderly associations, mental health networks)?
- Have you ensured that stakeholder input meaningfully shaped system design, especially in areas such as:
  - Use case definition and scope,
  - Risk identification and mitigation,
  - Definitions of fairness, transparency, and usability?
- Is there a continuing mechanism for feedback and consultation during post-deployment monitoring and updates?

Requirement #6: Societal and Environmental Well-Being

Trustworthy healthcare AI should contribute to the well-being of the healthcare human and environmental ecosystem and society at large. It should be designed and deployed in ways that promote public health outcomes, support preventive care, empower healthcare professionals (not displace or de-skill them), and ensure equitable access to healthcare services. In doing so, it must be designed to reduce healthcare disparities, particularly those affecting underserved or vulnerable populations, and promote democratic, fair, and inclusive healthcare resources. Trustworthy healthcare AI must be ecologically responsible, minimizing its environmental impact through energy-efficient design, responsible data use, and sustainable deployment practices both for the environment and for individuals. This entails training healthcare staff and users (including patients) in system use and limitations, ensuring that decision-making remains human, and that any redesign healthcare workflow and task distributions be transparent, participatory, and sensitive to the needs of workers at all levels of the healthcare system.

Assessment list for trustworthy AI in healthcare:

(1) Environmental Well-being

- Have you considered whether the AI system could have negative environmental impacts across its development, deployment, and use in healthcare environments?

- Which specific impacts have you identified, such as:
    - High energy and water consumption from large-scale training or model inference,
    - Carbon emissions associated with data storage and computation in clinical infrastructure,
    - Electronic waste or increased hardware demands in medical settings?

- Where feasible, have you implemented mechanisms to evaluate the AI system's environmental footprint, including:
    - Energy and water usage during training, deployment, and continuous operation (e.g., in diagnostic imaging systems or hospital triage tools),
    - Cloud infrastructure or data center sustainability profiles,
    - The resource consumption associated with real-time or large-scale inference?

- Have you adopted measures to minimize environmental impact throughout the system's lifecycle, such as:
    - Model optimization techniques (e.g., pruning, distillation, or efficient architectures),
    - Deploying models on low-power hardware or using green data centers,
    - Encouraging responsible update cycles and modular system design to reduce waste?

- Have you balanced the system's clinical benefit with its environmental cost, particularly in settings where sustainability is a concern (e.g., mobile or rural healthcare units, or hospitals with limited energy resources)?

(2) Impact on Healthcare Providers' Work and Skills

- Does the AI system impact the work, roles, or responsibilities of healthcare professionals (e.g., clinicians, nurses, lab technicians, administrative staff)?

- Before deployment, did you inform and consult with affected workers and their representatives – such as clinical leadership, hospital unions, medical staff associations, or healthcare work councils – to prepare for changes introduced by the system?

- Have you taken steps to ensure that the impacts of the AI system on human work are clearly understood, including:
    - How the system integrates into existing clinical workflows and decision-making chains,
    - Which tasks are automated or augmented, and what remains under human oversight or judgment?
- Have you ensured that healthcare workers:
    - Understand how the AI system operates, including its capabilities, limitations, and appropriate use cases?
    - Know when and how to override or question AI recommendations?

- Could the AI system lead to de-skilling of healthcare professionals (e.g., overreliance on automated diagnosis or decision support)?
    - If so, have you implemented measures to counteract de-skilling, such as:
        - Ensuring continued engagement in core clinical reasoning,
        - Promoting human-in-the-loop processes,
        - Facilitating case review and reflective learning?

    - Does the AI system require or encourage new digital or data literacy skills?

- Have you provided training materials, upskilling opportunities, and ongoing support to help staff use AI systems safely, navigate related challenges, and adapt?
- Are these resources accessible to all staff, including those in lower-resourced or rural care settings?

(3) Impact on Society at large

- Could the deployment of the AI system in healthcare have negative consequences for society at large, such as:
  - Exacerbating health inequalities,
  - Undermining public trust in medical institutions,
  - Or contributing to the centralization of power and decision-making in opaque, unaccountable systems?
- Have you assessed the societal impact of the AI system beyond the immediate users and patients, including:
  - Indirectly affected communities (e.g., populations excluded from care, stigmatized by algorithmic decisions, or subject to public health surveillance),
  - The influence of the system on public discourse, social norms, or expectations about medical care?
- Have you taken proactive steps to minimize societal harm, such as:
  - Avoiding the use of AI in coercive or punitive health policies (e.g., insurance exclusion, mandatory treatment recommendations without oversight),
  - Promoting equitable access to AI-supported healthcare services,
  - And ensuring transparency and contestability of decisions made with AI support?
- Have you implemented safeguards to ensure that the AI system:
  - Does it not erode democratic values, such as liberty, justice, equality, and participation, by safeguarding, for example, individual autonomy, informed consent, and institutional accountability?
  - Does it support discriminatory practices, profiling, or surveillance that could undermine civic trust or marginalize vulnerable groups?

Requirement 7: Accountability

Mechanisms must be in place to ensure responsibility and accountability for healthcare AI systems and their outcomes, including auditability, reporting, and redress mechanisms when harm for patients, caregivers, and HCPs (including clinicians and medical staff) occurs. Trustworthy healthcare AI systems must be supported by clear use-case frameworks that assign responsibility and accountability for their deployment and outcomes. These frameworks should include mechanisms for auditability, structured incident reporting protocols, and HCPs/patient-centered accessible procedures for redress in cases of harm or failure (including clear pathways for patients and HCPs to challenge or correct erroneous decisions). The goal is to ensure that accountability is both proactive – through auditability and reporting – and reactive – by enabling meaningful responses when adverse outcomes occur.

Assessment list for trustworthy AI in healthcare:

(1) Auditability in healthcare AI

- Have you established mechanisms that ensure the AI system is auditable, particularly in ways that support clinical safety, regulatory compliance, and public trust?
- Does auditability cover the entire development and deployment lifecycle, including:
  - Traceability of decisions and outputs,
  - Documentation of design and development processes,
  - Sourcing and composition of training and validation data (e.g., origin, representativeness, inclusion criteria),
  - Logging of model updates, performance metrics, and clinical outcomes?

- Have you ensured that the AI system is auditable by independent third parties, including:
  - Healthcare regulators, ethics boards, or accreditation bodies (e.g., FDA, EMA, MHRA),
  - External auditors or academic experts,
  - Or patient advocacy groups and watchdog organizations, when appropriate?

- Are audit mechanisms designed to respect patient confidentiality while enabling:
  - Post hoc review of adverse outcomes,
  - Evaluation of bias, fairness, and safety claims,
  - And transparent reporting to stakeholders, including clinicians, patients, and the public?

(2) Risk management in healthcare AI

- Have you foreseen the involvement of external guidance or third-party auditing bodies to oversee ethical concerns and accountability measures related to the healthcare AI system?
  - Does this oversight extend beyond the development phase, covering deployment, clinical integration, post-market surveillance, and continuous monitoring?

- Have you provided risk training to relevant stakeholders (e.g. developers, clinical staff, compliance officers), including education on:
  - Potential ethical and clinical risks,
  - The applicable legal and regulatory frameworks, such as GDPR, HIPAA, the EU AI Act, or medical device regulations?
  - Have you considered establishing an AI ethics review board, multidisciplinary oversight committee, or similar mechanism within your healthcare organization to:
    - Review, discuss, and advise on ethical issues, accountability practices, and value conflicts;
    - Provide structured input on "grey areas", such as balancing clinical benefit against patient autonomy or data privacy?

  - Have you put in place a process for continuous monitoring and evaluation of the AI system's adherence to this healthcare AI ethical framework?
    - Does this process involve:
      - Tracking conflicts or tensions between key principles (e.g. accuracy vs fairness, explainability vs privacy),
      - Documenting trade-offs and justifications for decisions made during development and deployment?

  - Have the individuals involved in governance and monitoring been given appropriate training, including:
    - Familiarity with the ethical, legal, and clinical dimensions of AI in healthcare?
    - Understanding of how to handle ethical dilemmas and ensure proportional responses?

- Have you established a clear and trusted channel for third parties (e.g. clinicians, patients, suppliers, vendors, distributors) to:
    - Report potential vulnerabilities, safety risks, ethical issues, or algorithmic bias?
    - Trigger reviews and revisions to the risk management process based on that input?

- For AI applications that may adversely affect individuals (e.g. treatment eligibility, risk scoring, diagnostics), have you implemented "redress by design" mechanisms, such as:
    - Appeals processes or human review options
    - Transparent error correction pathways,
    - And clear communication of patient rights and recourse when harmed by or disagreeing with AI decisions?

Glossary

This glossary explains the main technical and ethical terms used in the EthicAI4CARE 'Ethics by Design Framework and Self-Assessment Tool for Trustworthy AI in Healthcare' to support understanding among non-experts and project's stakeholders.

Word	**Meaning**
AI system	It means a machine-based system that, for explicit or implicit objectives, infers from the input it receives how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Such systems are designed to operate with varying levels of autonomy and may exhibit adaptiveness after deployment.
Artificial Intelligence Act (EU AI Act)	The European Union's first comprehensive AI law. The Act classifies certain AI-driven medical software as high-risk and imposes obligations on providers and deployers (e.g., risk-management, data-governance, transparency and human-oversight mechanisms), while requiring generative models to label AI-generated content and summarise their training data.
Auditability	The capacity to examine and reconstruct an AI system's decisions and processes through traceable documentation and logs. Auditability supports regulatory oversight, ethical accountability, and public trust.
Automation Boundary	The dividing line between the tasks assigned to the AI system and those that are and should be handled by humans. It is essential to define these boundaries in every context in which the AI system is employed, in order to understand how the system integrates into existing clinical workflows and decision-making chains.
Autonomous	Describes AI systems that perform tasks or make decisions without continuous human intervention. Autonomy in healthcare is ethically sensitive, as it may shift accountability from clinicians to machines. Regulatory frameworks require that such systems remain subject to meaningful human control.

Bias	Systematic and unjustified distortion in AI performance resulting from imbalanced data, flawed model design, or contextual misapplication. In healthcare, bias may lead to disparities in diagnosis or treatment, especially across gender, racial, or socioeconomic groups. Mitigation requires representative datasets and fairness auditing.
Data governance	The structured management of data quality, access, storage, and sharing in accordance with ethical and legal norms. Effective governance ensures accountability and compliance with data protection regulations.
Data Poisoning	A type of cyber-attack that involves the manipulation of AI training datasets derived from clinical records for malicious purposes.
De-skilling	The erosion of professional expertise resulting from excessive reliance on automation. In healthcare, this can weaken clinicians' diagnostic reasoning and critical thinking.
European Health Data Space (EHDS)	The European Union's sector specific regulation for electronic health data, adopted in 2025, that gives individuals more control over their health information and sets harmonised rules for the primary and secondary use of health data. Once fully implemented, it aims to enable secure, high quality data sharing and reuse, thereby providing the datasets needed for developing and monitoring AI driven medical software.
Fairness	In this context, fairness has both a substantive and a procedural dimension. Substantively, fairness requires that AI systems contribute to the equitable distribution of healthcare benefits and burdens, and actively prevent harm from unfair bias, discrimination, or stigmatization. This is particularly critical in healthcare, where disparities in diagnosis, treatment, and access to services can have life-changing consequences. Equal access to AI-driven healthcare innovations, digital tools, and services must be fostered to avoid deepening existing healthcare inequalities. Procedurally, fairness requires that decisions made by or with the assistance of AI systems are contestable and accountable. Patients and healthcare professionals must have clear avenues to seek redress, challenge AI-supported outcomes, and understand the rationale behind clinical recommendations.
Foundation models	A foundation model is a type of artificial intelligence that is trained using machine learning or deep learning on broad datasets. Due to their extensive training on large amounts of data, foundation models can operate in a wide range of use cases. Well-known examples of foundation models include Large Language Models, such as ChatGPT, as well as DALL-E for image generation from text and MusicGen for music generation. Unlike general-purpose models, which are also trained on large amounts of data, foundation models are typically pre-trained and fine-tuned for specific applications, whereas the former are designed for a wide range of tasks without extensive customization.

General-purpose AI models	General-purpose AI models are trained on a huge amount of data and are designed to perform a wide range of tasks without requiring a task-specific pre-training. They can also generate content and operates across multiple domains such as translation, problem-solving, data analysis and so on. Advanced AI assistants, such as CoPilot, are well-known examples of this kind of AI models. The European AI Act regulates extensively this kind of AI model requiring transparency about training data, limitations and capabilities, as well as risk classification and compliance with EU fundamental rights and values.
Generative AI	This term refers to a specific feature of existing AI models—their ability to generate new content, including text, images, audio, and video. These models learn statistical patterns from existing data and use that learning to generate realistic outputs. These outputs differ from the training data that inspired them. While general-purpose AI models are always generative, generative AI models can be limited to specific tasks and are not necessarily general purpose.
GDPR (General Data Protection Regulation)	The European Union's legal framework for protecting personal data, establishing principles such as lawfulness, fairness, transparency, and purpose limitation, which directly apply to AI in healthcare.
Human-in-command (HIC)	The highest level of human authority over an AI system, ensuring that humans retain the capacity to define goals, control functioning, and terminate operations if necessary. It embodies the principle of human accountability.
Human-in-the-loop (HITL)	A design configuration where human experts actively participate in AI decision-making. HITL ensures that automated suggestions are reviewed, validated, or overridden by professionals, preserving clinical judgment and ethical accountability.
Human-on-the-loop (HOTL)	A supervision model where humans monitor AI systems' operation and outcomes but intervene only when anomalies are detected. This mode supports efficiency while maintaining oversight responsibility.
Marginalized/Vulnerable Groups	AI must be designed to recognize and address systemic inequalities, ensuring that no patient group – such as those from marginalized ethnic, gender, socioeconomic, or geographic communities – is disproportionately misdiagnosed, undertreated, or excluded from the benefits that can stem from AI deployment and use. The Charter of Fundamental Rights of the EU encompasses under Article 21 on non-discrimination the following grounds, which can be a reference point amongst others: namely sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age and sexual orientation. Additionally, what constitutes a vulnerable person or group is often context-specific. In the healthcare sector, special attention must be given to vulnerable individuals and populations – such as the elderly, persons with disabilities,

	children, and those with complex or rare conditions, ensuring they are not exposed to disproportionate risk.
Model hallucination	In artificial intelligence, a model hallucination is a phenomenon observed especially in generative AI and large language models where the AI system generates false or nonsensical information, presented as accurate. These phenomena happen for several co-reasons: the model lacks sufficient information to generate the output; it prioritizes optimization and coherence over accuracy, and its probabilistic nature recognizes patterns not facts. AI systems cannot distinguish between correct and incorrect information but they fully depend on the training dataset that can be incomplete or biased. There are several documented examples of AI hallucinations in the healthcare sector uses such as miscalculations in anaesthesia dosages due to insufficient exposure to diverse patient populations during the training phase. More information about AI Hallucination cases in healthcare can be found at the following link: https://bhmpc.com/2024/12/ai-hallucination/
Noise	Random or irrelevant variations in data that can interfere with an AI system's accuracy. A robust system should perform well even when data contain noise.
Nudge	A concept from behavioral economics describing subtle design interventions that steer user behavior without restricting freedom of choice. In AI interfaces, nudges can encourage beneficial actions (e.g., treatment adherence) but risk manipulation when used without transparency or consent.
Ongoing adaptation	The dynamic modification of an AI system's behavior through exposure to new data or contexts. This property raises governance challenges, as system performance and risk profiles may change over time.
Opt-out	Opting out means choosing not to participate in, or withdrawing from, a specific activity or service. Patients should always be able to identify the healthcare services for which they wish to withdraw consent, and they should be provided with instructions on how to do so. It is crucial to ensure that the opt-out process is straightforward and accessible to respect patient rights.
Performance/Model Drift	AI model drift occurs when the performance of an AI model degrades over time because the real-world data it encounters differs from the data it was originally trained on. This can cause the model to make inaccurate or even harmful predictions. To mitigate AI drift, it's important to continuously monitor model

	performance, retrain models with fresh data, and use techniques like ensemble methods or anomaly detection to adapt to changing conditions.
Post-deployment monitoring	Continuous observation and evaluation of an AI system after its introduction into real-world environments, to detect performance drift, emergent risks, or unintended social consequences.
Potential attack vectors	Weak points or methods that hackers could exploit to manipulate or damage an AI system, such as altering training data or system access.
Privacy-by-design	A design methodology that integrates privacy and data protection principles throughout all stages of AI development. It aligns with the GDPR requirement that systems safeguard personal information as a default setting.
Prohibited practices	Article 5 of the European Regulation AI Act lists the types and uses of AI systems that must be prohibited. These prohibited AI uses are the following: • deploying subliminal, manipulative, or deceptive techniques to distort behaviour and impair informed decision-making, causing significant harm. • exploiting vulnerabilities related to age, disability, or socio-economic circumstances to distort behaviour, causing significant harm. • biometric categorisation systems inferring sensitive attributes (race, political opinions, trade union membership, religious or philosophical beliefs, sex life, or sexual orientation), except labelling or filtering of lawfully acquired biometric datasets or when law enforcement categorises biometric data. • social scoring, i.e., evaluating or classifying individuals or groups based on social behaviour or personal traits, causing detrimental or unfavourable treatment of those people. • assessing the risk of an individual committing criminal offenses solely based on profiling or personality traits, except when used to augment human assessments based on objective, verifiable facts directly linked to criminal activity. • compiling facial recognition databases by untargeted scraping of facial images from the internet or CCTV footage. • inferring emotions in workplaces or educational institutions, except for medical or safety reasons. • 'real-time' remote biometric identification (RBI) in publicly accessible spaces for law enforcement. (source: https://artificialintelligenceact.eu/high-level-summary/)

Proxy variables	Indirect features used by AI systems that may unintentionally encode sensitive attributes (e.g., using geographic data as a proxy for ethnicity). Their use requires scrutiny to prevent discriminatory outcomes.
Red-Teaming	Red teaming is the practice whereby a "red team" or independent group challenges an organisation to improve its effectiveness by assuming an adversarial role or point of view. It is particularly used to help identifying and addressing potential security vulnerabilities. (Definition from European Commission's Ethics Guidelines for Trustwhorty AI – pg.37)
Regulatory sandbox	"AI regulatory sandbox means a controlled framework set up by a competent authority which offers providers or prospective providers of AI systems the possibility to develop, train, validate and test, where appropriate in real-world conditions, an innovative AI system, pursuant to a sandbox plan for a limited time under regulatory supervision." (Definition from EU AI Act – Art. 3, comm. 55)
Reproducibility	Reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions. This enables scientists and policy makers to accurately describe what AI systems do. Replication files (files that will replicate each step of the AI system's development process, from research and initial data collection to the results) can facilitate the process of testing and reproducing behaviours. (Definition from European Commission's Ethics Guidelines for Trustwhorty AI – pg.17)
Risk management system	"The risk management system shall be understood as a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic review and updating. It shall comprise the following steps: (a) the identification and analysis of the known and the reasonably foreseeable risks that the high-risk AI system can pose to health, safety or fundamental rights when the high-risk AI system is used in accordance with its intended purpose; (b) the estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose, and under conditions of reasonably foreseeable misuse; (c) the evaluation of other risks possibly arising, based on the analysis of data gathered from the post-market monitoring system referred to in Article 72; (d) the adoption of appropriate and targeted risk management measures designed to address the risks identified pursuant to point (a)." (Definition from European AI Act – Article 9 comm. 2)

Robustness	Robustness of an AI system encompasses both its technical robustness (appropriate in a given context, such as the application domain or life cycle phase) and as well as its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no unintentional harm can occur. Robustness is the third of the three components necessary for achieving Trustworthy AI. (Definition from European Commission's Ethics Guidelines for Trustwhorty AI – pg.37)
Self-learning AI	Refers to adaptive AI models that update or refine their internal parameters using new data after deployment. While this capacity can improve performance, it also introduces unpredictability and requires continuous monitoring to prevent harmful or biased behavior.
Societal Harm	In the healthcare sector, uses of AI systems that cause societal harm can be the ones enforcing coercive or punitive health policies (e.g. insurance exclusion, mandatory treatment recommendations without oversight) and the ones that do not promote equitable access to AI-supported healthcare services.
Trustworthy AI	AI that is lawful, ethical, and technically robust, consistent with the European Commission's Ethics Guidelines for Trustworthy AI.
Universal Design	Universal design is the design of products or environments to make them accessible to people, regardless of age, disability, or other factors. This kind of design should follow seven main principles: equitable use (design that does not disadvantage or stigmatize any group of users), flexibility in use (design that accommodates a wide range of individual preferences and abilities), simple/intuitive use (use is easy to understand, regardless of the user's experience, knowledge, language skills, or current concentration level), perceptible information (the design communicates necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities), tolerance for error (the design minimizes hazards and the adverse consequences of accidental or unintended actions), low physical effort (the design can be used efficiently and comfortably, and with a minimum of fatigue), and size and space for approach and use (the design is appropriate size and space).
Vulnerabilities	In computer security, a vulnerability is a flaw in the design, implementation, or management of a system that can be exploited by malicious actors to compromise its security. The likelihood of a vulnerability being exploited varies, and the actual risk depends on the nature of the vulnerability, as well as the value of the surrounding system. While some vulnerabilities can only be used for denial-of-service attacks, more dangerous ones allow attackers to perform code injection without the victim's awareness. Without a vulnerability, an exploit — a method or piece of code that takes advantage of vulnerabilities — typically cannot gain access. Malware can also be installed directly through

	social engineering or poor physical security, such as an unlocked door or exposed IT port.